This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2958030, IEEE Access

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Deformable Non-local Network for Video Super-Resolution

HUA WANG[1], DEWEI SU[1], CHUANGCHUANG LIU[1], LONGCUN JIN[1], (Member, IEEE), XIANFANG SUN[2] AND XINYI PENG[1]

[1]School of Software Engineering, South China University of Technology, Guangzhou 510006, China
[2]School of Computer Science and Informatics, Cardiff University, Cardiff CF24 3AA, U.K.

Corresponding author: Longcun Jin (lcjin@scut.edu.cn)

This work was supported in part by the Pearl River Technology Nova Project under Grant 201710010020, the Fundamental Research Funds for the Central Universities under Grant 2019MS086 and in part by the National Science Foundation of China under Grant 61300135.

ABSTRACT The video super-resolution (VSR) task aims to restore a high-resolution (HR) video frame by using its corresponding low-resolution (LR) frame and multiple neighboring frames. At present, many deep learning-based VSR methods rely on optical flow to perform frame alignment. The final recovery results will be greatly affected by the accuracy of optical flow. However, optical flow estimation cannot be completely accurate, and there are always some errors. In this paper, we propose a novel deformable non-local network (DNLN) which is a non-optical-flow-based method. Specifically, we apply the deformable convolution and improve its ability of adaptive alignment at the feature level. Furthermore, we utilize a non-local structure to capture the global correlation between the reference frame and the aligned neighboring frames, and simultaneously enhance desired fine details in the aligned frames. To reconstruct the final high-quality HR video frames, we use residual in residual dense blocks to take full advantage of the hierarchical features. Experimental results on benchmark datasets demonstrate that the proposed DNLN can achieve state-of-the-art performance on VSR task. The code is available at https://github.com/wh1h/DNLN.

INDEX TERMS Convolutional neural networks, deep learning, deformable convolution, non-local operation, video super-resolution.

I. INTRODUCTION

THE target of super-resolution (SR) is to generate a corresponding high-resolution (HR) image or video from its low-resolution (LR) version. As an extension of single image super-resolution (SISR), video super-resolution (VSR) provides a solution to restore the correct content from the degraded video, so that the reconstructed video frames will contain more details with higher clarity. Such kind of technology with important practical significance can be widely used in many fields such as video surveillance [1], ultra-high definition television [2] and so on.

Different from SISR which only considers one single low-resolution image as input at a time, VSR devotes to effectively making use of intrinsic temporal information among multiple low-resolution video frames. Although vanilla SISR approaches can be directly applied to video frames by treating them as single images, abundant detail information available from neighboring frames will be wasted. Such practice is hard to reconstruct promising video frames, so they are not well adapted to VSR task.

To overcome the limitation of the SISR, existing VSR methods [3]–[7] usually take a LR reference frame and its multiple neighboring frames as inputs to reconstruct a corresponding HR reference frame. Due to the motion of the camera or objects, the neighboring frames should be spatially aligned first for utilizing the temporal information. To this end, most traditional VSR methods [8]–[11] generally calculate the optical flow and estimate the sub-pixel motion between LR frames to achieve the alignment operation. However, fast and reliable flow estimation still remains a challenging problem. The brightness constancy assumption, which most motion estimation algorithms rely on, may be invalid due to the existence of motion blur and occlusion. Incorrect motion compensation will introduce artifacts in aligned neighboring frames and affect the quality of final reconstructed video frames. Hence, explicit flow estimation and motion compensation methods could be sub-optimal for VSR task.

In this paper, we propose a novel deformable non-local network (DNLN), which is non-optical-flow-based, to per-

form both implicit motion estimation and video super-resolution. Our network mainly consists of four modules: feature extraction module, alignment module, non-local attention module and SR reconstruction module. Inspired by TDAN [12], we apply the deformable convolution [13] in our alignment module and enhance its ability of adaptively warping frames. Specifically, we introduce a hierarchical feature fusion block (HFFB) [14] to effectively handle the videos with large and complex motions. Through the stacks of deformable convolutions, we align the neighboring frame to the reference frame at the feature level and gradually improve the alignment accuracy. Then in the non-local attention module, we exploit a non-local structure to capture the global correlation between the reference feature and each aligned neighboring feature, which assesses the importance of different regions in neighboring feature. Such operation is expected to highlight the features complementary to the reference frame and exclude regions with improper alignment. The features with attention guidance are fused and then fed into the final SR reconstruction module. Here, we use residual in residual dense blocks (RRDB) [15] to generate the SR reference frame. RRDBs help to make full use of the information from different hierarchical levels and retain more details of the input LR frame.

In summary, the main contributions of this paper can be concluded as follows:

- We propose a novel deformable non-local network (DNLN) to accomplish high quality video super-resolution. Our method achieves the most advanced VSR performance on several benchmark datasets.
- We design an alignment module based on deformable convolution, which can realize the feature level alignment in a coarse to fine manner without explicitly motion compensation.
- We propose a non-local attention module to select significant features from neighboring frames which are conducive to the recovery of the reference frame.

The rest of the paper is organized as follows: Section 2 introduces the related work. Section 3 elaborates the structure of the proposed network. Section 4 shows our experimental results on benchmark datasets, including visual comparisons with other methods. The effectiveness of each components in our network is analyzed in Section 5. Finally, Section 6 draws conclusions.

II. RELATED WORK

A. SINGLE IMAGE SUPER-RESOLUTION

Dong et al. [16] first proposed SRCNN for single image super-resolution to learn the nonlinear mapping between LR and HR images in an end-to-end manner, which achieves better performance than previous work. Kim et al. [17] further improved SRCNN by stacking more convolution layers and using residual learning to increase network depth. Tai et al. introduced recursive blocks in DRRN [18], which employs parameters sharing strategy to make the training stable. All of these methods first upscale the LR input to the desired size and the reconstruction process is based on the upscaled products. Such pre-processing step inevitably results in loss of details and additional computation cost. To avoid these problems, extracting features from the original LR input and upscaling spatial resolution at the end of the network become the main direction of SR network. Dong et al. [19] directly took the original LR image as input and brought in the transpose convolution layer (also known as the deconvolution layer) for upsampling features to high resolution outcomes. Shi et al. [20] proposed an effective sub-pixel convolution layer for amplifying the final LR feature map to SR output and accelerating the network.

Afterwards, Timofte et al. [21] provided a new large dataset (DIV2K) in the NTIRE 2017 challenge that consists of 1000 2K resolution images. This dataset enables researchers to train deeper and wider networks which leads to various development of SR methods. The most advanced SISR networks, such as EDSR [22], DBPN [23], RDN [24] and RCAN [25], have far better training performance on this dataset than previous networks.

B. VIDEO SUPER-RESOLUTION

Liao et al. [26] proposed DECN and made use of two classical optical flow methods: TV-L1 and MDP flow to generate SR drafts with different parameters, and then produced the final result through a deep network. Kappeler et al. [27] proposed VSRnet, which uses a hand-designed optical flow algorithm to perform motion compensation on the input LR frame, and takes the warped frame as the CNN input to predict the HR video frame. Caballero et al. [11] introduced the first end-to-end VSR network: ESPCN, which studies early fusion, slow fusion, and 3D convolution to learn temporal relationships. They applied a multi-scale spatial transformer to warp the LR frame and eventually generated a HR frame through another deep network. Tao et al. [9] proposed a sub-pixel motion compensation layer for frame alignment and used a convolutional LSTM architecture in following SR reconstruction network. Recently, Haris et al. [28] proposed RBPN which learns from the idea of back-projection to iteratively extract temporal features between frames. They treated frames independently rather than concatenated them together.

