Signal Processing: Image Communication 93 (2021) 116167

Contents lists available at ScienceDirect

Signal Processing: Image Communication

journal homepage: www.elsevier.com/locate/image

Real-time video super resolution network using recurrent multi-branch dilated convolutions

Yubin Zeng [a,1], Zhijiao Xiao [a,1], Kwok-Wai Hung [b,*], Simon Lui [b]

[a] College of Computer Science and Software Engineering, Shenzhen University, China
[b] Tencent Music Entertainment, Shenzhen, China

ABSTRACT

Recent developments of video super-resolution reconstruction often exploit spatial and temporal contexts from input frame sequence by making use of explicit motion estimation, e.g., optical flow, which may introduce accumulated errors and requires huge computations to obtain an accurate estimation. In this paper, we propose a novel multi-branch dilated convolution module for real-time frame alignment without explicit motion estimation, which is incorporated with the depthwise separable up-sampling module to formulate a sophisticated real-time video super-resolution network. Specifically, the proposed video super-resolution framework can efficiently acquire a larger receptive field and learn spatial–temporal features of multiple scales with minimal computational operations and memory requirements. Extensive experiments show that the proposed super-resolution network outperforms current state-of-the-art real-time video super-resolution networks, e.g., VESPCN and 3DVSRnet, in terms of PSNR values (0.49 dB and 0.17 dB) on average in various datasets, but requires less multiplication operations.

1. Introduction

As one of the fundamental problems in signal processing and computer vision, video or multi-frame super-resolution (VSR) focuses on reconstructing a high resolution (HR) image by utilizing information from a sequence of low resolution (LR) multi-frame inputs. This problem is drawing more and more attentions in recent academic research and industrial applications. From the industry aspect, there is a growing need for generating a visually pleasing high resolution video from a given low resolution video clip, such as HDTV [1], surveillance [2] and satellite imaging [3]. In research perspective, processing the spatial–temporal signal in a consecutive image sequence is a challenging problem.

With the benefit of deep learning, single image super-resolution (SISR) has been intensively studied in recent years [4–9]. While SISR relies on internal spatial information in a single LR input image and prior knowledge learned in training samples, an ideal video super-resolution algorithm should be able to make use of temporal redundancy of multiple input frames to further improve the reconstruction. The challenge for video super-resolution is how to exploit and fuse redundancy in multiple input frames to generate better reconstructions.

Early multi-frame video super-resolution approaches generally model the temporal information by simply concatenating all consecutive frames and up-sampling using a convolutional neural network [10,11]. Some approaches generates high resolution images without explicitly alignment [11–13]. Other approaches align frames explicitly with motion compensation module and then concatenate all aligned frames as input to a multi-frame super-resolution network [14–17].

Recently proposed approaches [12,16,18] introduce RNN architecture to address temporal coherence in video super-resolution, in which all frames are fed into a RNN architecture in temporal order. In [16], each SR reconstruction is generated with consideration of previous timestep SR estimation. The alignment can be processed either explicitly by a trainable CNN motion compensation module [16], or implicitly [12,18] by passing high dimension feature maps through RNN architecture.

However, these aforementioned VSR methods have various drawbacks. The concatenation approaches [11,14] process all consecutive inputs simultaneously and each reconstruction is generated independently without considering temporal correlation of LR frame sequence, leaving the network difficult to exploit temporal information or generate temporally consistent results. Approaches [15,16] with explicit motion compensation module generate optical flow and align neighbor frames at the cost of extra computational consumptions. In RNN approaches, vanilla RNN architecture cannot address long term dependency issue. Even with LSTM-like mechanism, networks cannot efficiently capture both subtle and significant changes (i.e. fast and slow object motions) which simultaneously exist in the same video clip.

Moreover, most of the current state-of-the-art VSR methods achieve superior performance with extremely computational expensive network structures, which are not applicable in industry perspective.

To address drawbacks of previous approaches, we propose an efficient recurrent video super-resolution network, namely **ERVSR**. Our proposed network achieves state-of-the-art performance under comparatively lower computational cost without explicit motion estimation to

* Corresponding author.
E-mail address: guoweihung@tencent.com (K.-W. Hung).
[1] Contribute equally in this work.

https://doi.org/10.1016/j.image.2021.116167
Received 23 May 2020; Received in revised form 30 November 2020; Accepted 25 January 2021
Available online 6 February 2021

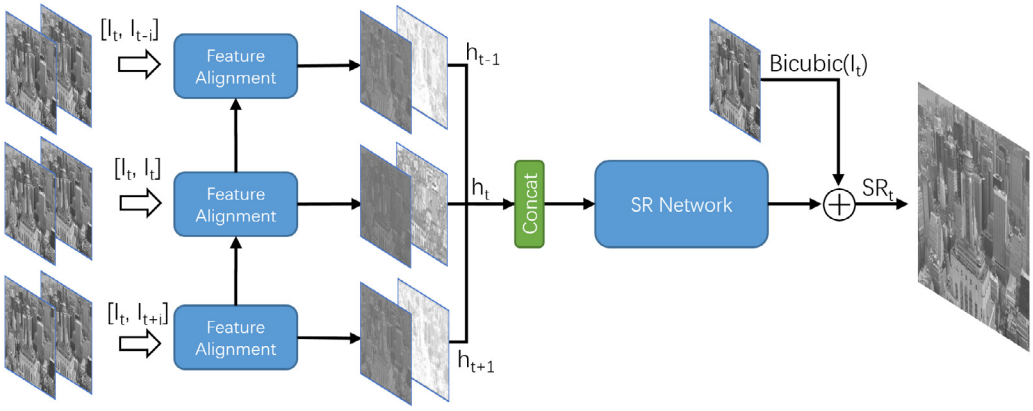

Fig. 1. Architecture of our proposed video super resolution network, where I denotes input LR images, h are featuremaps from feature alignment block and subscripts $t-1, t$ *and* $t+1$ are temporal position. Note that intensities of feature maps are normalized for better visual effects.

interpret spatial–temporal information in image sequence, in order to generate state-of-the-art image quality.

The main contributions of this paper are:

- In **ERVSR**, we adopt a multi-branch dilated module, namely MDB that can effectively improve receptive field of the network by extracting spatial–temporal features at different scales in parallel, resulting in superior performance with minimal computational costs.
- We introduce a new recurrent architecture that can process a consecutive multi-frame sequence to extract temporal and spatial features simultaneously. Instead of explicitly generating optical flow or aligning neighbor frames to the reference one, our proposed network extracts and concatenates the aligned feature maps with respect to the reference frame for the super-resolution reconstruction.
- The proposed **ERVSR** network is based on depthwise separable convolutions, in order to reduce the overall computational complexity, so that the proposed network can reconstruct high definition video clip up to 50 frames per second when inferencing on a single NVIDIA 980Ti GPU.

Overall, our model achieves the best performance among state-of-the-art real-time video super-resolution networks which require $1.5\times$ –$5\times$ more floating-point operations, in terms of objective and subjective evaluations on various standard datasets, such as Vid4, SPMC, etc.

