electronics

MDPI

Article

Bidirectional Temporal-Recurrent Propagation Networks for Video Super-Resolution

Lei Han, Cien Fan, Ye Yang and Lian Zou *

School of Electronic Information, Wuhan University, Wuhan 430072, China; 2013301200216@whu.edu.cn (L.H.); fce@whu.edu.cn (C.F.); yangye@whu.edu.cn (Y.Y.)

* Correspondence: zoulian@whu.edu.cn; Tel.: +86-13971579950

Received: 10 November 2020; Accepted: 3 December 2020; Published: 7 December 2020

Abstract: Recently, convolutional neural networks have made a remarkable performance for video super-resolution. However, how to exploit the spatial and temporal information of video efficiently and effectively remains challenging. In this work, we design a bidirectional temporal-recurrent propagation unit. The bidirectional temporal-recurrent propagation unit makes it possible to flow temporal information in an RNN-like manner from frame to frame, which avoids complex motion estimation modeling and motion compensation. To better fuse the information of the two temporal-recurrent propagation units, we use channel attention mechanisms. Additionally, we recommend a progressive up-sampling method instead of one-step up-sampling. We find that progressive up-sampling gets better experimental results than one-stage up-sampling. Extensive experiments show that our algorithm outperforms several recent state-of-the-art video super-resolution (VSR) methods with a smaller model size.

Keywords: convolutional neural network; video super-resolution; temporal-recurrent propagation; progressive up-sampling

1. Introduction

Super-resolution (SR) is a class of image processing techniques that generates a high-resolution (HR) image or video from its corresponding low-resolution (LR) image or video. SR is widely used in various fields, such as surveillance imaging [1], medical imaging [2], and satellite imaging [3]. With the improvement of display technology, the video super-resolution (VSR) becomes more and more critical for LR video.

Recently, neural networks have made remarkable achievements in the single-image super-resolution (SISR) [4–8]. One way to perform VSR is to run the SISR frame by frame. However, SISR methods do not consider the inter-frame temporal relationship. The output HR videos usually lack temporal consistency, which results in the flickering artifact [9]. Most existing VSR methods [10–14] consist of similar steps: motion estimation and compensation, feature fusion and up-sampling. They usually use optical flow to estimate the motion between the reference frame and supporting frames, and then align all other frames to the reference with warping operations. Therefore, the results of these methods depend heavily on the accuracy of optical flow estimation. Inaccurate motion estimation and alignment may introduce artifacts around image structures in the aligned supporting frames. Furthermore, it takes a lot of computational resources to compute the optical flow on every pixel between frames.

To alleviate the above issues, we propose an end to end bidirectional temporal-recurrent propagation network (BTRPN). We design a bidirectional temporal-recurrent propagation unit (BTRP unit). The BTRP unit can implicitly utilize motion information without explicit estimation and alignment. Therefore, the reconstructed HR video frames will have fewer artifacts due to inaccurate motion estimation and alignment. In addition, instead of using multiple consecutive video frames

to predict an intermediate frame, we use the reconstruction results of the previous frame to predict the next frame in an RNN-like manner. Consider that a frame is associated with its before and after frames, we exploit a bidirectional network to fully extract temporal information. One subnetwork processes the positive sequence on the time axis, the other processes the reverse sequence. To better integrate the forward and backward TRP unit, we take the channel attention mechanism [15] to better fuse the extracted temporal and spatial information. Additionally, we use a progressive up-sampling method [16] to replace one-step up-sampling.

Experimental results on the widely-used VSR benchmark: VID4 [17] show that our network achieves promising performance beyond 0.4 dB improvements in terms of signal-to-noise ratio (PSNR) over recent methods using optical flow such as DRVSR [11], FRVSR [14]. Compared to recent implicit frame alignment methods such as RBPN [18], DUF [19], we also outperform them in terms of PSNR and structural similarity index (SSIM) [20].

The contributions of this paper are summarized as follows:

1. We propose a novel end to end bidirectional temporal-recurrent propagation network, which avoids the complicated combination network of optical estimation and super-resolution. To better integrate the two subnetworks, we take the channel attention mechanism to fuse the extracted temporal and spatial information.
2. We propose a progressive up-sampling version of BTRPN. Compared to one-step up-sampling, progressive up-sampling means solving the SR optimization issue in a small solution space, which decreases the difficulty of learning and boosts the performance of reconstructed images.

