Video Super-Resolution with Recurrent Structure-Detail Network

Takashi Isobe[1,2⋆], Xu Jia[2✉], Shuhang Gu[3], Songjiang Li[2], Shengjin Wang[1✉], and Qi Tian[2]

[1] Department of Electronic Engineering, Tsinghua University
jbj18@mails.tsinghua.edu.cn, wgsgj@tsinghua.edu.cn
[2] Noah's Ark Lab, Huawei Technologies
{x.jia, songjiang.li, tian.qi1}@huawei.com
[3] School of Eie, The University of Sydney
shuhanggu@gmail.com

Abstract. Most video super-resolution methods super-resolve a single reference frame with the help of neighboring frames in a temporal sliding window. They are less efficient compared to the recurrent-based methods. In this work, we propose a novel recurrent video super-resolution method which is both effective and efficient in exploiting previous frames to super-resolve the current frame. It divides the input into structure and detail components which are fed to a recurrent unit composed of several proposed two-stream structure-detail blocks. In addition, a hidden state adaptation module that allows the current frame to selectively use information from hidden state is introduced to enhance its robustness to appearance change and error accumulation. Extensive ablation study validate the effectiveness of the proposed modules. Experiments on several benchmark datasets demonstrate superior performance of the proposed method compared to state-of-the-art methods on video super-resolution. Code is available at https://github.com/junpan19/RSDN.

Keywords: Video Super-Resolution, Recurrent Neural Network, Two-Stream Block

1 Introduction

Super-resolution is one of the fundamental problem in image processing, which aims at reconstructing a high resolution (HR) image from a single low-resolution (LR) image or a sequence of LR images. According to the number of input frames, the field of SR can be divided into two categories, *i.e.*, single image super-resolution (SISR) and multi-frame super-resolution (MFSR). For SISR, the key issue is to exploit natural image prior for compensating missing details; while for MFSR, how to take full advantage from additional temporal information is of pivotal importance. In this work, we focus on the video super-resolution (VSR) task which belongs to MFSR. It draws much attention in both research and

⋆ The work was done in Noah's Ark Lab, Huawei Technologies.

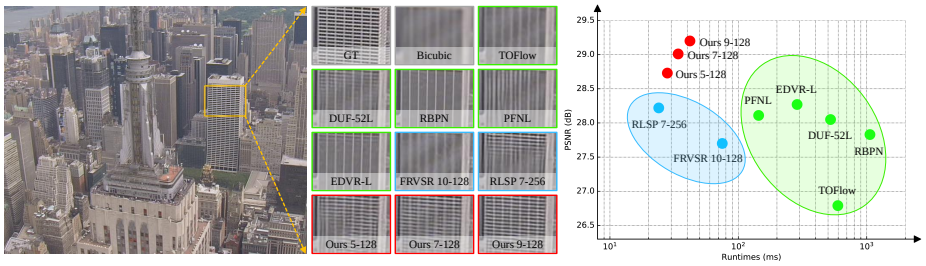

Fig. 1. VSR results on the **City** sequence in Vid4. Our method produces finer details and stronger edges with better balance between speed and performance than both temporal sliding window based [27,12,7,26,29] and recurrent based methods [23,4]. Blue box represents recurrent-based and green box represents sliding window based methods. Runtimes (ms) are calculated on an HR image of size 704×576.

industrial communities because of its great value on computational photography and surveillance.

In the last several years, great attempts have been made to exploit multi-frame information for VSR. One category of approaches utilize multi-frame information by conducting explicit motion compensation. These approaches [1,13,27,25,21] firstly compute optical flow between a reference frame and neighboring frames and then employ the aligned observations to reconstruct the high-resolution reference frame. However, estimating dense optical flow itself is a challenging and time-consuming task. Inaccurate flow estimation often leads to unsatisfactory artifacts in the SR results of these flow-based VSR approaches. In addition, the heavy computational burden also impedes the application of these applications in resource-constrained devices and time-sensitive scenarios. In order to avoid explicit motion compensation, another category of methods propose to exploit the motion information in an implicit manner. The dynamic upsampling filters [12] and the progressive fusion residual blocks [29] are designed to explore flow-free motion compensation. With these flexible compensation strategies, [12,29] not only avoid heavy motion estimation step but also achieve highly competitive VSR performance. However, they still suffer from the redundant computation for several neighboring frames within a temporal window and need to cache several frames in advance to conduct VSR. Recently, for the pursuit of efficiency, there is an emerging trend of applying recurrent connection to address the VSR task.

These approaches [23,4] make use of recurrent connection to conduct video super-resolution in a streaming way, that is, output or hidden state of previous time steps is used to help super-resolve future frames. In addition, they are able to exploit temporal information from many frames in the past. By simply propagating output and hidden state of previous steps with a recurrent unit, they achieve promising VSR performance with considerably less processing time.

In this paper, we propose a novel recurrent network for efficient and effective video super-resolution. Instead of simply concatenating consecutive three frames with previous hidden state as in [4], we propose to decompose each frame of a

sequence into components of structure and detail and aggregate both current and previous structure and detail information to super-resolve each frame. Such a strategy not only allows our method to address different difficulties in the structure and detail components, but also able to impose flexible supervision to recover high-frequency details and strengthen edges in the reconstruction.

In addition, we observe that hidden state in a recurrent network captures different typical appearances of a scene over time. To make full use of temporal information in hidden state, we treat the hidden state as a historical dictionary and compute correlation between the reference frame and each channel in hidden state. This allows the current frame to highlight the potentially helpful information and suppress outdated information such that information fusion would be more robust to appearance change and accumulated errors. Extensive ablation study demonstrates the effectiveness of the proposed method. It performs very favorably against state-of-the-art methods on several benchmark datasets, in both super-resolution performance and speed.