2. Related works

2.1. Deep-learning image super resolution

SISR has been extensively studied in recent years. Dong et al. [4] proposed SRCNN, which is the first fully convolutional neural network resolving super-resolution task and achieves state-of-the-art result. Following the pioneer work of SRCNN, many other researchers have proposed deep convolutional architectures advancing the field tremendously [5–9]. Meanwhile, some researches propose alternative loss functions to achieve better visually pleasing result [19]. Current state-of-the-art approaches [9,20] introduce generative adversarial network to generate the HR estimate. However, single image super-resolution methods can only aggregate information within a single image which contains very limited information for restoration.

2.2. Deep-learning video super-resolution with feedforward CNNs

Video and multi-frame super-resolution approaches interpret information from multiple low-resolution images and recover detail and texture that are missing in individual frames, which leads to a superior

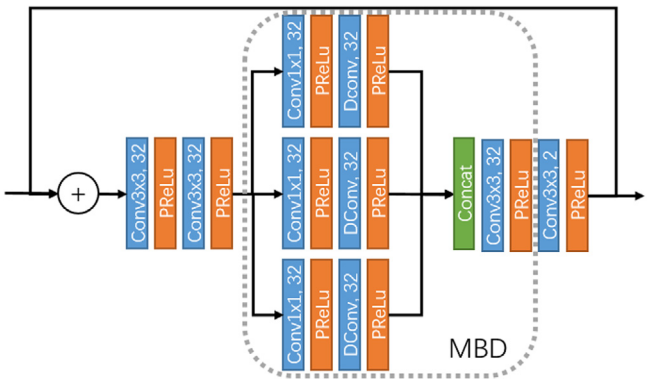

Fig. 2. Illustration of the feature alignment block. Kernel size and number of filters are demonstrated above.

reconstruction compared to SISR approaches. Most of the video and multiple frames super resolution approaches can be divided into two separate tasks: alignment and reconstruction. As the pioneer work resolving the VSR task, Kappeler et al. [14] propose VSRnet, which aligns all frames to the reference one by explicitly computing optical flow [21] and then concatenates all warped frames and feeds into an image super-resolution network. Caballero et al. [15] propose a CNN architecture to estimate optical flow of the reference frame with respect to each neighbor frame, resulting in faster inference runtime and better performance. Tao et al. [17] and Liu et al. [22] focus on designing a deeper and superior architecture to acquire a more precise alignment for better HR reconstruction. Instead of estimating optical flow, TDAN [23] and EDVR [24] align frames by deformable convolutions with estimated offsets. However, performance of alignment-based methods relies on the accuracy of alignment operation. Jo et al. [25] learn to generate a set of upsampling bilateral filters by simultaneously processing all input frames and reference frame is up-sampled by the bilateral filters. It avoids frame alignment operations, but it is difficult to learn inter-frame relation when processing all frames together.

2.3. Deep-learning video super-resolution with RNN architecture

Recently proposed VSR networks introduce RNN architecture to process sequential inputs by modeling temporal–spatial information. Huang et al. [12] propose a bidirectional convolutional RNN to model long-term contextual information of temporal sequence. Sajjadi et al. [16] adopt recurrent framework to propagate information from previous HR estimated frames for generating the subsequent frame. Haris et al. [18] integrate the spatial and temporal contexts from concurrent

video frames by using recurrent encoder–decoder module, in which embedded neighbor frame feature maps and recurrent feature maps are fused and passed through a traditional SISR architecture. However, their method is computational expensive, which is not applicable in real-time applications.

3. Methodology

In this section, we first present an overview of our network architecture in Section 3.1. In Section 3.2, we analyze the feature alignment block. In Section 3.3, we will further analyze the multi-branch dilation module. In Section 3.4, we explain our super-resolution network in details. In Section 3.5, we will present the implementation and training details.

3.1. Network architecture

The overview of ERVSR is illustrated in Fig. 1, which can be divided into two parts, i.e., the feature alignment block and the SR network. As shown in Fig. 1, at temporal position t, all input frames I_{t-1}, I_t, I_{t+1} are concatenated with the reference frame I_t. The feature alignment block is responsible for generating aligned feature maps, h_{t-1}, h_t *and* h_{t+1}, with respect to each input. The SR network block fuses all concatenated feature maps and processes spatial up-sampling.

At the rear of the network, we add the bicubic upscaled reference frame to the output of the SR network as global residual connection, so that our network can focus on refining high frequency details and textures efficiently.

3.2. Feature alignment block

The feature alignment block is illustrated in details as shown in Fig. 2. First, we concatenate all frames I_i with the reference one I_t respectively. Then all concatenated frame pairs $[I_t, I_i]$ are passed through the feature alignment block. By visualizing the feature map generated by the feature alignment block in Fig. 3, we can observe that all the feature maps are spatially aligned. And unlike ordinary motion compensation module generating a single channel image, our feature alignment module remains numbers of feature map channel rather than compressing channel-wise information. Therefore, in each recurrent iteration, the feature alignment block can effectively extract feature maps containing rich information from input frames and reused the extracted feature maps in the following procedure instead of processing feature extraction redundantly. Feature map visualization is demonstrated in Fig. 3.

The feature alignment block can be formulated as below

$$h_i = FA\left(\left[I_i, I_t\right] + h_{i-1}\right), \tag{1}$$

where $i \in \{t-n, \ldots, t, \ldots, t+n\}$

Different from previous RNN approaches, each of recurrent outputs in our network is not only taken as input to next recurrent iteration, but also preserved as input to the following part of our network. Feature maps passing through recurrent block consist of multiple convolutional layers could possibly suffer from unexpected information lost. Therefore, by preserving the feature maps corresponding to each input frames (in each recurrent iteration), we can maintain and make use of as much temporal information as possible in deep recurrent architecture (refer to Fig. 1). With such modification of standard recurrent network, our proposed **ERVSR** effectively addresses the long-term dependency issue and can avoid gradient explosions in training session, since each recurrent iteration is supervised by its output state.

Following the recurrent block, all the feature maps are concatenated and sent to the SR network for spatial up-sampling.