Most previous VSR methods exploit optical flow to estimate motion between frames and perform motion compensation to integrate effective features. While various approaches [29]–[33] are proposed to calculate the optical flow, it is still intractable to obtain precise flow estimation in the case of occlusion and large movement. Xue et al. [34] proposed task-oriented TOFlow with learnable task-oriented motion prompts. It achieved better VSR results than fixed flow algorithm, which reveals that standard optical flow is not the best motion representation for video recovery. To circumvent this problem, DUF [35] uses an adaptive upsampling with dynamic filters instead of the explicit estimation process. TDAN [12] uses deformable convolutions to adaptively

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2958030, IEEE Access

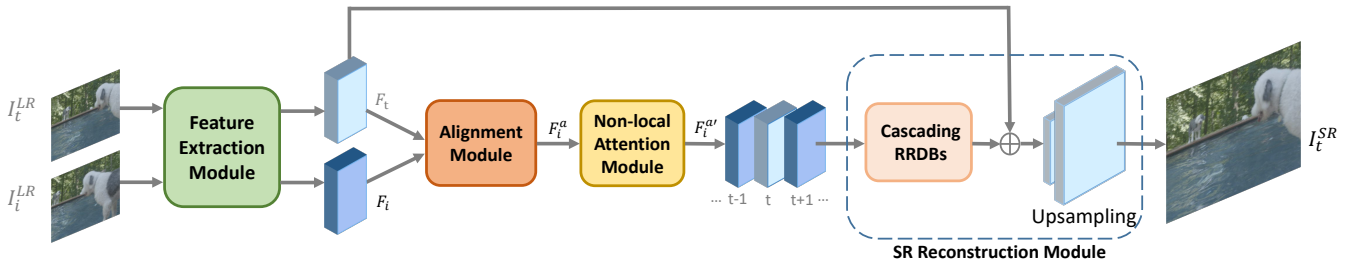


FIGURE 1. The architecture of the proposed DNLN framework. We only show one neighboring frame in the figure. Each neighboring frame will pass through feature extraction module, alignment module and non-local attention module. Then all the features are concatenated and fed into SR reconstruction module to generate HR reference frame.

align the video frame at the feature level without computing optical flow. These kind of methods transcend the flow-based approaches through implicit motion compensation.

C. DEFORMABLE CONVOLUTION

To enhance the CNNs' capability of modeling geometric transformations, Dai et al. [36] proposed deformable convolutions. It adds additional offsets to the regular grid sampling locations in the standard convolution and enables arbitrary deformation of the sampling grid. To further enhance the modeling capability, they proposed modulated deformable convolutions [13] which can further learn modulation scalar for sampling kernels. The modulation scalar lies in the range $[0, 1]$, which can adjust the weight for each sampling location. The deformable convolution is effective for high-level vision tasks such as object detection and semantic segmentation. TDAN [12] is the first to utilize deformable convolutions in the VSR task. It is an end-to-end network which adaptively aligns the input frames at the feature level without explicit motion estimation. EDVR [37] further exploits the deformable convolutions with a pyramid and cascading structure. It shows superior performance to previous optical-flow-based VSR networks.

D. NON-LOCAL BLOCK

Inspired by the classic non-local method in computer vision, Wang et al. [38] proposed a building block for video classification by virtue of non-local operations. For image data, long-range dependencies are commonly modeled via large receptive fields formed by deep stacks of convolutional layers. While the non-local operations capture long-range dependencies directly by computing interactions between any two positions, regardless of their positional distance. It computes the response at a position as a weighted sum of all positions in the input feature maps. The set of positions can be in space, time, or spacetime, so the non-local operations are applicable for image or video problems.

III. DEFORMABLE NON-LOCAL NETWORKS

A. NETWORK ARCHITECTURE

Given a sequence of 2N+1 consecutive low-resolution frames $\{I_{t-N}^{LR}, \ldots, I_{t-1}^{LR}, I_t^{LR}, I_{t+1}^{LR}, \ldots, I_{t+N}^{LR}\}$, where I_t^{LR} is the reference frame and the others are the neighboring frames, our goal is to recover the corresponding high quality video frame through the reference frame and its 2N neighboring frames. Therefore, our network takes $I_{[t-N,t+N]}^{LR}$ as inputs, and finally reconstructs I_t^{SR}. The overall network structure is shown in Fig.1, which can be divided into four parts, including feature extraction module, alignment module, non-local attention module and the final SR reconstruction module.

For all the input LR frames, we first extract their features via a shared feature extraction module. It consists of one convolutional layer and several residual blocks. The feature extraction can be represented as:

$$F_T = H_{fea}\left(I_T^{LR}\right), \tag{1}$$

where the output F_T denotes the extracted LR feature maps. Then each LR neighboring feature F_i will enter the alignment module along with the LR reference feature F_t. Our alignment module which consists of stacked deformable convolutions is responsible for performing adaptive feature level alignment:

$$F_i^a = H_{align}\left(F_i, F_t\right), i \in [t-N, t+N] \text{ and } i \neq t, \tag{2}$$

where F_i^a denotes the neighboring feature after alignment. Subsequently, each aligned neighboring feature and the reference feature are fed into a non-local attention module. By calculating the global correlation between them, connections of pixels are established and informative regions in F_i^a will be further enhanced. The output $F_i^{a'}$ of the non-local attention module can be expressed as:

$$F_i^{a'} = H_{nl}\left(F_i^a, F_t\right). \tag{3}$$

The last part is the SR reconstruction module, here we use the residual in residual dense blocks (RRDB). We concatenate 2N+1 features and fuse them through a 3×3 convolution layer, then the fused feature maps are fed into RRDBs for further reconstruction. Besides, we use a skip connection to propagate LR reference feature to the end of the network and do an element-wise addition with the outcome of RRDBs. Finally, a high quality HR reference frame is recovered from the output feature. The reconstruction module is defined as