2 Related Work

2.1 Single Image Super-Resolution

Traditional SISR methods include interpolation-based methods and dictionary learning-based methods. However, since the rise of deep learning, most traditional methods are outperformed by deep learning based methods. A simple three-layer CNN is proposed by Dong [2], showing great potential of deep learning in super-resolution for the first time. Since then, plenty of new network architectures [14,18,19,31,6,30] have been designed to explore power of deep learning to further improve performance of SISR. In addition, researchers also investigate the role of losses for better perceptual quality. More discussions can be found in a recent survey [28]. A very relevant work is the DualCNN method proposed by Pan *et al.* [22], where authors proposed a network with two parallel branches to reconstruct structure and detail components of an image, respectively. However, different from that work, our method aims at addressing the video super-resolution task. It decomposes the input frames into structure and detail components and propagates them with a recurrent unit that is composed of two interleaved branches to reconstruct the high-resolution targets. It is motivated by the assumption that structure and detail components not only suffer from different difficulties in high-resolution reconstruction but also take benefit from other frames in different ways.

2.2 Video Super-Resolution

Although SISR methods can also be used to address the video super-resolution task, they are not very effective because they only learn to explore natural prior and self-similarity within an image and ignore rich temporal information in a sequence. The key to video super-resolution is to make full use of complementary information across frames. Most video super-resolution methods can be

roughly divided into two categories according to whether they conduct motion compensation in an explicit way or not.

Explicit motion compensation. Most methods with explicit motion compensation follow a pipeline of motion estimation, motion compensation, information fusion and upsampling. VESPCN [1] presents a joint motion compensation and video super-resolution with a coarse-to-fine spatial transformer module. Tao *et al.* [25] proposed an SPMC module for sub-pixel motion compensation and used a ConvLSTM to fuse information across aligned frames. Xue *et al.* [27] proposed a task-oriented flow module that is trained together with a video processing network for video denoising, deblock or super-resolution. In [23], Sajjadi *et al.* proposed to super-resolve a sequence of frames in a recurrent manner, where the result of previous frame is warped to the current frame and two frames are concatenated for video super-resolution. Haris *et al.* [7] proposed to use a recurrent encoder-decoder module to exploit explicitly estimated inter-frame motion. Wang *et al.* [26] proposed to align multiple frames to a reference frame in feature space with a deformable convolution based module and fuse aligned frames with a temporal and spatial attention module. However, the major drawback of such methods is the heavy computational load introduced by motion estimation and motion compensation.

Implicit motion compensation. As for methods with implicit motion compensation [12,29,7,4], they do not estimate motion between frames and align them to a reference frame, but focus on designing an advanced fusion module such that it can make full use of complementary information across frames. Jo *et al.* [12] proposed to use a 3D CNN to exploit spatial-temporal information and predict a dynamic upsampling filter to reconstruct HR images. In [29], Yi *et al.* proposed to fuse spatial-temporal information across frames in a progressive way and use a non-local module to avoid explicit motion compensation. Video super-resolution with implicit motion can also be done with recurrent connection. Huang *et al.* [9] proposed a bidirectional recurrent convolutional network to model temporal information across multiple frames for efficient video super-resolution. In [4], Fuoli *et al.* proposed to conduct temporal information propagation with a recurrent architecture in feature space. Our method also adopts the recurrent way to conduct video super-resolution without explicit motion compensation. However, different from the above methods, we proposed to decompose a frame into two components of structure and detail and propagate them separately. In addition, we also compute correlation between the current frame and the hidden state to adaptively use the history information in the hidden state for better performance and less risk of error accumulation.

2.3 Recurrent Networks for Video-based Tasks

Recurrent networks have been widely used in different video recognition tasks. Donahue *et al.* [15] proposed a class of recurrent convolutional architectures which combine convolutional layers and long-range temporal information for

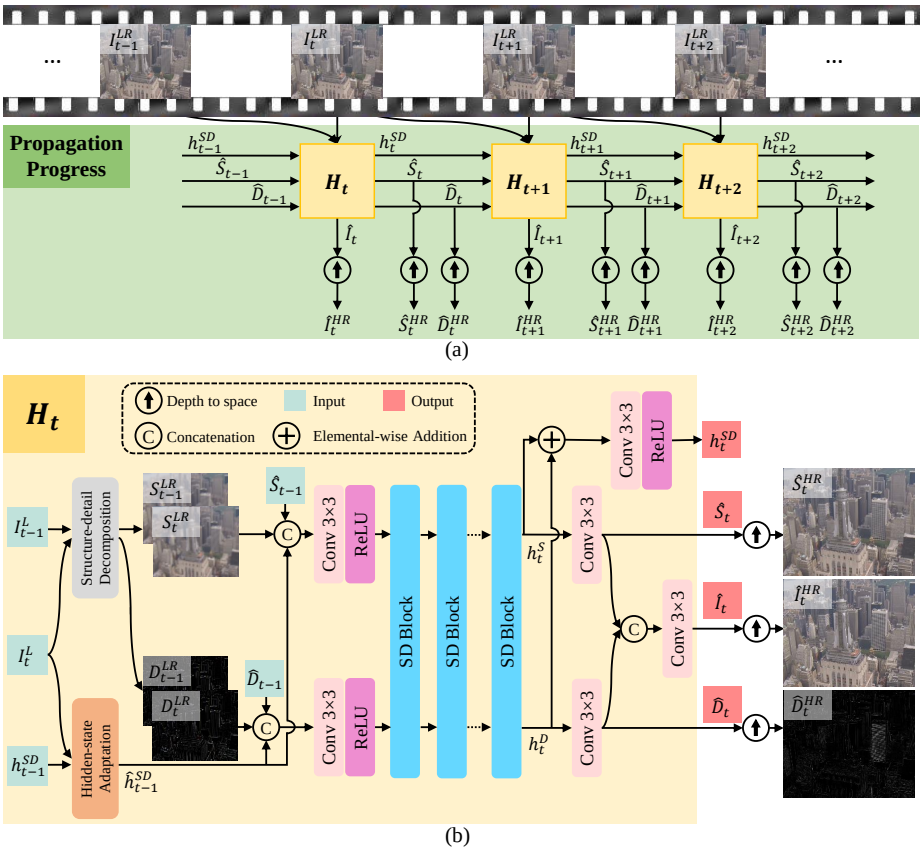

Fig. 2. (a) The overall pipeline of the proposed method; (b) architecture of the recurrent structure-detail unit.