2. Related Work

2.1. Single-Image Super-Resolution

Since Dong et al. first proposed the SRCNN [21], neural networks have made promising achievements in SISR. New improvements included sub-pixel convolution [9], residual learning [22], recursive layers with skip connection [23], back-projection [24]. Recently, state-of-the-art SISR networks [24–26] outperformed previous works by a large margin when trained on the DIV2K [27]. A recent survey was conducted in [28]. Many VSR methods use sub-pixel convolution [29,30] for up-sampling and residuals [18,19] for feature extraction.

2.2. Video Super-Resolution

Temporal alignment, either explicitly or implicitly, plays an essential role in the performance of VSR. Previous explicit methods, such as [10], split temporal alignment into two stages. They compute optical flow in the first stage and perform motion compensation in the second stage. VESCPN [31] is the first end-to-end VSR network that jointly trains optical flow estimation and spatial-temporal networks. SPMC [11] proposed a new sub-pixel motion compensation layer (SPMC), which can simultaneously achieve sub-pixel motion compensation and resolution enhancement. FRVSR [14] introduced a frame-recurrent structure to process video super-resolution reconstruction, which avoided the repeated redundant operation of the same frame image in some multiple-input VSR methods and improved the computing efficiency of the network. Reference [12] achieved temporal alignment through a proposed task-oriented flow (ToFlow), which achieved better VSR results than fixed flow algorithms. However, all these methods rely on the accuracy of optical flow estimation. At present, even state-of-the-art optical flow estimation algorithms are not easy to obtain sufficient high-quality motion estimation. Even with accurate motion fields, the image warping for motion compensation will also produce artifacts around the LR frames, which may affect the final reconstructed HR frames. Our proposed BTRPN performs an implicit temporal alignment without depending on optical flows, which will alleviate the issues caused by optical flow based methods.

Recently, some implicit algorithms were proposed. Reference [32] exploited a 3D convolution-based residual network for VSR instead of explicit motion alignment. Reference [19] proposed a dynamic filter

network for VSR. Reference [30,33] utilized deformable convolution to perform temporal alignment. These methods used implicit temporal assignment and avoided the issues in optical flow. However, they all used seven consecutive input frames to predict an intermediate frame, which led to huge training costs.

The work most related to ours is FRVSR [14], which also used frame-recurrent. However, in [14], the optical flow was used for explicit motion estimation, which may lead to artifacts around the image structure. In addition, our BTRPN uses a bidirectional structure, which ensures full utilization of temporal information. Compared to [14], our method achieves better VSR results with a smaller network.

3. The Progressive Up-Sampling Bidirectional Temporal-Recurrent Propagation Network

3.1. Network Architecture

The overall network framework is shown in Figure 1. The structure and weight of the two subnetworks of BTRPN are precisely the same, and they process, respectively, the positive and reverse LR input video on the time axis, thus allowing the bidirectional temporal information flow. In the TRP unit, the method of progressive sampling is adopted to avoid the large one-step scale sampling. At the end of BTRPN is a fusion module with the channel attention mechanism, through which the features of the two sub-networks are combined, and the reconstructed video frames are output.