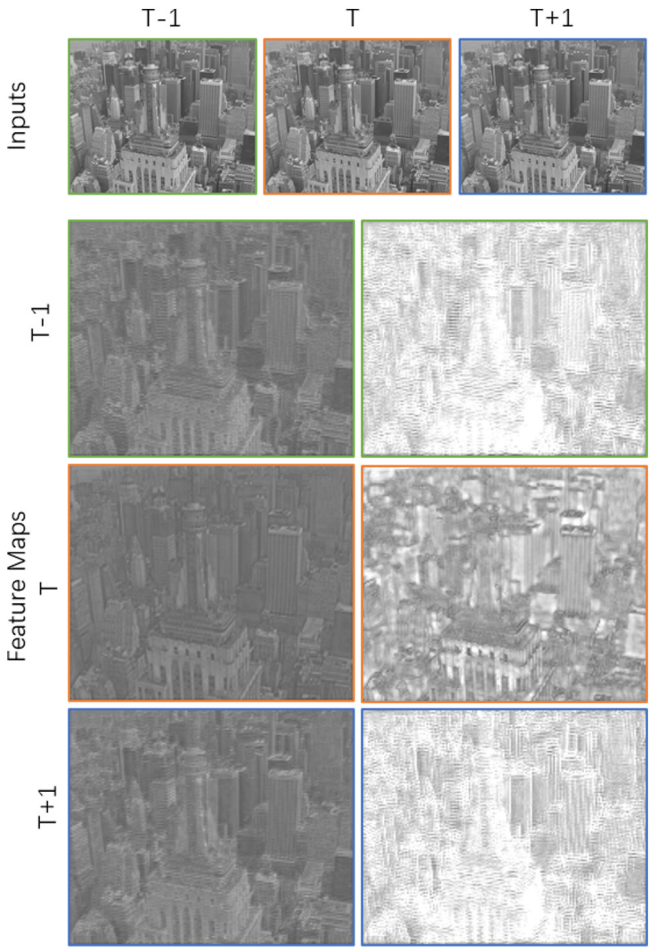

Fig. 3. Feature map visualization of recurrent block. Note that intensities of feature maps are finetuned for better visual effect.

3.3. Multi branch dilation module

Receptive field is a critical property of a convolutional neural network and has been intensively studied [26,27]. The size of receptive field has significant influence on performance of tasks such as object detection, classification and super-resolution etc. Dilated convolution [26] modifies standard convolutional layer by inserting defined gaps between kernel elements and achieves broader receptive view with less parameters. However, sampling pixels discretely with dilated convolution is not suitable for processing fine-details tasks. Moreover, empirically improving receptive field by simply stacking up convolutional layers cannot achieve a superior performance and may cause unnecessary computational consumptions.

On the other hand, there are both subtle and significant motions in video clips, leaving ordinary convolutional operations difficult to extract features of different spatial variations. Previous researches [15, 17] settle this problem by precisely align each frame to the reference one. However, the precisely aligned low resolution neighbor frames are unnecessary in VSR task and the alignment is processed at the cost of high computational consumptions for obtaining optical flow.

Based on these intuitions, we propose a Multi-Branch Dilation (MBD) module. The overview of MBD module is illustrated in Fig. 2. In MBD module, input feature maps are embedded with a pointwise convolution and processed in parallel with convolutional operations of different dilation factors. Then MBD module fuses all extracted feature with a convolutional layer. The MBD module can be formulated as follow

$$MBD(x) = Conv([DConv(x,1), DConv(x,2), DConv(x,3)]) \tag{2}$$

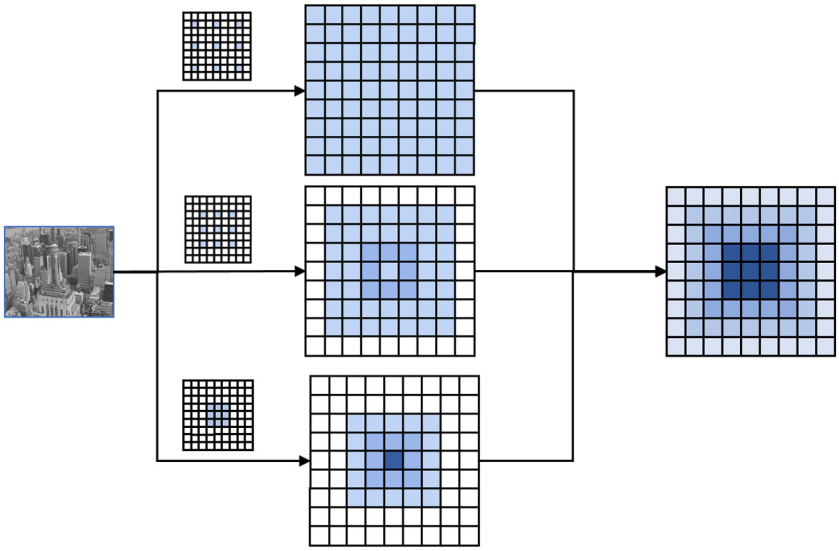

Fig. 4. Illustration of our proposed MBD module. The intensity of color denotes the frequency of convolution operation executed with respect to each pixel.

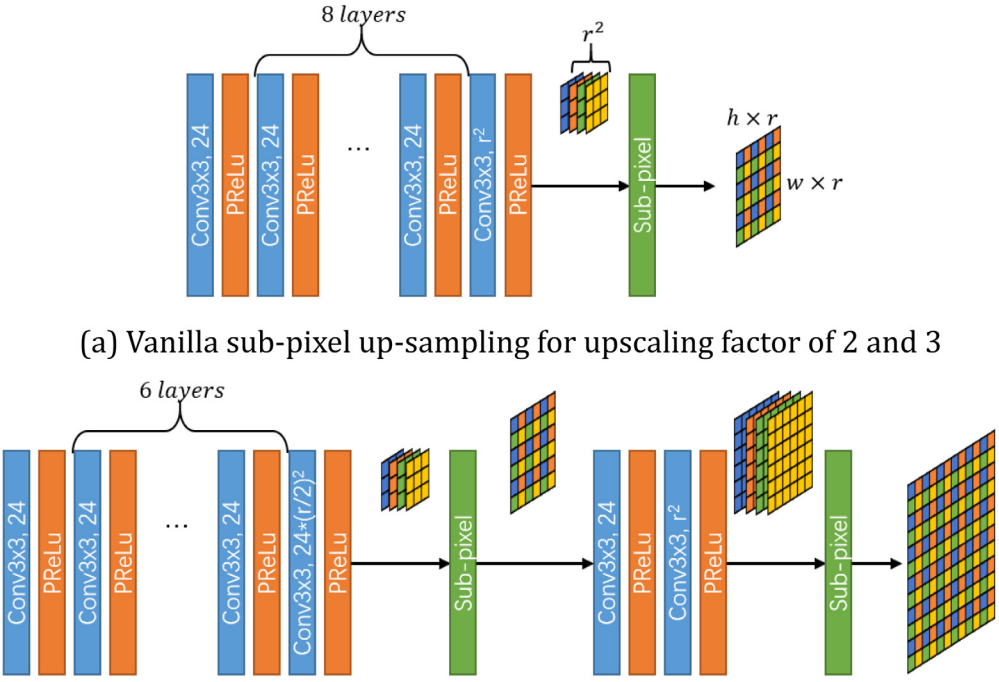

Fig. 5. Illustration of SR network. (a) the final convolutional layer outputs feature map of r^2 channels and followed by a sub-pixel operation while upscale factor $r \in \{2, 3\}$. (b) gradually up-sampling for upscale factor $r = 4$.

where $Conv(x)$ denotes standard convolution of input x, $DConv(x, d)$ denotes a pointwise convolution and a dilated convolution of input x with dilation factor d.

With the benefit of the multi branch architecture, MBD module has receptive fields of 3, 5 and 7. The receptive field of MBD module is demonstrated in Fig. 4. By fusing feature maps from multiple branches, MBD module acquires an overall receptive field of 9×9, which is equivalent to a single standard convolution of kernel size of 7 or a stack 3 standard convolutional layers of kernel size of 3.