action recognition and image captioning. In [24], a bi-directional LSTM is applied after a multi-stream CNN to fully explore temporal information in a sequence for fine-grained action detection. Du *et al.* [3] proposed a recurrent network with a pose attention mechanism which exploits spatial-temporal evolution of human pose to assist action recognition. Recurrent networks are capable of processing sequential information by integrating information from each frame in their hidden states. They can not only be used for high-level video recognition tasks but are also suitable for low-level video processing tasks.

3 Method

3.1 Overview

Given a low-resolution video clip $\{I^{LR}_{1:N}\}$, $N \geq 2$, the goal of VSR is to produce a high-resolution video sequence $\{\hat{I}^{HR}_{1:N}\}$ from the corresponding low-resolution

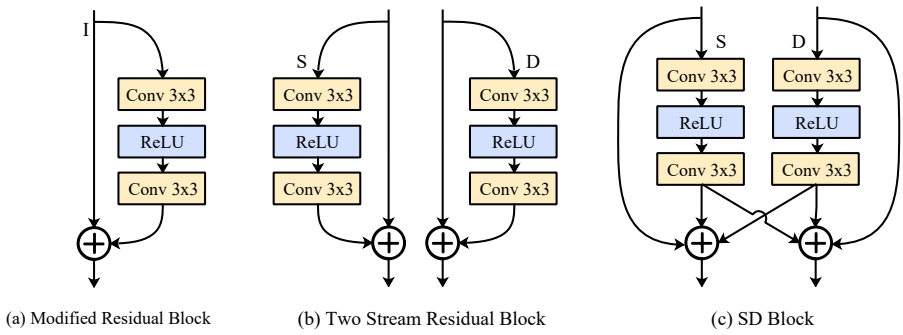

Fig. 3. Variant design for Structure-Detail block.

one by filling in missing details for each frame. In order to process a sequence efficiently, we conduct VSR in a recurrent way similar to [23,4]. However, instead of feeding a whole frame to a recurrent network at each time step, we decompose each input frame into two components, *i.e.*, a structure component and a detail component, to the following network. Two kinds of information interact with each other in the proposed SD blocks over time, which is not only able to sharpen the structure of each frame but also manages to recovers missing details. In addition, to make full use of complementary information stored in hidden states, we treat hidden state as a history dictionary and adapt this dictionary to the demand of the current frame. This allow us to highlight the potential helpful information and suppress outdated information. The overall pipeline is shown in Fig. 2(a).

3.2 Recurrent Structure-Detail Network

Recurrent unit. Each frame can be decomposed into a structure component and a detail component. The structure component models low-frequency information in an image and motion between frames. While the detail component captures fine high-frequency information and slight change in appearance. These two components suffer from different difficulty in high-resolution reconstruction and take different benefit from other frames, hence should be processed separately.

In this work, we simply apply a pair of bicubic downsampling and upsampling operations to extract structural information from a frame I_t^{LR}, which is denoted as S_t^{LR}. The detail component D_t^{LR} can be then computed as the difference between the input frame I_t^{LR} and the structure component S_t^{LR}. In fact, we can also use other ways such as low-pass filtering and high-pass filtering to get these two components. For simplicity, we adopt a symmetric architecture for the two components in the recurrent unit, as shown in Fig. 2 (b). Here we only take D-branch at time step t as an example to explain its architecture design. Detail components of the previous and current frames $\{D_{t-1}^{LR}, D_t^{LR}\}$ are concatenated with the previously estimated detail map $\hat{D}_{t-1}$ and hidden state $\hat{h}_{t-1}^{SD}$ along the channel axis. Such information is further fused by one 3×3 convolutional layer

and several structure-detail (SD) blocks. In this way, this recurrent unit manages to integrate together information from two consecutive input frames, output of the previous time step and historical information stored in the hidden state. h_t^D denotes the feature computed after several SD blocks. It goes through another 3×3 convolutional layer and an upsampling layer to produce the high resolution detail component $\hat{D}_t^{HR}$. The S-branch is designed in a similar way. h_t^S and h_t^D are combined to produce the final high resolution image $\hat{I}_t^{HR}$ and new hidden state h_t^{SD}. The D-branch focuses on extracting complementary details from past frames for the current frame while the S-branch focuses on enhancing existed edges and textures in the current frame.

Structure-Detail block. Residual block [18] and dense block [8] are widely used in both high-level and low-level computer vision tasks because of their effectiveness in mining and propagating information. In this section, we compare several variants of blocks in propagating information in a recurrent structure-detail unit. For comparison, we also include a modified residual block as shown in Fig. 3(a), which only has one branch and takes the whole frames as input. To adapt it to address two branches, the easiest way is to have two modified residual blocks that process two branches separately, as shown in Fig. 3(b). However, in this way each branch only sees the component-specific information and can not makes full use of all information in the input frames. Therefore, we propose a new module called structure-detail (SD) block, as shown in Fig. 3(c). The two components are first fed to two individual branches and then combined with an addition operation. In this way, it not only specializes on each component but also promotes information exchange between structure and detail components. Its advantage over the other two variants is validated in the experiment section.