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2958030, IEEE Access

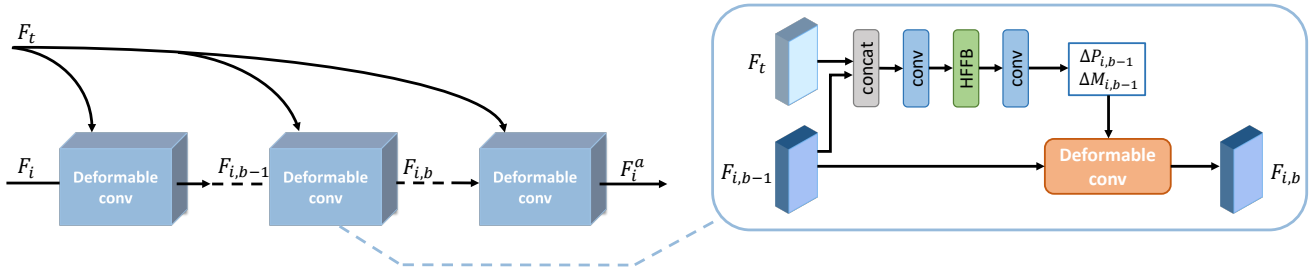
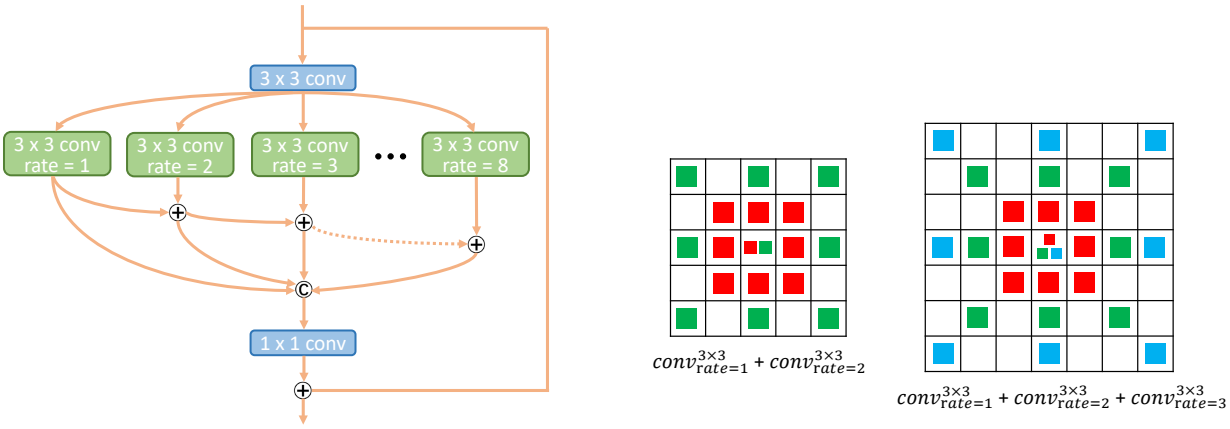


FIGURE 2. The proposed alignment module and the detailed illustration of deformable convolution operation.



(a) Hierarchical feature fusion block (HFFB)

(b) Receptive field of multiple dilated convolutions addition

FIGURE 3. (a) The structure of hierarchical feature fusion block (HFFB). It contains 8 3×3 dilated convolutions with a dilation rate from 1 to 8. The feature maps obtained using kernels of different dilation rates are hierarchically added before being concatenated. (b) A diagrammatic sketch of multiple dilated convolutions addition.

follows:

$$F_{\text{fusion}} = \text{Conv}\left(\left[F_{t-N}^{a'}, \ldots, F_{t-1}^{a'}, F_t, F_{t+1}^{a'}, \ldots, F_{t+N}^{a'}\right]\right), \tag{4}$$

$$I_t^{SR} = H_{rec}\left(H_{RRDBs}\left(F_{fusion}\right) + F_t\right), \tag{5}$$

where $[\cdot,\cdot,\cdot]$ denotes concatenation of the features. H_{rec} contains an upscaling layer and a reconstruction layer.

B. ALIGNMENT MODULE

In order to make use of temporal information from consecutive frames, traditional VSR methods are based on optical flow to perform frame alignment. However, explicit motion compensation method could be sub-optimal for video super-resolution task. We use modulated deformable convolutions [13] in the alignment module to get rid of such limitation.

For each location p on the output feature map Y, a normal convolution process can be expressed as:

$$Y(p) = \sum_{k=1}^{K} \omega_k \cdot X\left(p + p_k\right), \tag{6}$$

where p_k represents the sampling grid with K sampling locations and ω_k denotes the weights for each location. For example, $K = 9$ and $p_k \in \{(-1,-1),(-1,0),\ldots,(1,1)\}$ defines a 3×3 convolutional kernel. In the modulated deformable convolution, predicted offsets and modulation scalar are added to the sampling grid making deformable kernels spatially-variant. Here, we utilize the deformable convolution for temporal alignment. Let $F_{i,b-1}$ and $F_{i,b}$ denote the input and output of the deformable convolution in our module, respectively. The operation of modulated deformable convolution is as follows:

$$F_{i,b}(p) = \sum_{k=1}^{K} \omega_k \cdot F_{i,b-1}\left(p + p_k + \Delta p_{i,k}\right) \cdot \Delta m_{i,k}, \tag{7}$$

where $\Delta p_{i,k}$ and $\Delta m_{i,k}$ are the learnable offset and modulation scalar for the k-th location, respectively. The convolution will be operated on the irregular positions with dynamic weights to achieve adaptive sampling on input features. Since the offsets and modulation scalar are both learned, each input neighboring feature will be concatenated with the reference one to generate the corresponding deformable sampling parameters:

$$\Delta P_i, \Delta M_i = f\left(\left[F_i, F_t\right]\right), \tag{8}$$

where $[\cdot,\cdot]$ denotes the concatenation operation. And $\Delta P = \{\Delta p_k\}$, $\Delta M = \{\Delta m_k\}$. As the Δp_k may be fractional,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2958030, IEEE Access

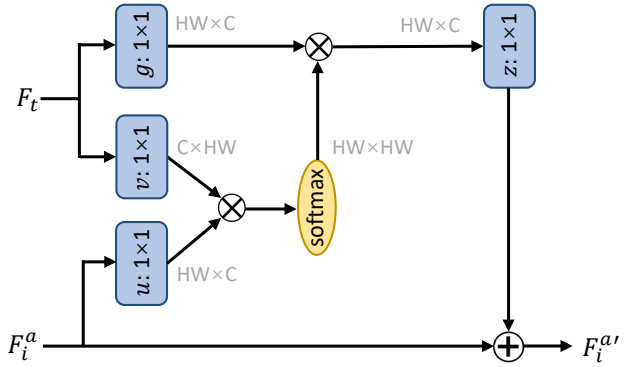


FIGURE 4. The non-local attention module.

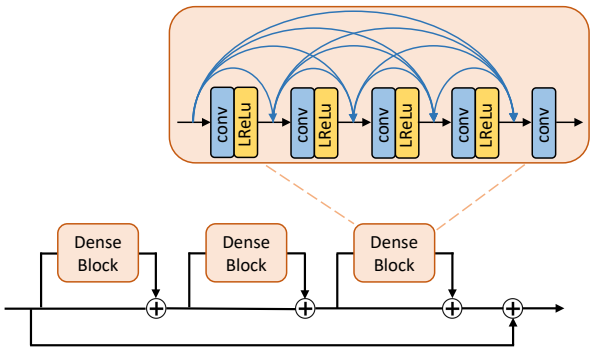


FIGURE 5. The residual in residual dense block (RRDB).

we use the bilinear interpolation, which is the same as that proposed in [36].