3.4. SR network

For SR network, we adopt a 9-layer ESPCN backbone network. The overview of the SR network is illustrated in Fig. 5(a). SR network can be formulated as

$$SR_t = SRnetwork\left(\left[h_{i-n}, \ldots, h_i, \ldots, h_{i+n}\right]\right) + I_t^{hr}, \tag{3}$$

where h_i denotes feature map from FA at temporal position i and I_t^{hr} denotes bicubic interpolation of low-resolution input image.

We introduce sub-pixel convolution (also known as periodic shift operation) for efficient spatial up-sampling. Let r to be the upscaling factor and H, W, C represent height, width, depth of tensor respectively. As illustrated in Fig. 5, generating I^{SR} of size $rH \times rW \times C$

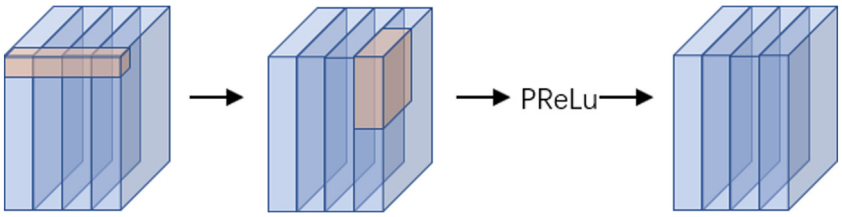

Fig. 6. Illustration of depth-wise separable convolution. Gray denotes filters and blue denotes feature maps.

Table 1
Details of network ERVSR, where $r = 2$ for 4× super-resolution.

Modules	Layers	Kernel	Num	Stride
Feature alignment modules	Conv	3 × 3	32	1
	Depth separable conv	3 × 3	32	1
	MBD module	–	–	–
	Depth separable conv	3 × 3	32	1
	Depth separable conv	3 × 3	32	1
	Depth separable conv	3 × 3	32	1
Up-sampling network	Conv	3 × 3	24	1
	6 × Depth separable conv	3 × 3	24	1
	Depth separable conv	3 × 3	r ∗ r ∗ 24	1
	Sub-pixel upscale × 2	–	–	–
	Depth separable conv	3 × 3	r ∗ r	1
	Sub-pixel upscale × 2	–	–	–

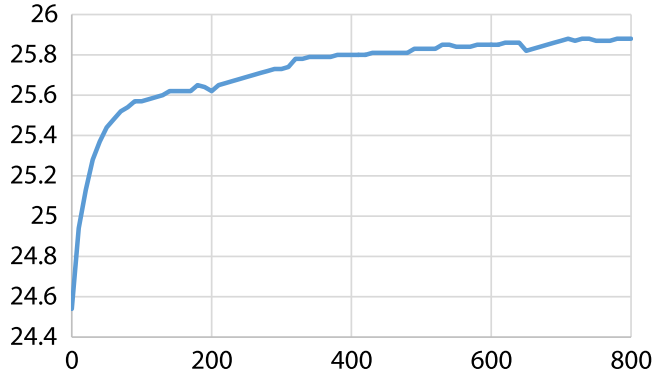

Fig. 7. Convergence curve of our proposed ERVSR.

with sub-pixel operation from given I^{LR} of size $H \times W \times Cr^2$ can be formulated as follow:

$$I^{SR} = ps\left(I^{LR}\right), \tag{4}$$

Moreover, for upscaling factor $r = 4$, we adopt pyramid up-sampling which operates twice up-sampling of scale factor of 2. The demonstration of pyramid up-sampling is illustrated in Fig. 5(b). With the benefit of pyramid up-sampling, the network has a simpler objective to learn, resulting in a superior performance.

Table 3
Average PSNR and SSIM comparisons of different context lengths.

Methods	ERVSR-3F	ERVSR-5F	ERVSR-7F	ERVSR-Static-3F
PSNR (dB)	25.88	25.91	25.94	24.97
SSIM	0.76	0.77	0.77	0.71
GFlops	9.7	14.5	19.4	9.7

Table 4
Average PSNR and SSIM of our proposed ERVSR/3F on single image super-resolution test set Set5 and Set14 for scale factor $r = 4$.

	Set5	Set14
PSNR (dB)	30.93	27.81
SSIM	0.86	0.76

3.5. Depth-wise separable convolution

In order to obtain a more efficient real time video super-resolution network, we adopt depth-wise separable convolution [28,29] in our proposed network to decrease computational costs and speed up inference runtime.

As illustrated in Fig. 6, point-wise convolution applies a 1 × 1 convolution to fuse channel-wise information and depth-wise convolution applies a single filter to each input channel to learn spatial features. By separating the standard convolutional layer into two layers, depth-wise separable convolution dramatically reduces computations by 8 to 9 times while suffering minor decline of performance.

We remove the activation between point-wise convolution and depth-wise convolution. Such modification reduces nonlinearity of our model but also avoid information loss from activation, resulting in better performance. The experimental results are further demonstrated in Section 4.

3.6. Detail of ***ERVSR***

Our **ERVSR** can be decomposed into two parts, feature alignment module and upsampling network. We display the details of ERVSR in Table 1. Feature alignment block consists of 5 convolutional layers and the MBD module and Up-sampling Network consists of 10 convolutional layers.

Since we purpose to design a real-time network, we try to minimize the number of convolutional filters. Therefore, we choose 32 as number of convolutional filters by grid search.

Table 2
Experimental results of ablation study. Average PSNR performances are evaluated on the standard Vid4 dataset for scaling factor $r = 4$. Computations (GFlops) indicate number of giga floating point operations for input resolution of 512 × 383.

Models	VESPCN	ERVSR-a	ERVSR-b	ERVSR-c	ERVSR-d	**ERVSR**
Depth separable conv		√	√	√	√	√
Feature alignment block			√	√	√	√
MBD module				√	√	√
Pyramid upsampling					√	√
Global residual						√
PSNR	25.35	24.57	25.56	25.71	25.76	**25.88**
Computations	14	1.6	5.8	8.7	9.7	9.7

Table 5
Average PSNR and SSIM of our proposed **ERVSR** on different scale factors. We implemented VESPCN on scale factor of 2 since the performance is not provided in paper. Other results of VESPCN, VSRenet and 3DSRnet are referenced from the corresponding papers respectively.

	Scale	Bicubic	VSRnet [14]	VESPCN [15]	3DSRnet [13]	ERVSR
PSNR (dB)/SSIM	2	28.43/0.87	31.30/0.93	31.60/0.93*	32.25/0.94	**32.36/0.94**
	3	25.38/0.73	26.79/0.81	27.25/0.85	27.70/0.85	**27.78/0.85**
	4	22.53/0.63	24.84/0.71	25.35/0.73	25.71/0.76	**25.88/0.76**

Table 6
Average PSNR and SSIM of our proposed **ERVSR** on different temporal orders. Let us define 2 as the reference frame, 1 as the first frame of the video clip and 3 as the last frame. Our proposed network adopts temporal order 2-1-3 by default.