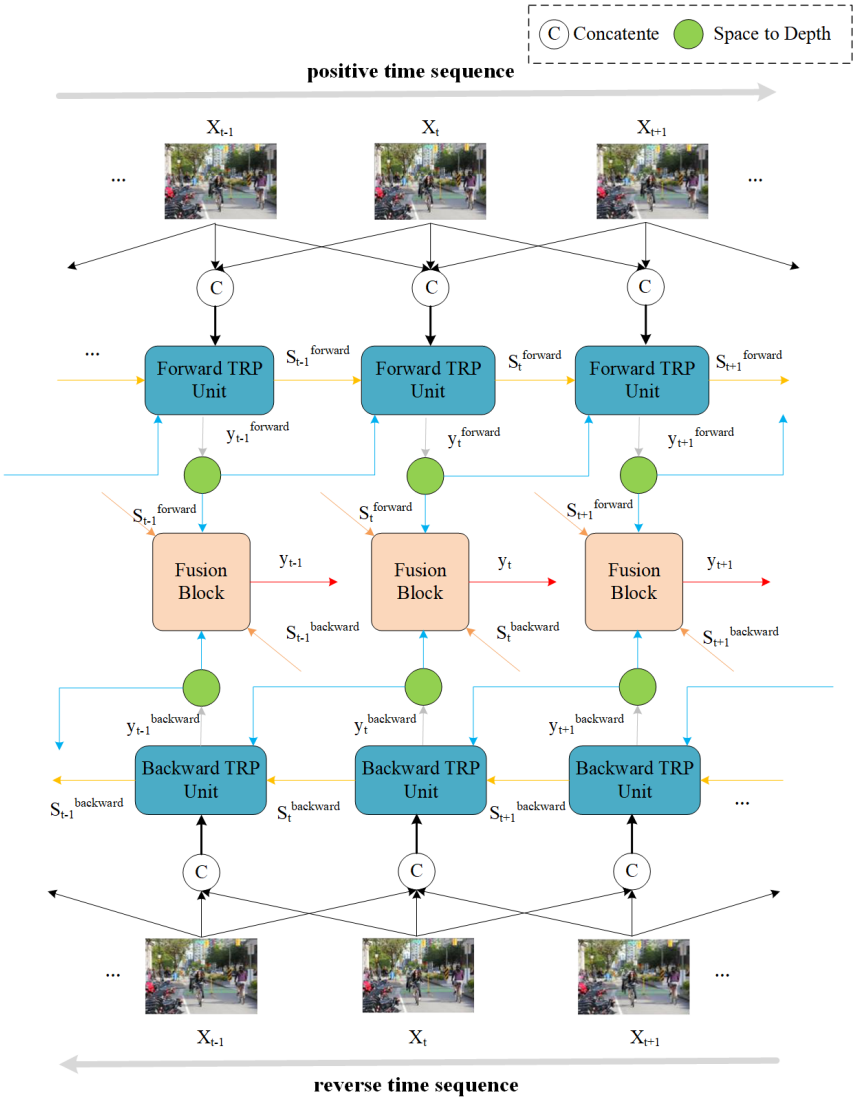

Figure 1. Bidirectional temporal-recurrent propagation network (BTRPN) network architecture.

3.2. TRP Unit

The TRP unit is illustrated in Figure 2. The input of the TRP unit is composed of a three section cascade: consecutive video frames $X_{t-1:t+1}$ (the current frame is in the middle), the temporal status of the last moment S_{t-1}, the result of Space to Depth processing of the reconstructed output from the last moment y_{t-1}. The output of the TRP unit is the temporal status of the current moment S_t and the SR result of the current frame y_t.

$$S_t, y_t = TRP(X_{t-1:t+1}, S_{t-1}, Space2Depth(y_{t-1})) \tag{1}$$

The TRP unit is composed of two branches, which output temporal status S_t and SR reconstruction results y_t, respectively. These two branches share the feature extraction module, which consists of multiple convolutional layers followed by the Rectified Linear Unit (Relu) activation layers. The Relu activation layers can make convergence much faster while still present good image quality. The branch of output y_t can be regarded as a residual network with the number of channels r^2. The output of the residual network is up-sampled through Depth to Space to obtain the reconstructed frame y_t. Space to Depth is the inverse of Depth to Space proposed by FRVSR [14]. It is illustrated in Figure 3.