The alignment module proposed in DNLN is composed of several deformable convolutions as shown in Fig.2. In each deformable convolution, a reference feature F_t and a neighboring feature F_i are concatenated as an input. Then they pass through a 3×3 convolution layer to reduce channels and a hierarchical feature fusion block (HFFB) [14] to increase the size of receptive field. The following 3×3 convolution layer is used to obtain the offset ΔP_i and modulation scalar ΔM_i for the deformable kernel. The structure of HFFB is depicted in Fig.3. It introduces a spatial pyramid of dilated convolutions to effectively enlarge receptive field with relatively low computational cost, which contributes to deal with complicated and large motions between frames. In HFFB, the feature maps obtained using kernels of different dilation rates are hierarchically added before being concatenated. With the same size of receptive field, the multiple dilated convolutions addition is more dense than just one dilated convolution. The use of HFFB is beneficial to acquire an effective receptive field, so we can more efficiently exploit the temporal dependency of pixels to generate the sampling parameters.

According to Eq.(7), the deformable kernel can adaptively select sampling positions on neighboring features, learn implicit motion compensation between two frames, and complete the alignment of features. With a cascade of deformable convolutions, we can gradually align the neighboring features and improve the alignment accuracy of sub-pixels. It is noticed that when passing through a deformable convolution layer, the reference feature always keeps unchanged, only to provide a reference for the alignment of neighboring features. Through such a coarse to fine process, the neighboring frames can be well warped at the feature level.

C. NON-LOCAL ATTENTION MODULE

Due to the factors such as occlusion, blurring and parallax problems, even after the alignment module, the neighboring frames still have some areas that are not well aligned or don't contain the missing details needed for the reference frame. Therefore, it is essential to dynamically select valid inter-frame information before merging features. The proposed DNLN introduces a non-local attention module to achieve this goal. By capturing the global correlation between the aligned neighboring feature F_i^a and the reference one F_t, the non-local module can effectively enhance desirable fine details in F_i^a which can be complementary to the reference frame, and suppress the misaligned areas.

Let $\mathbf{x}$ and $\mathbf{y}$ denote the input feature F_i^a and F_t in Fig.4, respectively. The non-local operation in our module can be defined as:

$$\mathbf{z}_p = \mathbf{x}_p + W_z \sum_{n=1}^{N} \frac{f\left(\mathbf{x}_p, \mathbf{y}_n\right)}{\mathcal{C}(\mathbf{y})} \left(W_g \cdot \mathbf{y}_n\right), \tag{9}$$

where $\mathbf{z}$ denotes the module output $F_i^{a'}$. Here p is the index of an output position, and n is the index that enumerates all positions on $\mathbf{y}$. $W_g \mathbf{y}_n$ computes the expression of input $\mathbf{y}$ at position n. The function $f\left(\mathbf{x}_p, \mathbf{y}_n\right)$ calculates the relationship between $\mathbf{x}_p$ and $\mathbf{y}_n$. We use embedded Gaussian function to represent this pairwise relationship and it is normalized by a factor $\mathcal{C}(\mathbf{y})$:

$$\frac{f\left(\mathbf{x}_p, \mathbf{y}_n\right)}{\mathcal{C}(\mathbf{y})} = \frac{\exp\left(\langle W_u \mathbf{x}_p, W_v \mathbf{y}_n \rangle\right)}{\sum_n \exp\left(\langle W_u \mathbf{x}_p, W_v \mathbf{y}_n \rangle\right)}. \tag{10}$$

$W_u \mathbf{x}_p$, $W_v \mathbf{y}_n$ are used to linearly embed the input and pairwise relationship is obtained from such a softmax computation. Then we calculate a value of position p by using these relationships and the corresponding expression of all positions on $\mathbf{y}$. The value is added to the input $\mathbf{x}_p$ to get the final output $\mathbf{z}_p$. Through non-local operation, the neighboring features can make full use of the correlation with the reference feature at the pixel level and enhance the desired missing details.

D. SR RECONSTRUCTION MODULE

The output of the non-local attention module $F_i^{a'}$ is aggregated with the reference feature F_t through a feature fusion layer, and then fed into the following SR reconstruction module. The SR reconstruction module mainly consists of stacked residual in residual dense blocks (RRDB) and a

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2958030, IEEE Access

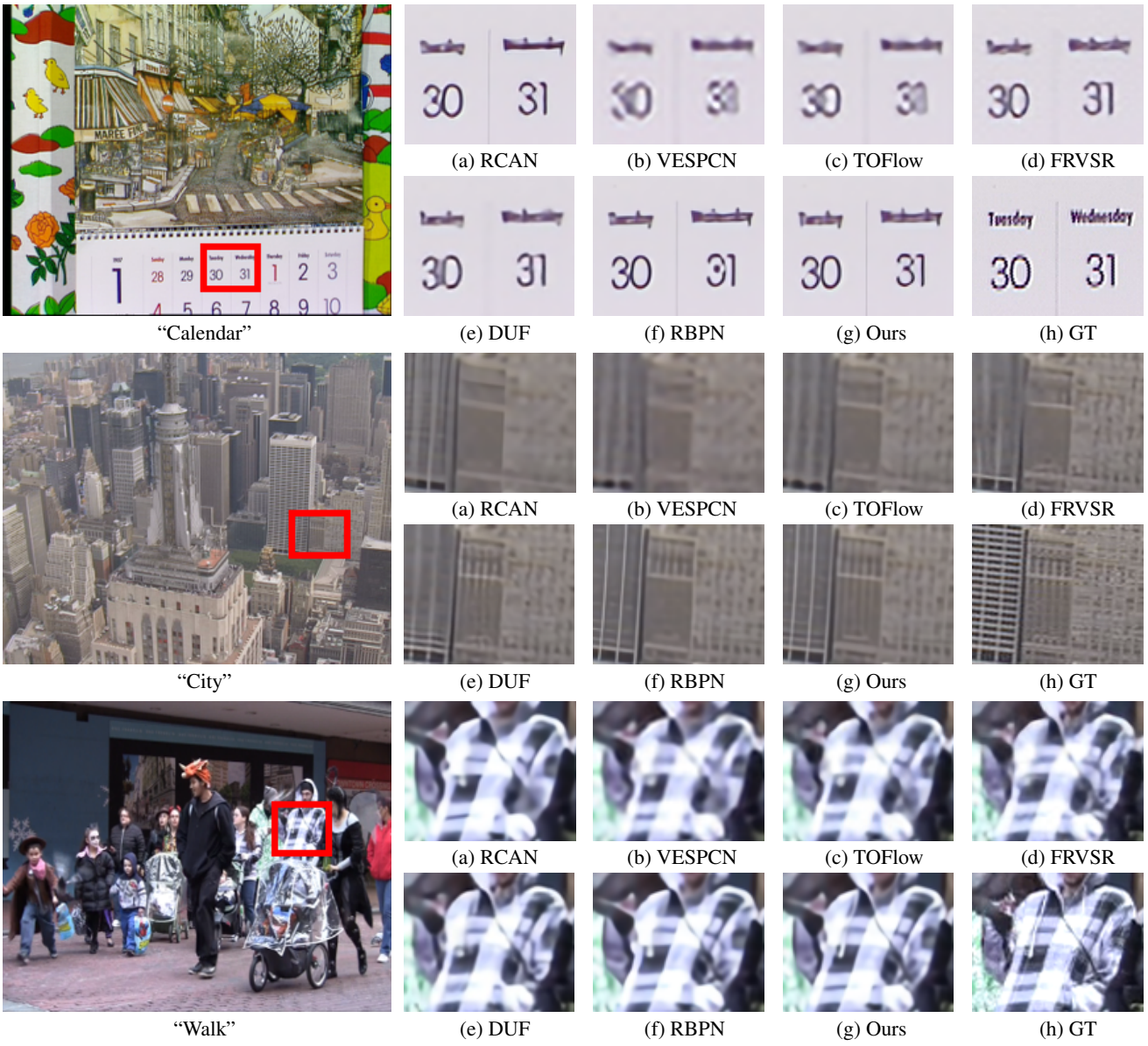


FIGURE 6. Visual results on Vid4 for $4\times$ scaling factor. Zoom in to see better visualization.

global skip connection.