Temporal order	2-1-3	1-2-3	3-2-1	1-3-2
PSNR (dB)	25.88	25.81	25.79	25.79

3.7. Training details

Following the pioneer works, we trained our network with the open source dataset Vimeo 90k with a training set of 64,612 7-frame sequences of fixed resolution of 448 × 256. In Vimeo 90k training set, there exists a certain amount of smooth blur images without rich texture or high frequency information, which is not useful for optimizing network weights. Therefore, we filter those smooth blur video clips with a standard deviation threshold of each video clip itself and remove about 30% of the complete Vimeo 90k dataset. By trimming the dataset, we speed up the training procedure without any decline of performance. We also adopt data augmentation strategies such as random flipping and rotation in training session.

Since human visual system is more sensitive to texture details preserved in Y channel rather than color information in CbCr channels, all training and testing images are transformed to YCbCr format and we use only Y channel in training and testing sessions.

In order to further speed up training session, we set a relatively higher learning rate to $1e{-}3$ and apply gradient clipping to stabilize the training procedure. Learning rate is declined by half for every 100 epochs. We optimize all layers with the same learning rate and ADAM optimizer with momentum of 0.9. Network is trained with batch size of 8 and LR patches are downscaled with bicubic interpolation. As shown in Fig. 7, our network converges in 800 epochs. All experiments are conducted using PyTorch framework on NVIDIA GPUs.

Table 9
PSNR and SSIM comparisons of different video SR methods on the open dataset SPMC for scale factor 4 ×.

Methods	Bicubic	VESPCN [15]	3DSRnet [13]	ERVSR
AMVTG_004	23.50	24.76	25.01	**25.10**
car05_001	27.65	29.20	29.43	**29.74**
hdclub_003_001	19.39	20.40	**20.98**	20.86
hitachi_isee5_001	19.58	21.83	21.56	**22.80**
hk004_006	28.30	30.16	30.13	**30.58**
HKVTG_004	27.38	28.24	28.40	**28.64**
jvc_004_001	26.13	28.86	29.67	**29.78**
jvc_009_001	25.34	27.21	27.56	**27.88**
NYVTG_006	28.44	29.53	30.22	**30.72**
PRVTG_012	25.62	26.34	26.53	**26.75**
RMVTG_011	23.98	25.53	25.66	**26.22**
veni3_011	29.39	32.45	32.61	**33.10**
Average PSNR	25.39	27.04	27.31	**27.68**
Average SSIM	–	0.73	0.76	0.76

Table 10
Average PSNR and SSIM values on REDS dataset.

Method	Bicubic	VESPCN [15]	TDAN [23]	SOF-VSR [30]	Proposed
PSNR	26.26	28.06	28.60	28.62	28.29
SSIM	0.738	0.792	0.814	0.822	0.802

4. Experimental results

4.1. Ablation study

By demonstrating how we make improvements step by step base on VESPCN architecture, we carry out the ablation study of **ERVSR** in this section.

First, we replace all standard convolutions in VESPCN with depthwise separable convolution to reduce the overall computational complexity, denote as ERVSR-a. Second, we replace the motion compensation module in VESPCN with our proposed feature alignment block

Table 7
PSNR and SSIM comparisons of different video SR methods on the standard Vid4 dataset for scaling factor $r = 4$. **ERVSR** achieves superior result among state-of-the-art SR methods. Computations (GFlops) indicate number of giga floating point operations for input resolution of 512×383.

Methods	Bicubic	VSRnet [14]	Robust VSR [22]	VESPCN [15]	Detail-revealing VSR [17]	3DSRnet [13]	SOF-VSR [30]	TDAN [23]	RBPN [18]	**ERVSR**
Calendar	19.82	20.99	21.61	21.92	–	22.41	**22.65**	22.33	23.99	22.51
City	24.93	24.78	26.29	26.22	–	26.81	25.48	**26.99**	27.73	25.30
Foliage	23.42	23.87	24.99	24.89	–	25.23	**26.87**	25.51	26.22	26.69
Walk	26.03	27.02	28.06	28.39	–	28.38	29.04	**29.50**	30.70	29.02
Average PSNR	23.53	24.84	25.24	25.35	25.52	25.71	26.01	26.24	**27.12**	25.88
SSIM	0.63	0.71	0.76	0.73	0.76	0.76	0.77	0.78	0.82	0.76
VMAF	–	–	–	78.51	24.80	35.45	80.24	68.69	75.86	**83.46**
GFlops	–	–	–	14	1322	46	387	**–**	21173	**9.7**

Table 8
Average runtime (ms) comparisons of different video SR methods on the open dataset Vid4 for scaling factor $r = 4$.

Set	VESPCN [15]	Detail revealing VSR [17]	SOF-VSR [30]	TDAN [23]	RBPN [18]	ERVSR
Calendar	35.3	86.6	39.1	6.9	440.8	**3.8**
Walk	36.3	86.5	36.1	7.2	363.6	**3.8**
City	35.9	86.2	38.7	6.9	435.0	**3.7**
Foliage	34.9	86.3	36.6	7.0	362.5	**3.7**
Average	35.6	86.4	39.0	7.0	400.5	**3.8**