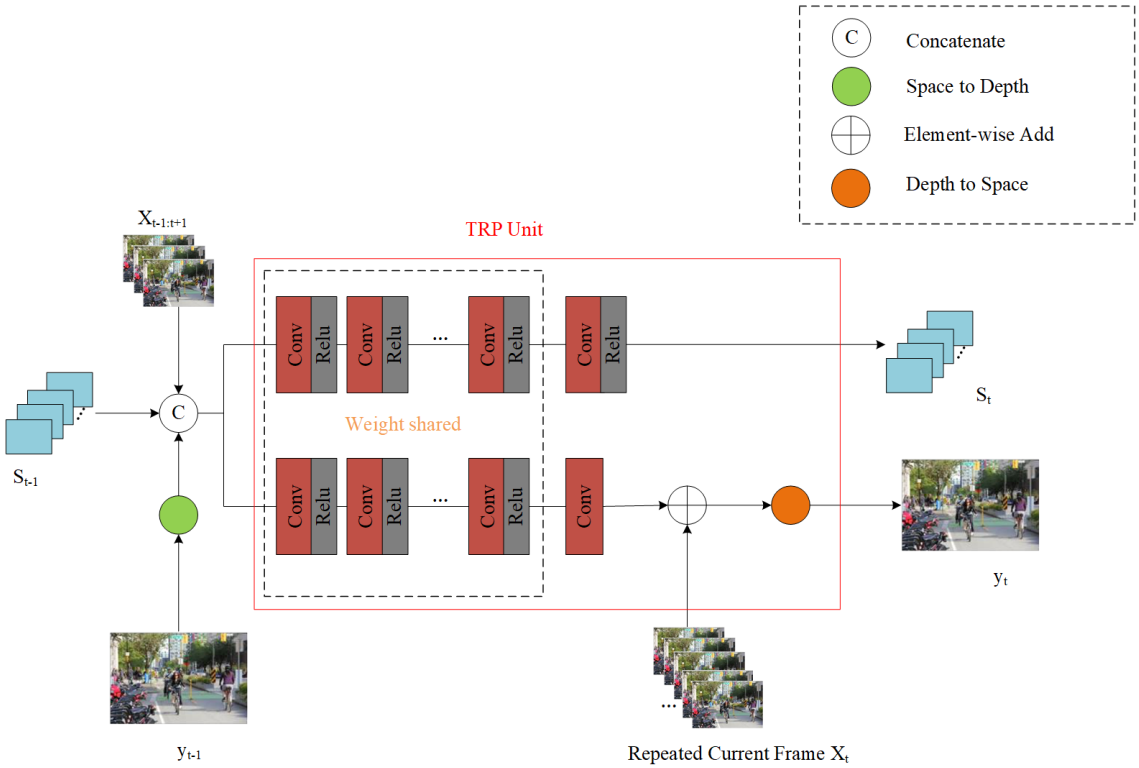

Figure 2. The architecture of the proposed TRP (Temporal-Recurrent Propagation) unit.

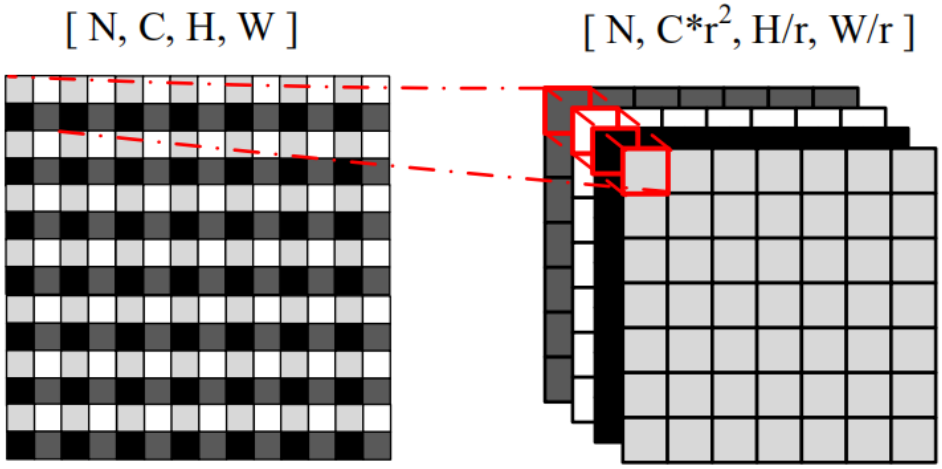

Figure 3. Illustration of the Space-to-Depth transformation.

3.3. Bidirectional Network

We found that VSR results obtained by positive and reverse sequence input are different, as Table 1 shows. We set up a small FRVSR [14] network: FRVSR 3-64. The optical flow estimation is obtained by stacking three residual blocks, and the number of channels in the hidden layer is 64. After the FRVSR 3-64 was trained to convergence, we tested the model on the VID4 dataset and recorded results. Then we processed the four videos on the VID4 dataset in reverse order on the timeline and tested the model on the reverse VID4 dataset. From Table 1, we can see the difference between the forward and reverse processing of the same video. Only unilateral information flows can provide limited temporal information. So we designed a bidirectional network to extract inter-frame temporal information fully.

Table 1. FRVSR3-64 SR results on the positive and reverse VID4 dataset.