The structure of RRDB can be seen in Fig.5. It combines multi-level residual network and dense connections. Benefiting from them, RRDBs can make full use of hierarchical features from input frames and obtain better restoration quality. More details about RRDB can be found in [15]. The global skip connection transfers the shallow features of the reference frame to the end of the network, making the reconstruction module focus on learning residual features from the neighboring frames. It can well keep spatial information of the input LR reference frame and ensure the input frame and the corresponding super-resolved one have more structural similarity. Finally, a high-resolution reference frame is produced by a sub-pixel upsampling layer and a reconstruction layer.

IV. EXPERIMENTS

A. TRAINING DATASETS AND DETAILS

Datasets To train high-performance VSR networks, a large video dataset is required. Xue [34] et al. collected videos from Vimeo and released a VSR dataset vimeo-90k after processing. The dataset contains 64612 training samples with various and complex real-world motions. Each sample contains seven consecutive frames with a fixed resolution of 448×256. We use the vimeo-90k dataset as our training dataset. To generate LR images, we downscale the HR images $4\times$ with MATLAB imresize function, which first blurs the input frames using cubic filters and then downsamples them using bicubic interpolation.

Training Details In our network, the convolutional layers have 64 filters and their kernel sizes are set to 3×3, if not specified otherwise. In the feature extraction module, we

TABLE 1. Quantitative comparison of state-of-the-art SR algorithms on Vid4 for $4\times$. Red indicates the best and blue indicates the second best performance (PSNR/SSIM). In the evaluation, the first and last two frames are not included and we do not crop any border pixels except DUF. Eight pixels near image boundary are cropped for DUF.

Clip Name	Flow Magnitude	Bicubic (1 Frame)	RCAN [25] (1 Frame)	VESPCN [11] (3 Frames)	TOFlow [34] (7 Frames)	FRVSR [10] (recurrent)	DUF [35] (7 Frames)	RBPN [28] (7 Frames)	DNLN(Ours) (7 Frames)
Calendar	1.14	20.39 / 0.5720	22.31 / 0.7248	-	22.44 / 0.7290	-	24.07 / 0.8123	23.95 / 0.8076	24.12 / 0.8141
City	1.63	25.17 / 0.6024	26.07 / 0.6938	-	26.75 / 0.7368	-	28.32 / 0.8333	27.74 / 0.8051	27.90 / 0.8111
Foliage	1.48	23.47 / 0.5666	24.69 / 0.6628	-	25.24 / 0.7065	-	26.41 / 0.7713	26.21 / 0.7578	26.28 / 0.7607
Walk	1.44	26.11 / 0.7977	28.64 / 0.8718	-	29.03 / 0.8777	-	30.63 / 0.9144	30.70 / 0.9111	30.85 / 0.9129
Average	1.42	23.79 / 0.6347	25.43 / 0.7383	25.35 / 0.7557	25.86 / 0.7625	26.69 / 0.822	27.36 / 0.8328	27.15 / 0.8204	27.29 / 0.8247

TABLE 2. Quantitative comparison of state-of-the-art SR algorithms on SPMCS-11 for $4\times$.

Clip Name	Flow Magnitude	Bicubic (1 Frame)	RCAN [25] (1 Frame)	TOFlow [34] (7 Frames)	DUF [35] (7 Frames)	RBPN [28] (7 Frames)	DNLN(Ours) (7 Frames)
car05_001	6.21	27.75 / 0.7825	29.86 / 0.8484	30.10 / 0.8626	30.79 / 0.8707	31.95 / 0.9021	31.96 / 0.9011
hdclub_003_001	0.70	19.42 / 0.4863	20.41 / 0.6096	20.86 / 0.6523	22.05 / 0.7438	21.91 / 0.7257	22.15 / 0.7366
hitachi_isee5_001	3.01	19.61 / 0.5938	23.71 / 0.8369	22.88 / 0.8044	25.77 / 0.8929	26.30 / 0.9049	26.60 / 0.9080
hk004_001	0.49	28.54 / 0.8003	31.68 / 0.8631	30.89 / 0.8654	32.98 / 0.8988	33.38 / 0.9016	33.46 / 0.9041
HKVTG_004	0.11	27.46 / 0.6831	28.81 / 0.7649	28.49 / 0.7487	29.16 / 0.7860	29.51 / 0.7979	29.53 / 0.7976
jvc_009_001	1.24	25.40 / 0.7558	28.31 / 0.8717	27.85 / 0.8542	29.18 / 0.8961	30.06 / 0.9105	30.65 / 0.9205
NYVTG_006	0.10	28.45 / 0.8014	31.01 / 0.8859	30.12 / 0.8603	32.30 / 0.9090	33.22 / 0.9231	33.35 / 0.9254
PRVTG_012	0.12	25.63 / 0.7136	26.56 / 0.7806	26.62 / 0.7788	27.39 / 0.8166	27.60 / 0.8242	27.68 / 0.8260
RMVTG_011	0.18	23.96 / 0.6573	26.02 / 0.7569	25.89 / 0.7500	27.56 / 0.8113	27.63 / 0.8170	27.75 / 0.8199
veni3_011	0.36	29.47 / 0.8979	34.58 / 0.9629	32.85 / 0.9536	34.63 / 0.9677	36.61 / 0.9735	36.33 / 0.9739
veni5_015	0.36	27.41 / 0.8483	31.04 / 0.9262	30.03 / 0.9118	31.88 / 0.9371	32.37 / 0.9409	33.04 / 0.9466
Average	1.17	25.73 / 0.7291	28.36 / 0.8279	27.87 / 0.8220	29.43 / 0.8664	30.05 / 0.8747	30.23 / 0.8782

TABLE 3. Quantitative comparison of state-of-the-art SR algorithms on Vimeo-90K-T for $4\times$.

Method	Slow	Medium	Fast	Average
Bicubic	29.34 / 0.8330	31.29 / 0.8708	34.07 / 0.9050	31.32 / 0.8684
RCAN [25]	32.93 / 0.9032	35.35 / 0.9268	38.47 / 0.9456	35.34 / 0.9249
TOFlow [34]	32.15 / 0.8900	35.01 / 0.9254	37.70 / 0.9430	34.84 / 0.9209
DUF [35]	33.41 / 0.9110	36.71 / 0.9446	38.87 / 0.9510	36.37 / 0.9386
RBPN [28]	34.26 / 0.9222	37.39 / 0.9494	40.16 / 0.9611	37.18 / 0.9456
DNLN(ours)	34.47 / 0.9246	37.59 / 0.9510	40.35 / 0.9621	37.38 / 0.9473
# of clips	1616	4983	1225	7824
Flow Mag.	0.6	2.5	8.3	3.0

utilize 5 residual blocks to extract shallow features. Then the alignment module adopts 5 deformable convolutions to perform feature alignment. The dilated convolutions in HFFB have 3×3 kernels and 32 filters. In the non-local attention module, the first three 1×1 convolutions have 32 filters and the last 1×1 convolution has 64 filters. Finally, in the reconstruction module, we use 23 RRDBs and set the number of growth channels to 32.