3.3 Hidden State Adaptation

In a recurrent neural network, hidden state at time step t would summarize past information in the previous frames. When applying a recurrent neural network to the video super-resolution task, hidden state is expected to model how a scene's appearance evolves over time, including both structure and detail. The previous recurrent-based VSR method [4] directly concatenates previous hidden state and two input frames and feeds it to several convolutional layers. However, for each LR frame to be super-resolved, it has distinct appearance and is expected to borrow complementary information from previous frames in different ways. Applying the same integration manner to all frames is not optimal and could hurt the final performance. As shown in Fig. 4, different channels in hidden state describe different scene appeared in the past. They should make different contribution to different positions of different frames, especially when there are occlusion and large deformation with some channels of the hidden state.

In this work, we proposed the Hidden State Adaptation (HSA) module to adapt a hidden state to the appearance of the current frame. As for each unit in hidden state, it should be highlighted if it has similar appearance as the current frame; otherwise, it should be suppressed if it looks very different. With this module the

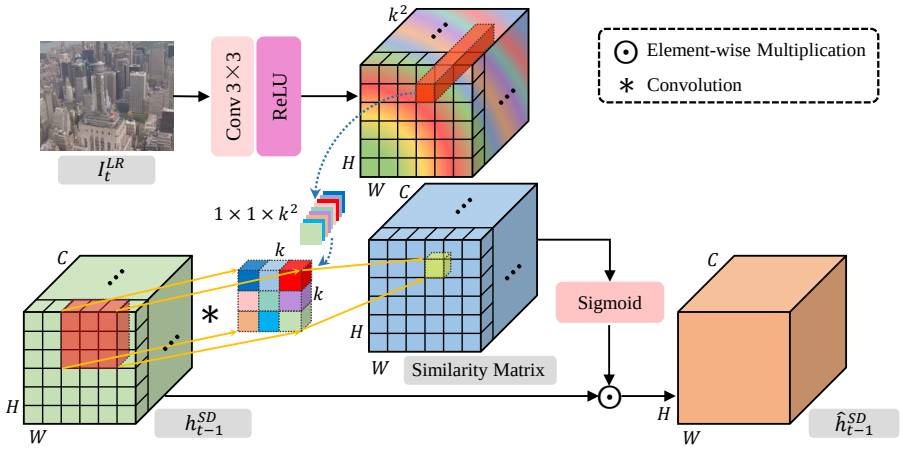

Fig. 5. Design of hidden state adaptation module.

proposed method carefully chooses only useful information in previous frames, hence alleviate the influence of drastic appearance change and error accumulation. Since response of a filter models correlation between the filter and a neighborhood on an image, here we take similar way to compute correlation between an input frame and hidden state. Inspired by [11], we generate spatially variant and sample specific filters for each position in the current frame and use those filters to compute their correlation with the corresponding positions in each channel of hidden state. Specifically, these spatially variant filters $F_t^\theta \in R^{H \times W \times (k \times k)}$ are obtained by feeding the current frame $I_t^{LR} \in R^{H \times W \times 3}$ into a convolutional layer with ReLU activation function [5], where H and W are respectively height and width of the current frame, and k denotes the size of filters. Then, each filter $F_t^\theta(i,j)$ are applied to a $k \times k$ window of h_{t-1}^{SD} centered at position (i,j) to conduct spatially variant filtering. This process can be formulated as:

$$M_t(i,j,c) = \sum_{u=-\lceil k/2 \rceil}^{\lfloor k/2 \rfloor} \sum_{v=-\lfloor k/2 \rfloor}^{\lfloor k/2 \rfloor} F_t^\theta(i,j,u,v) \times h_{t-1}^{SD}(i+u, j+v, c), \tag{1}$$

where $M_t(i,j,c)$ represents correlation between the current frame and the c-th channel of hidden state at position (i,j). It is further fed to a sigmoid activation function $\sigma(\cdot)$ that transforms it into a similarity value in range $[0,1]$. Finally, the adapted hidden state $\hat{h}_{t-1}^{SD}$ is computed by:

$$\hat{h}_{t-1}^{SD} = M_t \odot h_{t-1}^{SD}, \tag{2}$$

where '$\odot$' denotes element-wise multiplication.

3.4 Loss functions

Since the proposed recurrent network has two streams, the trade-off between supervision on structure and detail during training is very important. Imbalanced supervision on structure and detail might produce either sharpened frames but with less details or frames with many weak edges and details. Therefore, we propose to train the proposed network with three loss terms as shown in eq. 3, one for structure component, one for detail component, and one for the whole frame. α, β and γ are hyper-parameters to balance the trade-off of these three terms. The loss to train an N-frame sequence is formulated as:

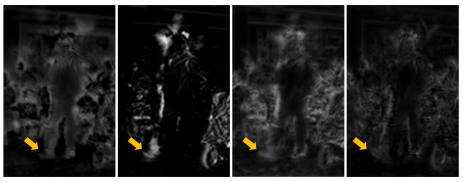

Fig. 4. Four channels in hidden state at a certain time step are selected for visualization. Yellow arrow denotes the difference in appearance among these four channels. Zoom in for better visualization.

$$\mathcal{L} = \frac{1}{N}\sum_{t=1}^{N}(\alpha\mathcal{L}_t^{\mathcal{S}} + \beta\mathcal{L}_t^{\mathcal{D}} + \gamma\mathcal{L}_t^{\mathcal{I}}). \tag{3}$$

Similar to [17], we use Charbonnier loss function to compute the difference between reconstruction and high-resolution targets. Hence, we have $\mathcal{L}_t^{\mathcal{S}} = \sqrt{\|S_t^{HR} - \hat{S}_t^{HR}\|^2 + \varepsilon^2}$ for structure component, $\mathcal{L}_t^{\mathcal{D}} = \sqrt{\|D_t^{HR} - \hat{D}_t^{HR}\|^2 + \varepsilon^2}$ for detail component, and $\mathcal{L}_t^{\mathcal{I}} = \sqrt{\|I_t^{HR} - \hat{I}_t^{HR}\|^2 + \varepsilon^2}$ for the whole frame. The effectiveness of these loss functions is validated in the experiment section.