		Calendar		**City**		**Foliage**		**Walk**		**Average**	
Time Axis	**Scale**	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**
positive	4	22.86	0.754	27.07	0.774	25.46	0.722	29.23	0.889	26.16	0.785
reverse	4	22.77	0.749	27.08	0.775	25.24	0.714	29.26	0.889	26.09	0.782

As shown in Figure 4, the network consists of two sub-networks, which input videos on the forward and backward time axis, respectively. Then the two subnetworks are combined by a fusion module to obtain the final output. The two subnetworks are identical in structure and parameters, distinguished only by the timeline order of the input video.

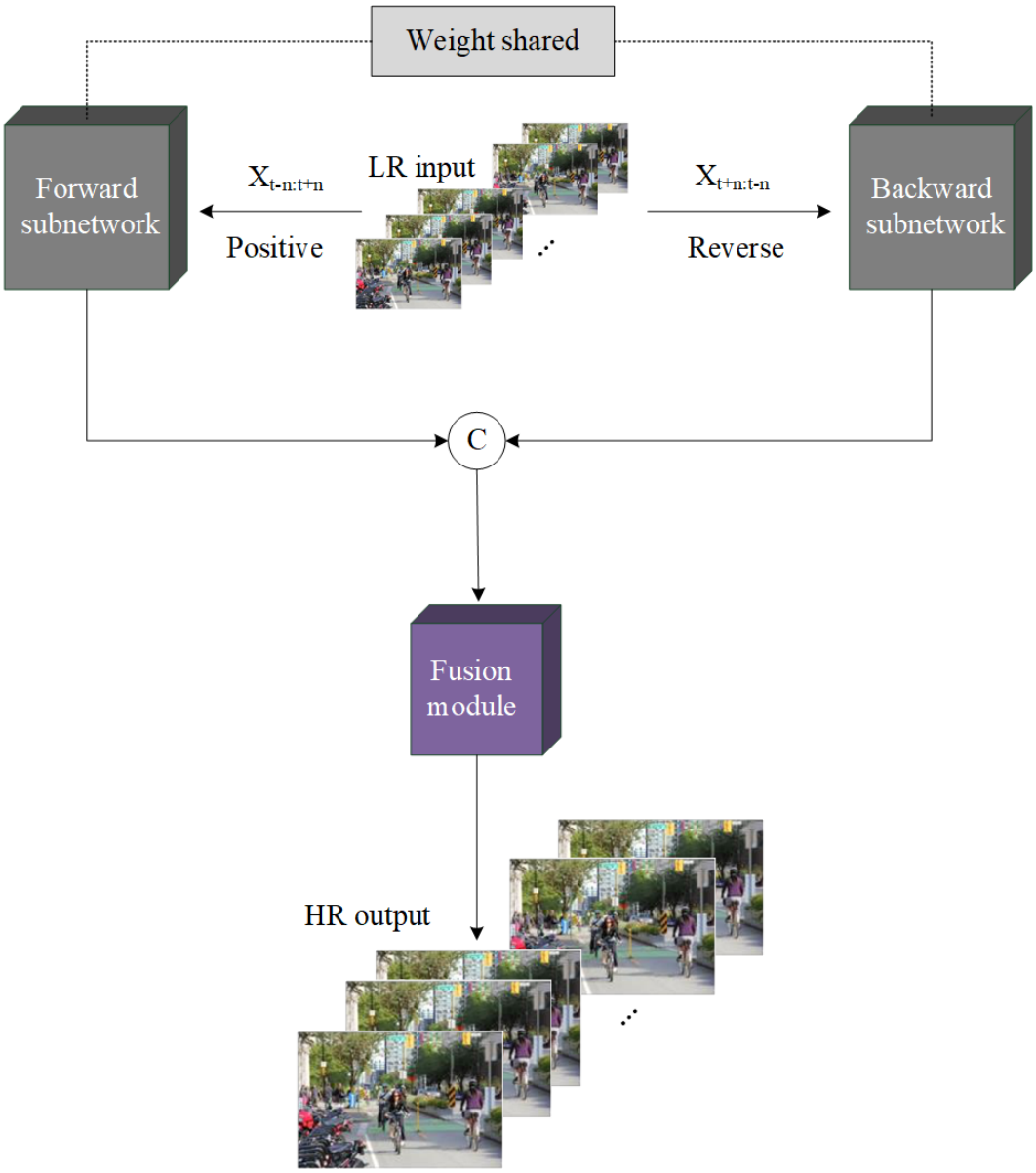

Figure 4. The bidirectional recurrent transmission mechanism.