In the training process, we perform data augmentation by doing horizontal or vertical flipping, 90° rotation and random cropping of the images. The batch size is set to 8. The network takes seven consecutive frames as inputs, and LR patches with the size of 50×50 are extracted for training. Our model is trained by Adam optimizer [39] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. The initial learning rate is 10^{-4} before 70 epochs and later decreases to half every 20 epochs. All experiments were conducted on two NVIDIA RTX 2080 GPUs using PyTorch 1.0 [40]. We train the network end-to-end by minimizing L1 loss between the predicted frame and the ground truth HR frame. And later we employ the L2 loss to finetune the model, which could result in better performance.

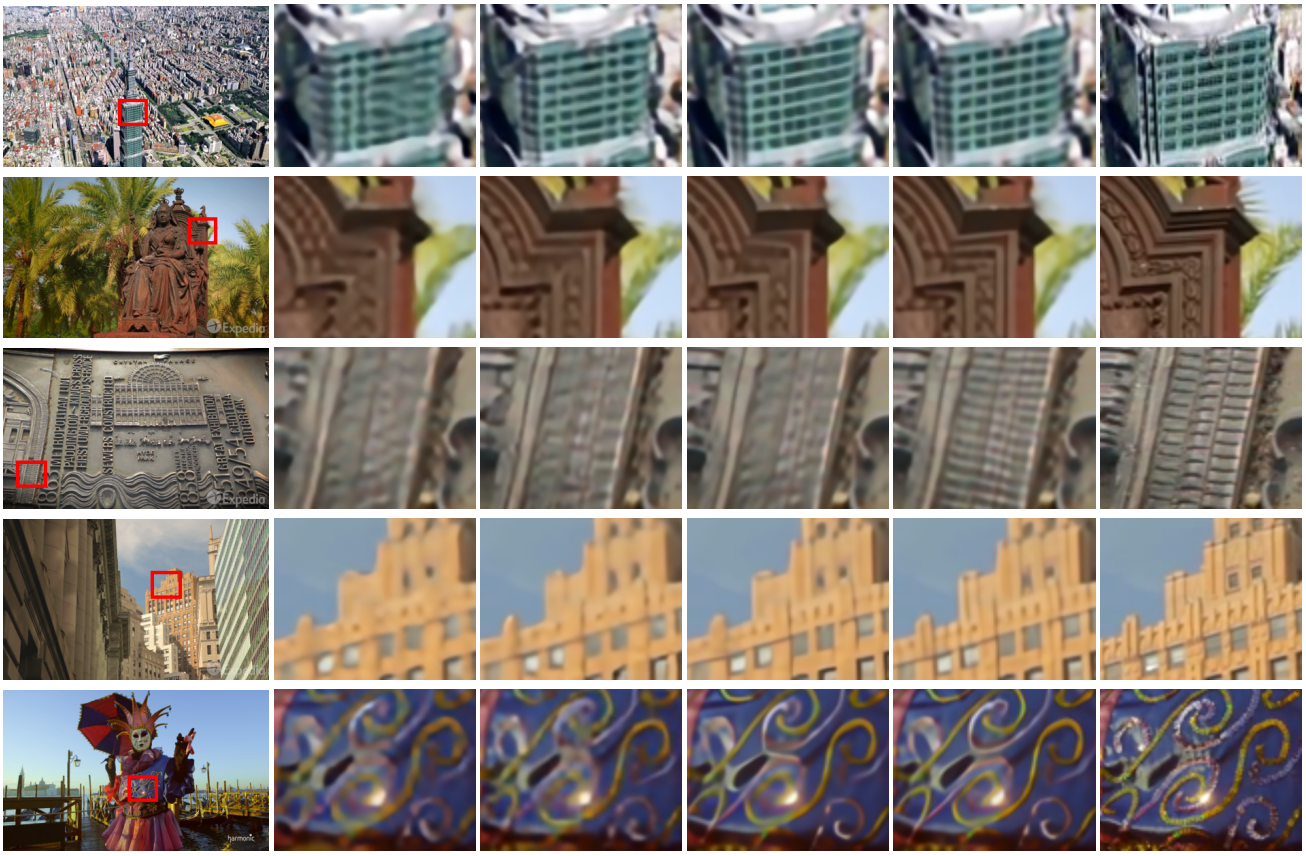
B. COMPARISON WITH THE STATE-OF-THE-ART METHODS

We compare our DNLN with several state-of-the-art SISR and VSR methods: RCAN [25], VESPCN [11], TOFlow [34], FRVSR [10], DUF [35] and RBPN [28]. Note that most previous methods are trained with different datasets and we just compare with the results they provided. The SR results are evaluated with PSNR and SSIM [41] quantitatively on Y channel (i.e., luminance) of transformed YCbCr space. In the evaluation, the first and last two frames are not included and we do not crop any border pixels except DUF [35]. Eight pixels near image boundary are cropped for DUF due to its severe boundary effects.

We evaluated our models on three datasets: Vid4 [5], SPMCS [9], and Vimeo-90K-T [34] with average flow magnitude (pixel/frame) provided in [28]. Vid4 is a commonly used dataset which contains four video sequences: calendar, city, foliage and walk. However, we can observe that Vid4 has limited inter-frame motion and there exists artifacts on its ground-truth frames. SPMCS consists of higher quality video clips with various motions and diverse scenes. Vimeo-90K-T is a much larger dataset. It contains a wide range of flow magnitude between frames which can well judge the performance of the VSR methods.