4 Experiments

In this section, we first explain the experiment datasets and implementation details of the proposed method. Then extensive ablation study is conducted to analyze the effectiveness of the proposed SD block and hidden state adaptation module. Furthermore, the proposed method is compared with state-of-the-art video super-resolution methods in terms of both effectiveness and efficiency.

4.1 Implementation Details

Datasets. Some works [23,29] collect private training data from youtube on their own, which is not suitable for fair comparison with other methods. In this work, we adopt a widely used video processing dataset Vimeo-90K to train video super-resolution models. Vimeo-90K is a recent proposed large dataset for video processing tasks, which contains about $90K$ 7-frame video clips with various motions and diverse scenes. About $7K$ video clips select out of $90K$ as the test set, termed as Vimeo-90K-T. To train our model, we crop patches of size 256×256 from HR video sequences as the target. Similar to [12,23,4,29], the corresponding

Method	One Stream 7-256		Two Stream 7-128			SD Block 7-128		
Model	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
HSA?	w/o	w/	w/	w/o	w/	w/	w/o	w/
Input	Image	Image	Image	S & D	S & D	Image	S & D	S & D
PSNR/SSIM	27.58/0.8410	27.65/0.8444	27.70/0.8452	27.64/0.8404	27.68/0.8429	27.73/0.8460	27.76/0.8463	27.79/0.8474

Table 1. Ablation study on different network architecture.

low-resolution patches are obtained by applying Gaussian blur with $\sigma = 1.6$ to the target patches followed by $\times 4$ times downsampling.

To validate the effectiveness of the proposed method, we evaluate our models on several popular benchmark datasets, including Vimeo-90K-T [27], Vid4 [20] and UDM10 [29]. As mentioned above, Vimeo-90K-T contains a lot of video clips, but each clip has only 7 frames. Vid4 and UDM10 are long sequences with diverse scenes, which is suitable to evaluate the effectiveness of recurrent-based method in information accumulation [23,4].

Training Details. The base model of our method consists of 5 SD blocks where each convolutional layer has 128 channels, *i.e.*, RSDN 5-128. By adding more SD blocks, we can obtain RSDN 7-128 and RSDN 9-128. The performance can be further boosted with only small increase on computational cost and runtime. We adopt $K = 3$ for HSA module for efficiency. To fully utilize all given frames, we pad each sequence by reflecting the second frame at the beginning of the sequence. When dealing with the first frame of a sequence, the previous estimated detail $\hat{D}_{t-1}$, structure $\hat{S}_{t-1}$ and hidden state feature h_{t-1}^{SD} are all initialized with zeros. The model training is supervised with Charbonnier penalty loss function and is optimized with Adam optimizer [16] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Each mini-batch consists of 16 samples. The learning rate is initially set to 1×10^{-4} and is later down-scaled by a factor of 0.1 every 60 epoch till 70 epochs. The training data is augmented by standard flipping and rotating. All experiments are conducted on a server with Python 3.6.4, PyTorch 1.1 and Nvidia Tesla V100 GPU.

Recurrent Unit. We compare three kinds of blocks for information flow in the recurrent unit, *i.e.*, the three blocks shown in Fig. 3. For fair comparison among these blocks, we keep these three networks with almost the same parameters by setting the channel of convolutional layers in model 1 to 256, and setting the one in model 4 and 7 to 128.

4.2 Ablation Study

In this section, we conduct several ablation studies to analyze the effectiveness of the proposed SD block and the hidden state adaptation module. In addition, we also investigate the influence of different supervision on structure and detail components on the reconstruction performance. As shown in Tab. 1, model 1 and model 4 achieves similar performance, with model 1 a little higher SSIM and model 4 a little higher PSNR. This implies that simply dividing the input into

(α, β, γ)	$(1, 0.5, 1)$	$(0.5, 1, 1)$	$(1, 1, 0)$	$(1, 1, 1)$
PSNR/SSIM	27.56/0.8440	27.77/0.8459	27.73/0.8453	27.79/0.8474

Table 2. Ablation study on influence of different loss items.

structure and detail components and processing each one individually does not work well. Although it seems that having two branches to process each component divides a difficult task into two easier ones, it makes each one blind to the other and can not make full use of the information in the input to reconstruct either component.

By introducing information exchange between structure and detail components, model 7 obtains better performance than model 1 and 4 in both PSNR and SSIM. Similar result can also found in comparison among model 2, 5 and 8. In addition, we experiment with taking the whole frames as input of both branches, that is, model 3 and model 6. By comparing model 3 and model 5 (and also model 6 and model 8), we show that the improvement comes not only from architecture of the two-stream block itself but also indeed from the decomposition into structure and detail components. The network with the proposed SD block allows each branch to explicitly focus on reconstructing a single component, which is easier than reconstructing a mixture of multiple components. Each branch makes use of the other one such that it can obtain enough information to reconstruct the high-resolution version for that component. The advantage of the proposed SD blocks can also be observed in the qualitative comparison as shown in Fig. 6.

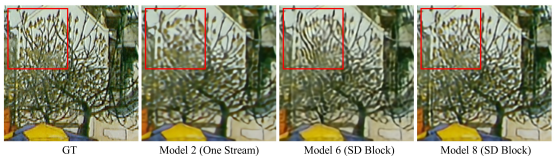

Fig. 6. Qualitative comparison between different network structures. Zoom in to see better visualization.

In addition, we show in Tab. 1 that each model can gain further boost in performance with the proposed HSA module, about 0.04 dB in PSNR and 0.002 in SSIM on average. This module does not only work for the proposed network with SD blocks but also helps improve the performance for the ones with one-stream and two-stream residual blocks. The hidden state adaptation module allows the model to selectively use the history information stored in hidden state, which makes it robust to appearance change and error accumulation to some extent.