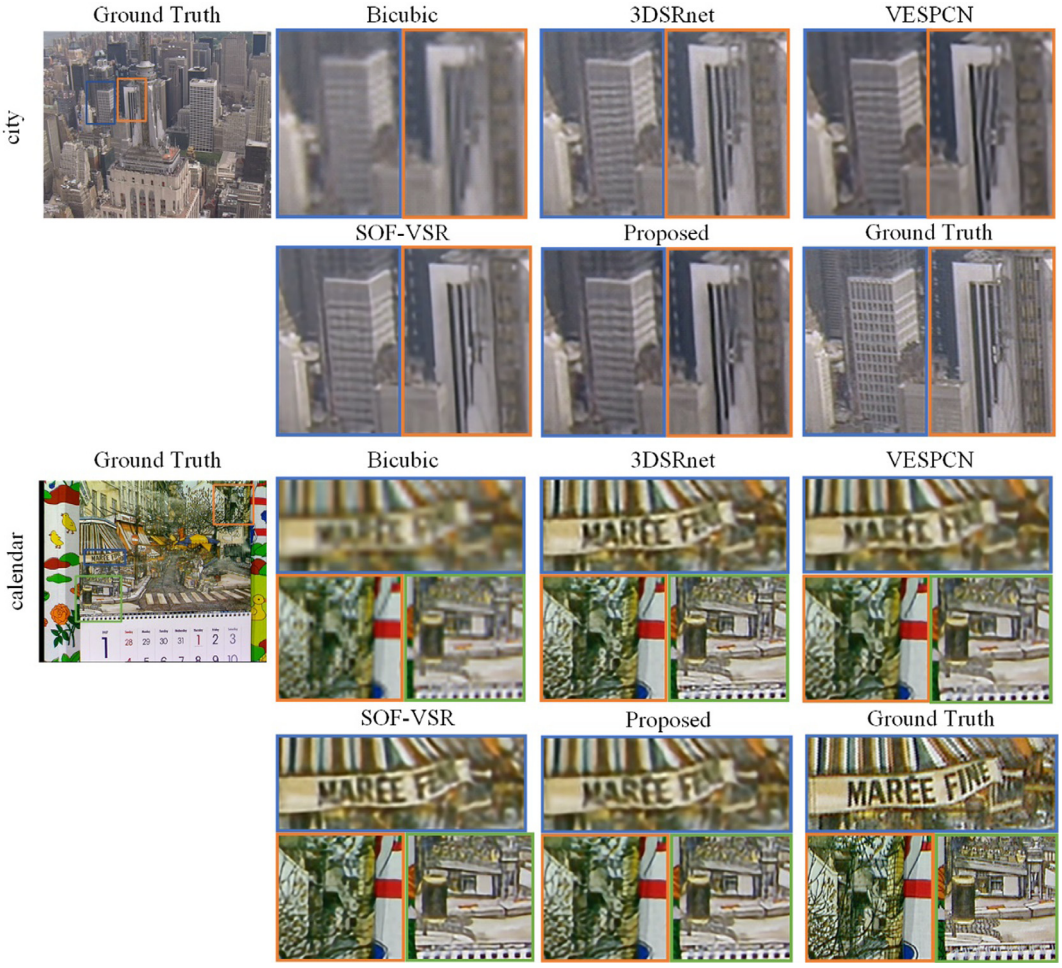

Fig. 8. Comparisons of state-of-the-art approaches on validation set Vid4 for scale factor 4 ×.

and denote it as ERVSR-b. In ERVSR-c, we insert the MBD module into feature alignment block. In ERVSR-d, we adopt the pyramid upsampling method. And in **ERVSR**, our have an additional global residual connection. The experimental results are shown in Table 2. When calculating PSNR, we remove the first and last frame of all tested video clips and 3 pixels of image border for fair comparison against other approaches.

According to experimental results, depthwise separable convolution draws significant reduction of performance when reducing computational complexity. Our proposed feature alignment block provides 0.99 dB increase of performance compared to ERVSR-a which uses motion compensation module from VESPCN. In ERVSR-b and ERVSR-c, we adopt MBD and pyramid upsampling, which introduces trivial computational complexity but provides a further 0.2 dB addition to PSNR. The global residual connection allows our network focusing on refining the high frequency details of an interpolated high-resolution image and provides an increasement of 0.12 dB without addition computational complexity.

4.2. Context length

Since **ERVSR** is a recurrent architecture that is adaptive to arbitrary number of input frames, we compare the influence of context length on its performance. We train the **ERVSR** with video clips of length $n \in \{3, 5, 7\}$, yielding average video PSNR of 25.88, 25.91 and 25.94 on Vid4 respectively. The effect of context length shows our network can generate better reconstruction with benefit of temporal information redundancy.

For further comparisons, we test our model on validation set Vid4 as a SISR network with a single frame input. In each recurrent iteration, we take the same frame pair $[I_t, I_t]$ as input, which is equivalent to feeding the network with a sequence of static LR frame without inter-frame information. Experimental result in Table 3 demonstrates that static frame model suffers a significant decline on PSNR. We suggest that **ERVSR** can extracting and utilizing interframe information as a VSR network instead of generating a better reconstruction as a SISR network. It shows that the superior performance of our network comes from the interpretation of multi-frame information instead of deeper convolutional network.

We also test our model on standard SISR validation sets, Set5 and Set14. According to the experimental results in Table 4, **ERVSR** can effectively generalize from video SR task to single image SR task.

4.3. Scale factor

In Table 5, we compare the performance of the proposed method with bicubic, VSRnet, VESPCN and 3DSRnet on Vid4 validation set of different scale factors. Our **ERVSR** surpasses other methods in both PSNR and SSIM for all different scale factors.

4.4. Temporal order

In our proposed ERVSR, the reference frame pair $[I_t, I_t]$ is the first frame pair fed into the network, while the rest pairs are fed in temporal order. We denote it as temporal order 2-1-3. Feeding the reference frame pair in the first place will provide an initial reference for feature extraction and alignment from the following neighboring frames. We compare the performance of different temporal input order of the

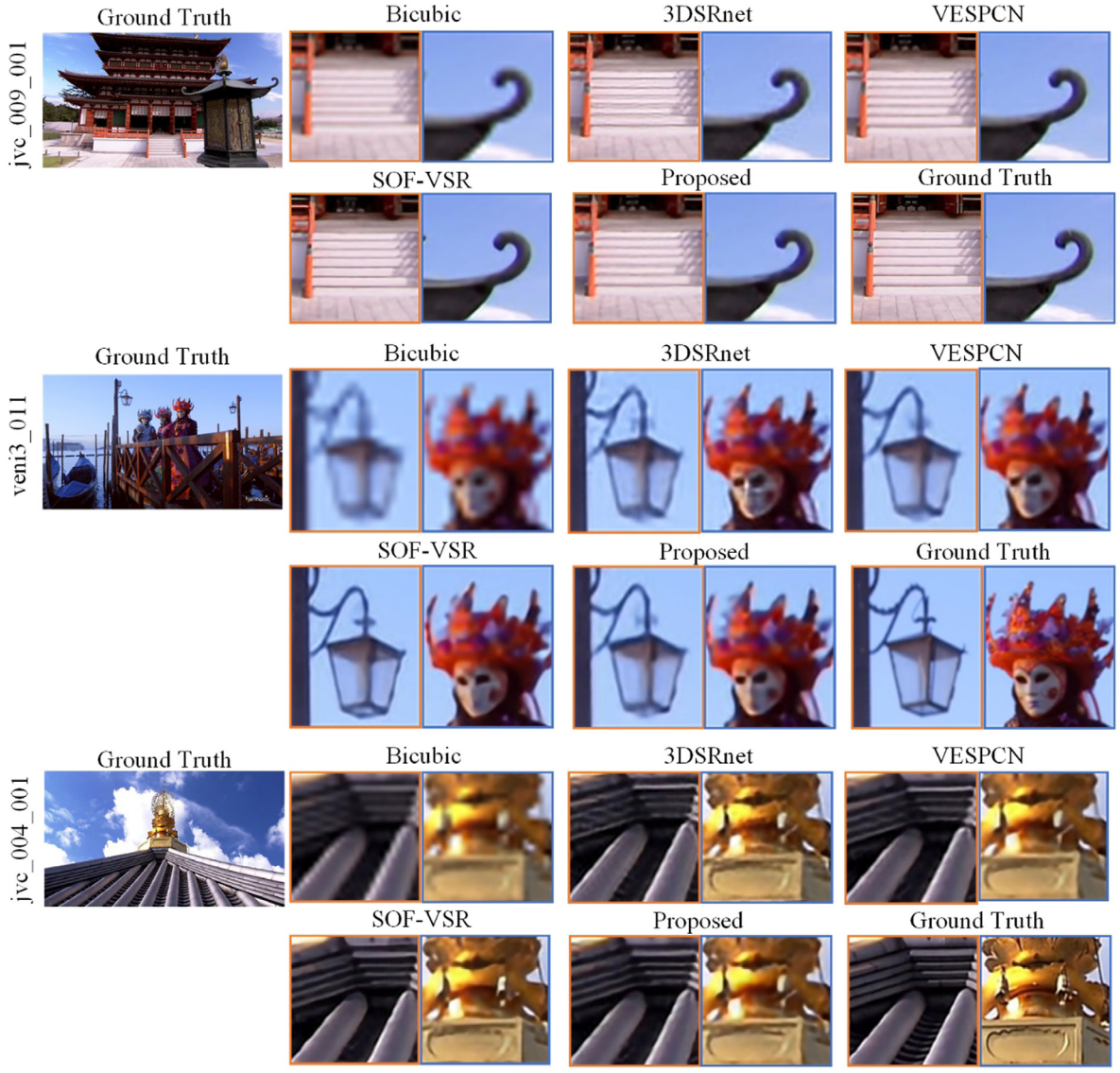

Fig. 9. Comparisons of state-of-the-art-approaches on SPMC dataset for scale factor 4 ×.