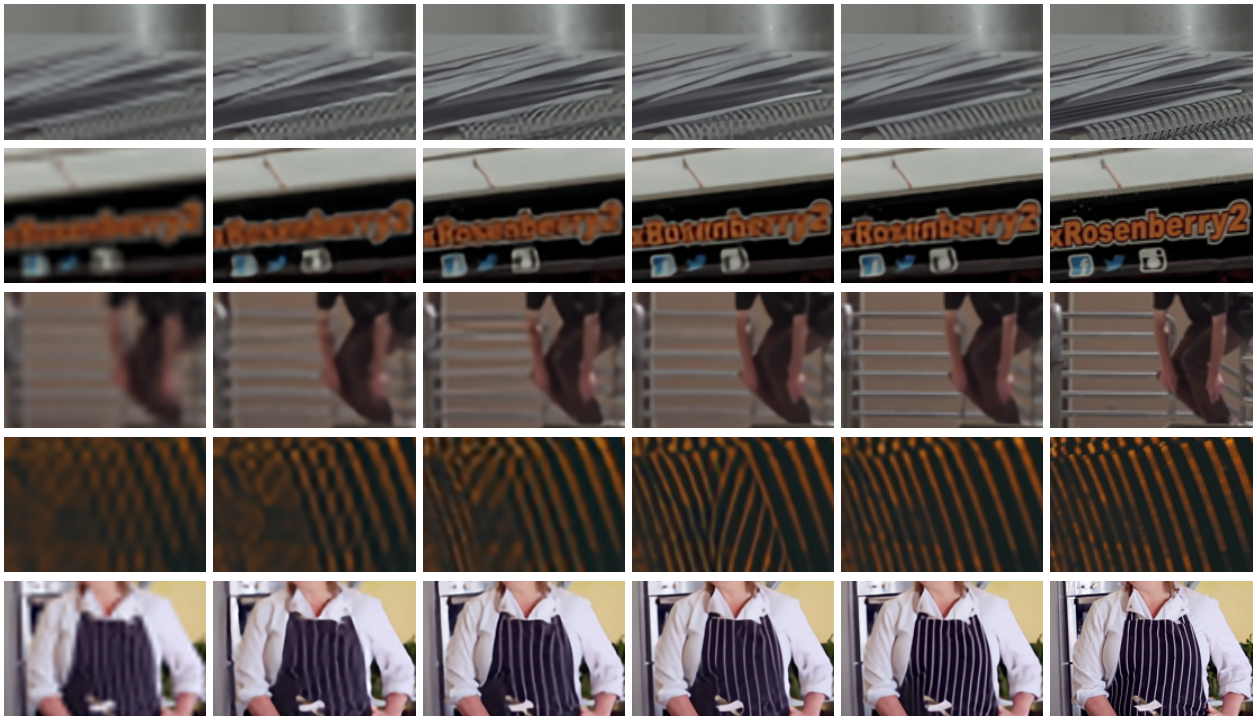
Table 1 shows the quantitative results on Vid4. Our model outperforms the optical-flow-based methods which demonstrates the effectiveness of our optical flow free alignment module. Qualitative results are shown in Fig.6. We mark out the positions which display obvious distinctions among different methods. For the date in "Calendar" clip, most compared methods produce images with blurring artifacts, while our method achieves a better result and alleviates the artifacts. In the "Walk" clip, the existing methods blur the rope and clothing together, only DNLN can clearly distinguish these two parts and restore the pattern closest to the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2958030, IEEE Access



(a) TOFlow (b) DUF (c) RBPN (d) Ours (e) GT

FIGURE 7. Visual results on SPMCS for 4× scaling factor. Zoom in to see better visualization.



(a) Bicubic (b) TOFlow (c) DUF (d) RBPN (e) Ours (f) GT

FIGURE 8. Visual results on Vimeo-90K-T for 4× scaling factor. Zoom in to see better visualization.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2958030, IEEE Access

(a) 1dconv (b) 2dconv (c) 3dconv (d) 4dconv (e) 5dconv (f) w/o HFFB (g) GT

FIGURE 9. Qualitative results of ablation on alignment module.

ground truth frame.

In comparison to Vid4, SPMCS contains more high frequency information with higher resolution, which requires the superb recovery abilities of algorithms. Results on SPMCS are shown in Table 2. DNLN achieves the best results and outperforms other methods by a large margin on PSNR, which proves the superiority of our model. Visual comparisons are depicted in Fig.7. Due to the abundance of textures, most methods cannot fully recover the frames and obviously produce blurring artifacts. Although DUF and RBPN could reproduce part of the HR patterns, it is obvious that our DNLN is the unique approach to restore the abundant details and clean edges. Such visual comparisons demonstrate that our network can extract more sophisticated features from LR space with the proposed modules.

Table 3 presents the quantitative outcomes of Vimeo-90K-T. As suggested in [28], we classified the video clips into three groups (e.g. slow, medium and fast) according to the motion velocity. While the motion velocity increases, video frames with larger motion amplitude will contain more useful temporal information but also make the recovery more challenging. Our DNLN ensures optimal performance on all three groups, surpassing RBPN by 0.21 dB, 0.20 dB and 0.19 dB on PSNR, respectively. Since the flow magnitude of Vimeo-90K-T in fast group is higher than Vid4 and SPMCS, the content between video frames varies greatly, which reflects that DNLN could take full advantage of temporal information among multiple frames. The qualitative evaluations are shown in Fig.8. For the railing texture in third row, only our method restores the correct and clear pattern while others suffer from varying degrees of blurring. In some cases, even the SR frames recovered by different methods have the same sharp edges, our DNLN is more accurate and faithful to the ground truth. Such as the images in fourth row, the results restored by RBPN and DNLN are equally clear, while the former produces the stripes with wrong directions.

C. MODEL SIZE AND RUNNING TIME ANALYSES

Table 4 shows comparisons about model size and and running time of the methods. The running time is test with input size 112×64. Our DNLN has the largest model size but also achieves the best performance. Here, we adopt a smaller model S-DNLN with only 3 deformable convolutions in alignment module and 14 RRDBs in reconstruction module. We can see that S-DNLN has a comparable number of parameters and running time with RBPN. Nevertheless, it can still get a better result which further validates the effectiveness of our proposed modules.

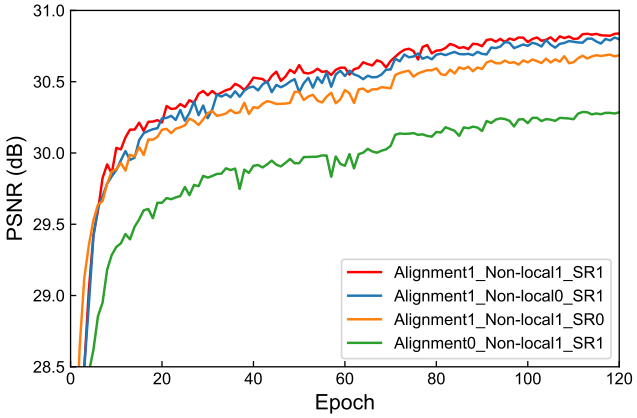


FIGURE 10. Convergence analysis on the three modules of proposed network. The curves for each combination are based on the PSNR with scaling factor $4\times$.

TABLE 4. Number of parameters and time cost on Vimeo-90K-T for $4\times$.

Methods	RCAN	TOFlow	DUF	RBPN	S-DNLN	DNLN
Parameter (M)	15.59	1.41	5.82	12.77	12.39	19.74
Time (s)	0.051	0.126	0.175	0.086	0.095	0.119
PSNR	35.34	34.84	36.37	37.18	37.23	37.38

V. ABLATION STUDY

To further investigate the proposed method, we conducted ablation experiments by removing the main components of our network. The results are shown in Table 5. First, we remove the alignment module, thus the shallow features would be directly fed into the following network without warping. The PSNR of the results on Vimeo-90K-T is relatively low, which indicates that the alignment operation is crucial for utilizing the inter-frame information. Second, we remove the non-local attention module and the performance decreases a lot. Third, we replace the RRDBs by simply stacking common residual blocks and it also harms the performance. We visualize the convergence process of these combinations in Fig.10. The results demonstrate the effectiveness and benefits of our proposed three modules.