3.4. Attentional Mechanism

We use a channel attention fusion module [15] in Figure 5 to fuse features from the two subnetworks. The attention mechanism [15] has been applied to the SISR and improved SR performance. Attention can be viewed as a guidance to bias the allocation of available processing resources towards the most informative components of an input. In our network, the output features of the two subnetworks can be regarded as a set of vectors in the space based on local priors. These different channels of the two subnetworks feature vectors contain different information, and different channels have different effects on the SR results. Adding channel attention in the fusion module in Figure 6 can help the network adaptively rescale and adjust the features of the channel so that the network can focus on more informative features and get better SR results. In the fusion module, firstly, features are scaled by an attention mechanism after the two subnetwork outputs are concatenated and input into two 3×3 kernel 2D convolutional layers. Secondly, the scaled features are added to the original input at the pixel level. Thirdly, the second-step result is channel-compressed through a 1×1 kernel convolutional layer.

Finally, the channel compressed result performs Depth to Space to get the final SR result. The formula for the whole process is as follows:

$$I_{SR} = D2S(W_3^{1\times1} \times (F^{input} + CA(W_2^{3\times3} \times (ReLu(W_1^{3\times3} \times F^{input}))))) \quad (2)$$

In the formula, F^{input} and I_{SR}, respectively, represent the input feature vectors of the fusion module and the final reconstruction results. CA(·) represents the channel attention mechanism. ReLu(·) represents ReLU nonlinear activation unit. D2S(·) represents Depth to Space. W represents the weight matrix. Subscripts 1, 2 and 3, respectively, represent the three convolutional layers from shallow to deep in the fusion module, and the superscript represents the size of the convolution kernel.

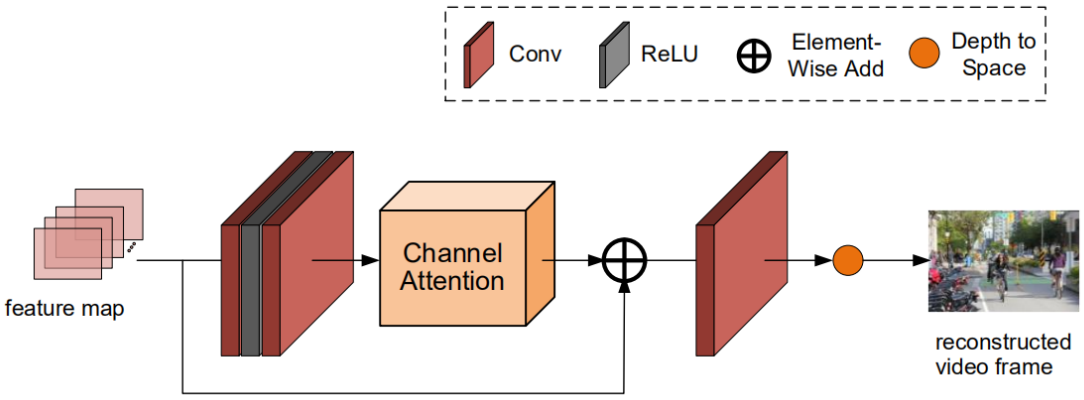

Figure 5. Fusion module.

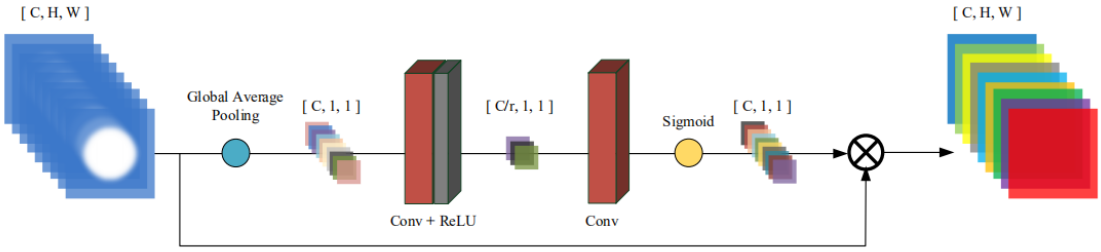

Figure 6. Channel attention mechanism with scaling factor r.