frame sequence simulating more complex object motions with different orders of motion trajectories. Experimental results are demonstrated in Table 6, where the temporal order 1-2-3 means feeding frames into the network in original frame temporal order. Temporal order 3-2-1 means feeding in the reverse order. Finally, 1-3-2 means feeding all neighbor frames in the first place while the referenced frame as the last frame. Table 6 shows that our proposed method achieves consistent PSNR results regardless of different temporal orders to be fed into the frame alignment network. Hence, it verifies that the proposed alignment network is flexible to different kinds of motion orders.

4.5. Experimental result

In this section, we will demonstrate both visual and quantitative performance on Vid4, SPMC and REDS datasets. Our experiments mainly focus on scale factor of 4. When testing PSNR on validation set, we crop 3 pixels of image border and remove the first and the last frames.

4.5.1. Testing on Vid4

We compare our proposed network with several state-of-the-art video SR algorithms on validation set Vid4 in Table 7. **ERVSR** achieves state-of-the-art performance compared to other state-of-the-art VSR approaches. Compared with VESPCN, our proposed network achieves 0.53 dB higher in PSNR while requires 33% less total floating-point multiplications. Our network's performance also surpasses Detail-revealing VSR and 3DSRnet by 0.36 dB and 0.17 dB respectively in terms of PSNR. Our network has the best performance on Video Multimethod Assessment Fusion (VMAF) [31], which predicts subjective quality by combining multiple elementary quality metrics.

Visual performance is demonstrated in Fig. 8. Comparing to VESPCN, our approach generates sharper edges and finer details. Edges of buildings are correctly reconstructed with **ERVSR.**

In Table 8, we evaluate the inference time between different approaches. ERVSR requires merely 3.8 ms per frame on average when inferencing Vid4 dataset.

4.5.2. Testing on SPMC

When compared on SPMC dataset, **ERVSR** performs better than VESPCN and 3DSRnet on numbers of video clips from SPMC set. In Fig. 9, comparing to 3DSRnet, **ERVSR** generates high resolution output with less flicker effects and artifacts. And we have similar visual effect as SOF-VSR while requires only 10% of its inference time. Details of performance comparison on SPMC are demonstrated in Table 9.

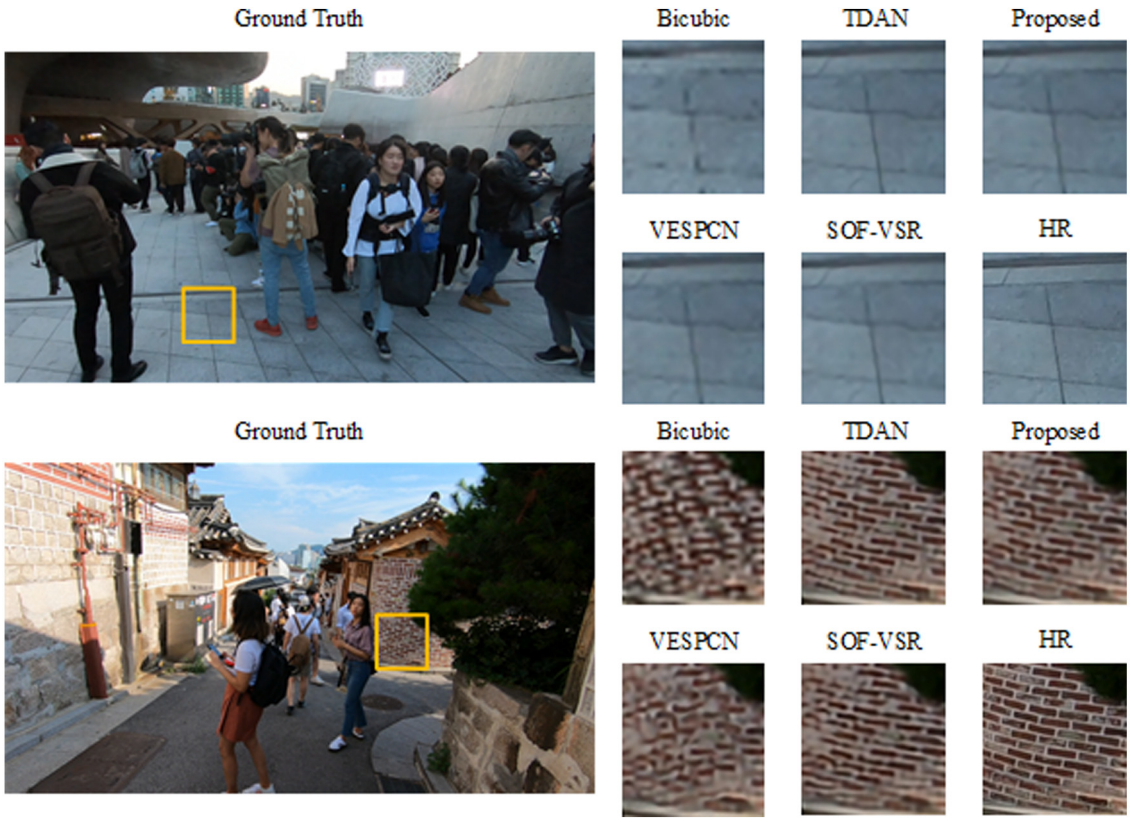

Fig. 10. Comparisons of images from validation set REDS for scale factor 4 ×.

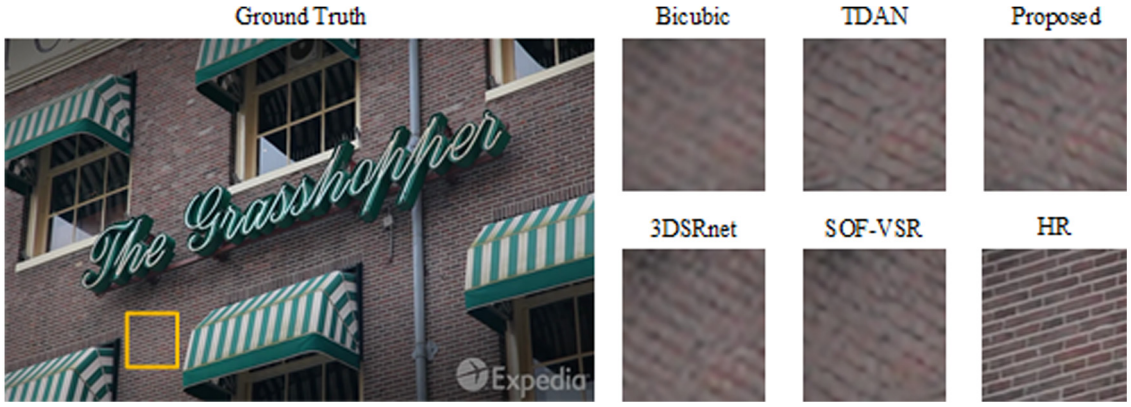

Fig. 11. Failure case. ERVSR failed to restore finer detail if there is insufficient high-frequency information in the low-resolution input.