Influence of different components. The above experiment shows that decomposing the input into two components and processing them with the proposed SD blocks brings much improvement. We also investigate the relative importance of these two components by imposing different levels of supervision on the reconstruction of two components. It implies that the relative supervision strength applied to different components also plays an important role in the

Vid4	#Frame	FLOPs	#Param.	Calendar (Y)	City (Y)	Foliage (Y)	Walk (Y)	Average (Y)	Average (RGB)
Bicubic	1	N/A	N/A	18.83/0.4936	23.84/0.5234	21.52/0.4438	23.01/0.7096	21.80/0.5426	20.37/0.5106
SPMC † [25]	3	-	-	-/-	-/-	-/-	-/-	25.52/0.76	-/-
Liu† [21]	5	-	-	21.61/-	26.29/-	24.99/-	28.06/-	25.23/-	-/-
TOFlow [27]	7	0.81T	1.41M	22.29/0.7273	26.79/0.7446	25.31/0.7118	29.02/0.8799	25.85/0.7659	24.39/0.7438
DUF-52L [12]	7	0.62T	5.82M	24.17/0.8161	28.05/0.8235	26.42/0.7758	30.91/0.9165	27.38/0.8329	25.91/0.8166
RBPN [7]	7	9.30T	12.2M	24.02/0.8088	27.83/0.8045	26.21/0.7579	30.62/0.9111	27.17/0.8205	25.65/0.7997
EDVR-L† [26]	7	0.93T	20.6M	24.05/0.8147	28.00/0.8122	26.34/0.7635	31.02/0.9152	27.35/0.8264	25.83/0.8077
PFNL† [29]	7	0.70T	3.00M	23.56/0.8232	28.11/0.8366	26.42/0.7761	30.55/0.9103	27.16/0.8365	25.67/0.8189
TGA [10]	7	0.23T	5.87M	24.50/0.8285	28.50/0.8442	26.59/0.7795	30.96/0.9171	27.63/0.8423	26.14/0.8258
FRVSR 10-128 [23]	recurrent (2)	0.14T	5.05M	22.67/0.7844	27.70/0.8063	25.83/0.7541	29.72/0.8971	26.48/0.8104	25.01/0.7917
RLSP 7-256 [4]	recurrent (3)	0.09T	4.21M	24.36/0.8235	28.22/0.8362	26.66/0.7821	30.71/0.9134	27.48/0.8388	25.69/0.8153
RSDN 5-128	recurrent (2)	0.08T	3.83M	24.34/0.8242	28.73/0.8374	26.66/0.7842	30.73/0.9149	27.61/0.8402	26.13/0.8238
RSDN 7-128	recurrent (2)	0.10T	5.01M	24.46/0.8305	29.01/0.8480	26.78/0.7921	30.92/0.9189	27.79/0.8474	26.30/0.8314
RSDN 9-128	recurrent (2)	0.13T	6.19M	24.60/0.8355	29.20/0.8527	26.84/0.7931	31.04/0.9210	27.92/0.8505	26.43/0.8349

Table 3. Quantitative comparison (PSNR (dB) and SSIM) on **Vid4** for 4× video super-resolution. Red text indicates the best and blue text indicates the second best performance. Y and RGB indicate the luminance and RGB channels, respectively. FLOPs (MAC) are calculated on an HR image of size 720×480. '†' means the values are either taken from paper or calculated using provided models.

UDM10	**Bicubic**	**TOFlow** [27]	**DUF-52L** [12]	**RBPN** [7]	**PFNL†** [29]	**FRVSR 10-128** [23]	**RLSP 7-256** [4]	**RSDN 7-128**	**RSDN 9-128**
FLOPs [TMAC]	N/A	2.17	1.65	24.81	1.88	0.36	0.24	0.28	0.35
Runtime [ms]	N/A	1693	1413	3567	295	137	49	79	94
Average (Y)	28.47/0.8523	36.26/0.9438	38.48/0.9605	38.66/0.9596	38.74/0.9627	37.09/0.9522	38.48/0.9606	39.13/0.9645	39.35/0.9653
Average (RGB)	27.05/0.8267	34.46/0.9298	36.78/0.9514	36.53/0.9462	36.78/0.9514	35.39/0.9403	36.39/0.9465	37.26/0.9548	37.46/0.9557
Vimeo-90K-T	**Bicubic**	**TOFlow** [27]	**DUF-52L** [12]	**RBPN** [7]	**EDVR-L†** [26]	**FRVSR 10-128** [23]	**RLSP 7-256** [4]	**RSDN 7-128**	**RSDN 9-128**
FLOPs [TMAC]	N/A	0.27	0.20	3.08	0.30	0.04	0.03	0.03	0.04
Runtime [ms]	N/A	215	167	470	99	28	11	13	15
Average (Y)	31.30/0.8687	34.62/0.9212	36.87/0.9447	37.20/0.9458	37.61/0.9489	35.64/0.9319	36.49/0.9403	37.05/0.9454	37.23/0.9471
Average (RGB)	29.77/0.8490	32.78/0.9040	34.96/0.9313	35.39/0.9340	35.79/0.9374	33.96/0.9192	34.56/0.9274	35.14/0.9325	35.32/0.9344

Table 4. Quantitative comparison (PSNR(dB) and SSIM) on **UDM10** and **Vimeo-90K-T** for 4× video super-resolution, respectively. Flops and runtimes are calculated on an HR image size of 1280×720 and 448×256 for UDM10 and Vimeo-90K-T, respectively. Red text indicates the best and blue text indicates the second best performance. Y and RGB indicate the luminance and RGB channels, respectively. '†' means the values are either taken from paper or calculated using provided models.

super-resolution performance. As shown in Tab. 2, when the weights for structure component, detail component and the whole frame are set to $(\alpha, \beta, \gamma) = (1, 1, 1)$, it achieves a good performance of 27.79/0.8474 in PSNR/SSIM. The performance degrades when the weigh for structure component more than the weight for detail component ($i.e.(\alpha, \beta, \gamma) = (1, 0.5, 1)$), and verse vise ($i.e.(\alpha, \beta, \gamma) = (0.5, 1, 1)$). The result of $(1, 1, 0)$ is 0.06dB lower than that of $(1, 1, 1)$, which means applying additional supervision on the combined image helps the training of the model.