From the ablation experiments above, we can observe that the network performance would be significantly affected by the alignment preprocessing. So we further validated the impact of deformable convolutions on the reconstruction capability. As shown in Table 6, with only one deformable convolution, the PSNR value can improve greatly. It demonstrates the importance of alignment operations for making efficient use of the neighboring frames. As the number of

TABLE 5. Ablation study of proposed network on Vimeo-90K-T for 4×.

Alignment module	Non-local module	SR module	PSNR	SSIM
	✓	✓	36.82	0.9418
✓		✓	37.34	0.9471
✓	✓		37.25	0.9462
✓	✓	✓	37.38	0.9473

TABLE 6. Ablation study on alignment module.

Model	PSNR	SSIM
w/o deform	36.81	0.9417
1dconv	37.09	0.9446
2dconv	37.19	0.9455
3dconv	37.33	0.9469
4dconv	37.36	0.9471
5dconv	37.38	0.9473
5dconv, w/o HFFB	37.09	0.9447

deformable convolutions increases, the network gains a better performance. The visual comparisons are shown in Fig.9. From left to right, the network alleviates the blurring artifacts of the office building and recovers more accurate details. In addition, we replaced the HFFB used in deformable convolutions with a 3×3 convolution layer. The performance of network decreases by roughly 0.29 dB. It proves that by enlarging the receptive field, the deformable convolution can more effectively cope with complex and large motions.

In order to study the influence of inter-frame information on the recovery results, we leveraged different number of frames to train our network. From Table 7, we can observe that there is a significant improvement in DNLN when switching from 3 frames to 5 frames, and the performance of DNLN/5 is even better than RBPN which uses 7 frames. When further switching to 7 frames, we can still get a better result but the improvement becomes minor.

VI. CONCLUSION

In this paper, we propose a novel deformable non-local network (DNLN), which is a non-flow-based method for effective video super-resolution. To deal with complicated and large motion compensation, we introduce the deformable convolution with HFFB in our alignment module, which can well align the frames at the feature level. In addition, we adopt a non-local attention module to further extract complementary features from neighboring frames. By making full use of the temporal information, we finally restore a high quality video frame through a reconstruction module. Extensive experiments on benchmark datasets illustrate the effectiveness of our DNLN in video super-resolution.

REFERENCES

[1] H. Seibel, S. Goldenstein, and A. Rocha, "Eyes on the target: Super-resolution and license-plate recognition in low-quality surveillance videos," IEEE access, vol. 5, pp. 20 020–20 035, 2017.

TABLE 7. Experimental results with a different number of input frames.

Input	3 frames	5 frames	7 frames
PSNR / SSIM	37.06 / 0.9435	37.29 / 0.9463	37.38 / 0.9473

[2] J. S. Park, J. W. Soh, and N. I. Cho, "High dynamic range and super-resolution imaging from a single image," IEEE Access, vol. 6, pp. 10 966–10 978, 2018.

[3] S. C. Park, M. K. Park, and M. G. Kang, "Super-resolution image reconstruction: a technical overview," IEEE signal processing magazine, vol. 20, no. 3, pp. 21–36, 2003.

[4] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super resolution," IEEE transactions on image processing, vol. 13, no. 10, pp. 1327–1344, 2004.

[5] C. Liu and D. Sun, "On bayesian adaptive video super resolution," IEEE transactions on pattern analysis and machine intelligence, vol. 36, no. 2, pp. 346–360, 2013.

[6] Z. Ma, R. Liao, X. Tao, L. Xu, J. Jia, and E. Wu, "Handling motion blur in multi-frame super-resolution," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5224–5232.

[7] W. Wang, C. Ren, X. He, H. Chen, and L. Qing, "Video super-resolution via residual learning," IEEE Access, vol. 6, pp. 23 767–23 777, 2018.

[8] D. Liu, Z. Wang, Y. Fan, X. Liu, Z. Wang, S. Chang, and T. Huang, "Robust video super-resolution with learned temporal dynamics," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2507–2515.

[9] X. Tao, H. Gao, R. Liao, J. Wang, and J. Jia, "Detail-revealing deep video super-resolution," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4472–4480.

[10] M. S. Sajjadi, R. Vemulapalli, and M. Brown, "Frame-recurrent video super-resolution," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6626–6634.

[11] J. Caballero, C. Ledig, A. Aitken, A. Acosta, J. Totz, Z. Wang, and W. Shi, "Real-time video super-resolution with spatio-temporal networks and motion compensation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4778–4787.

[12] Y. Tian, Y. Zhang, Y. Fu, and C. Xu, "Tdan: Temporally deformable alignment network for video super-resolution," arXiv preprint arXiv:1812.02898, 2018.

[13] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable convnets v2: More deformable, better results," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 9308–9316.

[14] Z. Hui, J. Li, X. Gao, and X. Wang, "Progressive perception-oriented network for single image super-resolution," arXiv preprint arXiv:1907.10399, 2019.

[15] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 0–0.

[16] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in European conference on computer vision. Springer, 2014, pp. 184–199.

[17] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1646–1654.

[18] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in Proceedings of the IEEE Conference on Computer vision and Pattern Recognition, 2017, pp. 3147–3155.

[19] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in European conference on computer vision. Springer, 2016, pp. 391–407.

[20] W. Shi, J. Caballero, F. Huszár, 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 Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1874–1883.

[21] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, and L. Zhang, "Ntire 2017 challenge on single image super-resolution: Methods and results," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 114–125.

[22] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2017, pp. 136–144.

[23] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 1664–1673.

[24] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2472–2481.
[25] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 286–301.
[26] R. Liao, X. Tao, R. Li, Z. Ma, and J. Jia, "Video super-resolution via deep draft-ensemble learning," in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 531–539.
[27] A. Kappeler, S. Yoo, Q. Dai, and A. K. Katsaggelos, "Video super-resolution with convolutional neural networks," IEEE Transactions on Computational Imaging, vol. 2, no. 2, pp. 109–122, 2016.
[28] M. Haris, G. Shakhnarovich, and N. Ukita, "Recurrent back-projection network for video super-resolution," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3897–3906.
[29] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 2758–2766.
[30] A. Ranjan and M. J. Black, "Optical flow estimation using a spatial pyramid network," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4161–4170.
[31] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2462–2470.
[32] T.-W. Hui, X. Tang, and C. Change Loy, "Liteflownet: A lightweight convolutional neural network for optical flow estimation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8981–8989.
[33] M. Zhai, X. Xiang, N. Lv, S. M. Ali, and A. El Saddik, "Skflow: Optical flow estimation using selective kernel networks," IEEE Access, vol. 7, pp. 98 854–98 865, 2019.
[34] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, "Video enhancement with task-oriented flow," International Journal of Computer Vision, pp. 1–20, 2017.
[35] Y. Jo, S. Wug Oh, J. Kang, and S. Joo Kim, "Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3224–3232.
[36] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 764–773.
[37] X. Wang, K. C. Chan, K. Yu, C. Dong, and C. Change Loy, "Edvr: Video restoration with enhanced deformable convolutional networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 0–0.
[38] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803.
[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.
[40] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.
[41] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli et al., "Image quality assessment: from error visibility to structural similarity," IEEE transactions on image processing, vol. 13, no. 4, pp. 600–612, 2004.