3.5. Progressive Up-Sampling

For the image SR work, one-step mapping on the large scale factor means that the optimization solution will be carried out in a more extensive solution space compared with small scale factors, which will increase the difficulty of model learning and affect the final image. Network design needs to avoid one-time large scale mapping, as much as possible in the form of multiple small scales (2×) mapping. Therefore, we propose a progressive improved TRP version, as Figure 7 shows. For 4× enhancement, We used two 2× TRP units level 1 and level 2. The TRP Unit level 1 is the same as Section 3.2 describes. The input of TRP Unit level 2 is the consecutive video frames $X_{t-1:t+1}$ after interpolation, the temporal status of the last moment S_{t-1} after interpolation, the result of Space to Depth processing of TRP Unit level 1 SR output $X_{t-1} \uparrow_2$. The final output is 4× SR result y_t.

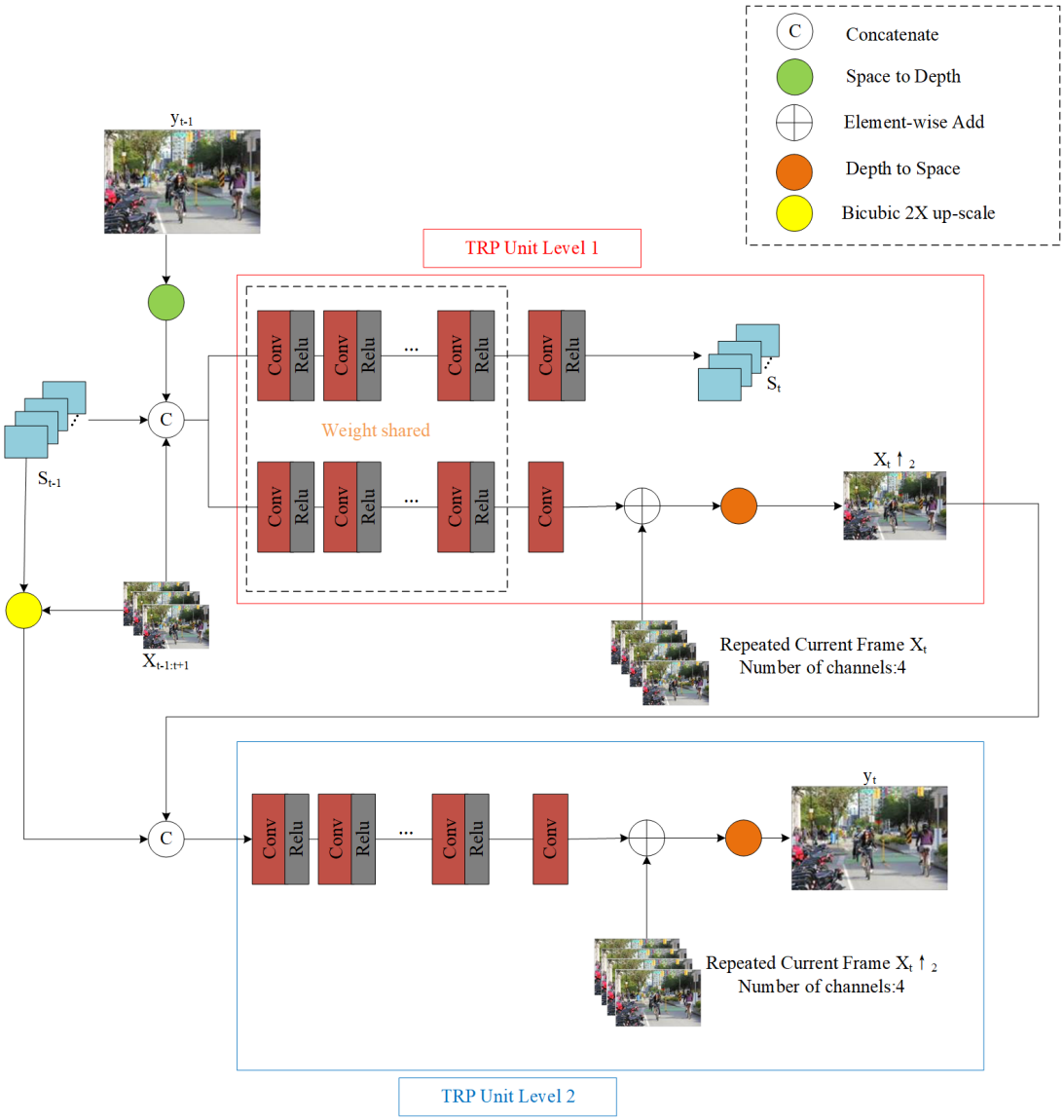

Figure 7. Progressive up-sampling TRP unit for 4× video super-resolution (VSR).

4. Experiments

4.1. Datasets and Training Details

We train all the networks using videos from the REDS [34] dataset proposed by NTIRE2019. REDS consists of 300 high-quality (720p) clips: 240 training clips, 30 validation clips, and 30 testing clips (each with 100 consecutive frames). We use the training clips and validation clips as a training dataset (270 clips). Limited by the device, the REDS raw data need to be compressed and sampled randomly before training to ensure storage space. Under the 4× VSR task, we firstly compress all the original videos of 1280 × 720 into 960 ×540 as the training HR videos. Then the videos are down-sampled four times to obtain the input LR videos of 240 × 135. The image resize function in Matlab (imresize) completes the above sampling operation. Furthermore, the training data of the original 100 consecutive video frames only takes the first 30 frames for the training to further save space.

We set the batch size as 8 with size 128 × 128 for HR patches. We set the learning rate as 1×10^{-4} and decrease it by a factor of 10 for every 200 K iterations for a total of 400 K iterations. We initialize all the weights based on a Xavier Initialization. For all the activation units following the convolutional layers, we use ReLu. We use Adam [35] with a momentum of 0.9 and weight decay of 1×10^{-4} for the