4.3 Comparison with State-of-the-arts

In this section, we compare our methods with several state-of-the-art VSR approaches, including SPMC [25], TOFlow [27], Liu [21], DUF [7], EDVR [26], PFNL [29], TGA [10], FRVSR [23] and RLSP [4]. The first seven methods super-resolve a single reference within a temporal sliding window. Among these methods, SPMC, TOFlow, Liu, RBPN and EDVR need to explicitly estimate the motion between the reference frame and other frames within the window, which requires redundant computation for several frames. DUF, PFNL and TGA skip the motion estimation process and partially ameliorate this issue. The last two

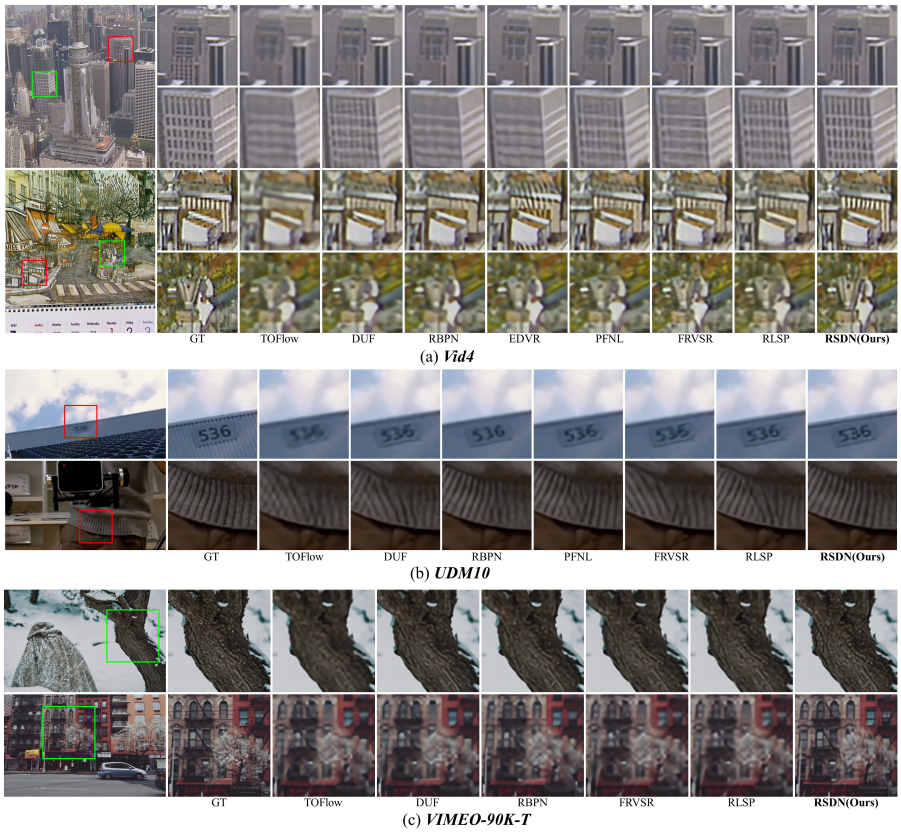

Fig. 7. Qualitative comparison on **Vid4**, **UDM10** and **Vimeo-90K-T** test set for 4× SR. Zoom in for better visualization.

methods FRVSR and RLSP super-resolve each frame in a recurrent way and are more efficient. We carefully implement most of these methods either on our own or by running the publicly available code, and manage to reproduce the results in their paper. The quantitative result of state-of-the-art methods on Vid4 is shown in Tab. 3, where the number is either reported in the original papers or computed with our implementation. In addition, we also include the number of parameters and FLOPs for most methods when super-resolution is conducted on an LR image of size 112×64 in Tab. 3.

On Vid4, our model with only 5 SD block achieves 27.61dB PSNR in Y channel and 26.13dB PSNR in RGB channels, which already outperforms most of the previous methods by a large margin. By increasing the number of SD block to 7 and 9, our methods respectively gain another 0.18dB and 0.31dB PSNR in Y channel while with only a little increase in FLOPs. We also evaluate our method on other three popular test sets. The quantitative results on UDM10 [29] and Vimeo-90K-T [27] two datasets are reported in Tab. 4. Our method achieves

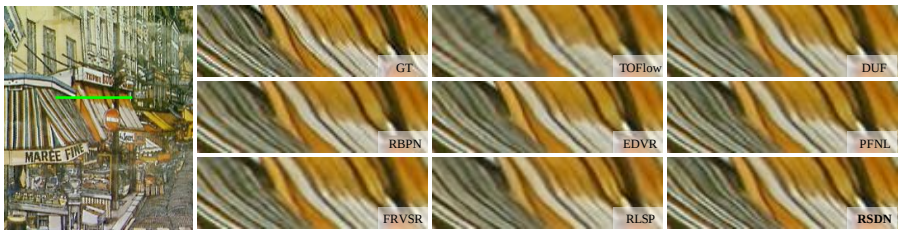

Fig. 8. Visualization of temporal profile for the green line on the calendar sequence.

a very good balance between reconstruction performance and speed on these datasets. On UDM10 test set, RSDN 9-128 achieves new state-of-the-art, and is about 15 and 37 times faster than DUF and RBPN, respectively. RSDN 9-128 outperforms the recent proposed PFNL, where this dataset is proposed by 0.61dB in PSNR in Y channel while being 3 times faster. The proposed method is also evaluated on Vimeo-90K-T, which only contains 7-frame in each sequence. In this case, although our method can not take full of its advantage because of the short length of the sequence, it only lags behind the large model EDVR-L but is 6 times faster.