4.5.3. Testing on REDS

We also make comparison on REDS [32] dataset, which is collected with handheld camera and contains a lot of video clips with large motions and object movements. Quantitative comparison on REDS validation set is shown in Table 10. Without explicit alignment, ERVSR still surpasses VESPCN, which shows the capability of ERVSR handling videos with large motions. Although ERVSR is not highest in PSNR, but as shown in Table 8, ERVSR is significantly faster than comparative approaches TDAN [23] and SOF-VSR [30]. Visual comparisons are demonstrated in Fig. 10.

4.6. Limitations

Although our network achieves superior performance at very low inference time, we notice that there is still limitation in our ERVSR. As shown in Fig. 11, if the high-frequency is lost and distorted by down-sampling operation, all methods fail to restore the texture of brick wall properly. All the approaches that are capable of aggregating inter-frame information are unable to give the desired restoration. Such limitation is very common among super-resolution methods [4,7,15].

4.7. Runtime evaluation

As show in Table 8, our network takes only 3.8 ms per frame inferencing Vid4 dataset. Compared with other methods with similar performance, our network is obviously faster. Moreover, when compared with other methods with similar runtime, our network has much higher performance.

5. Conclusion

In this paper, we propose a new video super-resolution network that can fully exploit the redundancy of inter-frame and intra-frame

information and generate better SR reconstruction with very low computational requirement. From the ablation study, we demonstrate that the effectiveness of the feature alignment block, pyramid upsampling method and other modules. Specifically, **ERVSR** reduces computational complexity with the help of feature alignment block and depthwise separable convolutions. Experiments on public datasets show that **ERVSR** achieves state-of-the-art performance among other video super-resolution approaches but the proposed network requires comparatively much lower computational requirement, which is essential to real-time applications in real scenarios.

CRediT authorship contribution statement

Yubin Zeng: Conceptualization, Methodology, Software, Writing. **Zhijiao Xiao:** Data curation, Writing - original draft preparation. **Kwok-Wai Hung:** Writing, Visualization, Investigation, Supervision, Reviewing and editing. **Simon Lui:** Editing, Reviewing, Manuscript correction.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work is funded by National Natural Science Foundation of China (NSFC) with grant no: 62002230.

References

[1] T. Goto, T. Fukuoka, F. Nagashima, S. Hirano, M. Sakurai, Super-resolution system for 4K-HDTV, in: 2014 22nd International Conference on Pattern Recognition, 2014, pp. 4453–4458.
[2] L. Zhang, H. Zhang, H. Shen, P. Li, A super-resolution reconstruction algorithm for surveillance images, Signal Process. 90 (3) (2010) 848–859.
[3] M.W. Thornton, P.M. Atkinson, D.a. Holland, Sub-pixel mapping of rural land cover objects from fine spatial resolution satellite sensor imagery using super-resolution pixel-swapping, Int. J. Remote Sens. 27 (3) (2006) 473–491.
[4] C. Dong, C.C. Loy, K. He, et al., Image super-resolution using deep convolutional networks, IEEE Trans. Pattern Anal. Mach. Intell. 38 (2) (2015) 295–307.
[5] C. Dong, C.C. Loy, X. Tang, Accelerating the super-resolution convolutional neural network, in: European Conference on Computer Vision, Springer, Cham, 2016, pp. 391–407.
[6] W. Shi, J. Caballero, F. Huszár, et al., 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.
[7] J. Kim, J. Kwon Lee, 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.
[8] M. Haris, G. Shakhnarovich, 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.
[9] C. Ledig, L. Theis, F. Huszár, et al., Photo-realistic single image super-resolution using a generative adversarial network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4681–4690.
[10] Y. Jo, S. Wug Oh, J. Kang, et al., 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.
[11] R. Liao, X. Tao, R. Li, et al., Video super-resolution via deep draft-ensemble learning, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 531–539.
[12] Y. Huang, W. Wang, L. Wang, Bidirectional recurrent convolutional networks for multi-frame super-resolution, in: Advances in Neural Information Processing Systems, 2015, pp. 235–243.
[13] S.Y. Kim, J. Lim, T. Na, et al., 3DSRnet: Video super-resolution using 3D convolutional neural networks, 2018, arXiv preprint arXiv:1812.09079.
[14] A. Kappeler, S. Yoo, Q. Dai, A.K. Katsaggelos, Video super-resolution with convolutional neural networks, in: IEEE Transactions on Computational Imaging, 2016.
[15] J. Caballero, C. Ledig, A. Aitken, et al., 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.
[16] M.S.M. Sajjadi, R. Vemulapalli, M. Brown, Frame-recurrent video super-resolution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6626–6634.
[17] X. Tao, H. Gao, R. Liao, et al., Detail-revealing deep video super-resolution, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4472–4480.
[18] M. Haris, G. Shakhnarovich, 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.
[19] J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution, in: European Conference on Computer Vision, Springer, Cham, 2016, pp. 694–711.
[20] X. Wang, K. Yu, S. Wu, et al., Esrgan: Enhanced super-resolution generative adversarial networks, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, 0.
[21] M. Drulea, S. Nedevschi, Total variation regularization of local–global optical flow, in: ITSC, 2011.
[22] D. Liu, Z. Wang, Y. Fan, et al., Robust video super-resolution with learned temporal dynamics, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2507–2515.
[23] Tian Yapeng, et al., TDAN: Temporally-deformable alignment network for video super-resolution, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.
[24] Wang Xintao, et al., Edvr: Video restoration with enhanced deformable convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019.
[25] Jo Younghyun, et al., 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.
[26] F. Yu, V. Koltun, Multi-scale context aggregation by dilated convolutions, 2015.
[27] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.
[28] A.G. Howard, M. Zhu, B. Chen, et al., Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017, arXiv preprint arXiv:1704.04861.
[29] M. Sandler, A. Howard, M. Zhu, et al., Mobilenetv2: Inverted residuals and linear bottlenecks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4510–4520.
[30] L. Wang, Y. Guo, Z. Lin, et al., Learning for video super-resolution through HR optical flow estimation, in: Asian Conference on Computer Vision, Springer, Cham, 2018, pp. 514–529.
[31] R. Rassool, VMAF reproducibility: validating a perceptual practical video quality metric, in: 2017 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), Cagliari, 2017, pp. 1–2.
[32] Nah Seungjun, Baik, et al., Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019.