We also show the qualitative comparison with other state-of-the-art methods. As shown in Fig. 7, our method produces higher quality high-resolution images on all three datasets, including finer details and sharper edges. Other methods are either prone to generate some artifacts (e.g., wrong stripes in an image) or can not recover missing details (e.g., small windows of the building). We also examine temporal consistency of the video super-resolution results in Fig. 8, which is produced by extracting a horizontal row of pixels at the same position from consecutive frames and stacking them vertically. The temporal profile produced by our method is not only temporally smoother but also much sharper, satisfying both requirements of the video super-resolution task.

5 Conclusion

In this work we have presented an effective and efficient recurrent network to super-resolve a video in a streaming manner. The input is decomposed into structure and detail components and fed to two interleaved branches to respectively reconstruct the corresponding components of high-resolution frames. Such a strategy allows our method to address different difficulties in the structure and detail components and to enjoy flexible supervision applied to each components for good performance. In addition we find that hidden state in a recurrent network captures different typical appearance of a scene over time and selectively using information from hidden state can enhance its robustness to appearance change and error accumulation. Extensive experiments on several benchmark datasets demonstrate its superiority in terms of both effectiveness and efficiency.

References

1. Caballero, J., Ledig, C., Aitken, A., Acosta, A., Totz, J., Wang, Z., Shi, W.: Real-time video super-resolution with spatio-temporal networks and motion compensation. In: CVPR (2017)
2. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: ECCV (2014)
3. Du, W., Wang, Y., Qiao, Y.: Rpan: An end-to-end recurrent pose-attention network for action recognition in videos. In: CVPR (2017)
4. Fuoli, D., Gu, S., Timofte, R.: Efficient video super-resolution through recurrent latent space propagation. CoRR **abs/1909.08080** (2019)
5. Glorot, X., Bordes, A., Bengio, Y.: Deep sparse rectifier neural networks. In: AISTATS (2011)
6. Haris, M., Shakhnarovich, G., Ukita, N.: Deep back-projection networks for super-resolution. In: CVPR (2018)
7. Haris, M., Shakhnarovich, G., Ukita, N.: Recurrent back-projection network for video super-resolution. In: CVPR (2019)
8. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR (2017)
9. Huang, Y., Wang, W., Wang, L.: Bidirectional recurrent convolutional networks for multi-frame super-resolution. In: NeurIPS (2015)
10. Isobe, T., Li, S., Jia, X., Yuan, S., Slabaugh, G., Xu, C., Li, Y.L., Wang, S., Tian, Q.: Video super-resolution with temporal group attention. In: CVPR (2020)
11. Jia, X., De Brabandere, B., Tuytelaars, T., Gool, L.V.: Dynamic filter networks. In: NeurIPS (2016)
12. Jo, Y., Wug Oh, S., Kang, J., Joo Kim, S.: Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In: CVPR (2018)
13. Kappeler, A., Yoo, S., Dai, Q., Katsaggelos, A.K.: Video super-resolution with convolutional neural networks. IEEE Transactions on Computational Imaging **2**(2), 109–122 (2016)
14. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: CVPR (2016)
15. Kim, J., Lee, J.K., Lee, K.M.: Deeply-recursive convolutional network for image super-resolution. In: CVPR (2016)
16. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
17. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Fast and accurate image super-resolution with deep laplacian pyramid networks. IEEE transactions on pattern analysis and machine intelligence **41**(11), 2599–2613 (2018)
18. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: CVPR (2017)
19. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: CVPR Workshops (2017)
20. Liu, C., Sun, D.: On bayesian adaptive video super resolution. IEEE transactions on pattern analysis and machine intelligence **36**(2), 346–360 (2013)
21. Liu, D., Wang, Z., Fan, Y., Liu, X., Wang, Z., Chang, S., Huang, T.: Robust video super-resolution with learned temporal dynamics. In: ICCV (2017)
22. Pan, J., Liu, S., Sun, D., Zhang, J., Liu, Y., Ren, J., Li, Z., Tang, J., Lu, H., Tai, Y.W., et al.: Learning dual convolutional neural networks for low-level vision. In: CVPR (2018)

23. Sajjadi, M.S., Vemulapalli, R., Brown, M.: Frame-recurrent video super-resolution. In: CVPR (2018)
24. Singh, B., Marks, T.K., Jones, M., Tuzel, O., Shao, M.: A multi-stream bi-directional recurrent neural network for fine-grained action detection. In: CVPR (2016)
25. Tao, X., Gao, H., Liao, R., Wang, J., Jia, J.: Detail-revealing deep video super-resolution. In: ICCV (2017)
26. Wang, X., Chan, K.C., Yu, K., Dong, C., Change Loy, C.: Edvr: Video restoration with enhanced deformable convolutional networks. In: CVPR Workshops (2019)
27. Xue, T., Chen, B., Wu, J., Wei, D., Freeman, W.T.: Video enhancement with task-oriented flow. International Journal of Computer Vision **127**(8), 1106–1125 (2019)
28. Yang, W., Zhang, X., Tian, Y., Wang, W., Xue, J.H., Liao, Q.: Deep learning for single image super-resolution: A brief review. IEEE Transactions on Multimedia **21**(12), 3106–3121 (2019)
29. Yi, P., Wang, Z., Jiang, K., Jiang, J., Ma, J.: Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In: ICCV (2019)
30. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: ECCV (2018)
31. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: CVPR (2018)