optimization. We use Huber loss [36] as the loss function for training BTRPN referring to DUF-VSR [19]. The expression for Huber Loss is as follows:

$$L(\hat{Y}, Y) = \begin{cases} \frac{1}{2}||\hat{Y} - Y||_2^2 & ||\hat{Y} - Y|| < \delta \\ \delta||\hat{Y} - Y|| - \frac{1}{2}\delta^2 & \text{otherwise} \end{cases} \quad (3)$$

When training, the δ is set as 0.01. All experiments are performed using Python3.7 and Pytorch1.1.0 on a 2.1 GHz CPU and NVIDIA 1080Ti GPU. All tensors involved in the training and testing process are interpolated using the bilinear interpolation function provided in the Pytorch. According to mainstream practice, training and testing are conducted only on the Y channel of the YCbCr space, PSNR and SSIM are only calculated on the Y channel.

4.2. Model Analysis

4.2.1. Depth and Channel Analysis

We construct multiple BTRPN networks of different depths and channels: BTRPN10-64, BTRPN10-128, BTRPN20-64, BTRPN20-128. 10/20 means that the network has 10/20 convolutional layers. 64/128 means that each convolutional layer channel is 64/128. Table 2 shows the performance of the four models for the 4× VSR. We can see that BTRPN20-128 has the best performance. The BTRPN20-64 has double-depth compared with BTRPN10-64, but the performance was not significantly improved. However, the BTRPN10-128 with double numbers of channels compared to BTRPN10-64 has a significant performance improvement. It indicates that it is more useful for the shallow network to increase the channel numbers in each layer than deepen the network.

Table 2. The comparison of different BTRPN models on the VID4 dataset for 4× VSR.

		Calendar		City		Foliage		Walk		Average	
Model	**Scale**	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**
BTRPN10-64	4	23.30	0.780	27.62	0.794	25.91	0.743	30.04	0.897	26.69	0.804
BTRPN20-64	4	23.39	0.786	27.68	0.799	25.99	0.746	30.26	0.900	26.83	0.808
BTRPN10-128	4	23.56	0.794	27.78	0.804	26.15	0.754	30.44	0.904	26.98	0.814
BTRPN20-128	4	**23.69**	**0.804**	**27.84**	**0.811**	**26.37**	**0.766**	**30.72**	**0.909**	**27.15**	**0.822**

Table 3 records training time for the four models. Table 4 records test time for the four models. Figure 8 shows the convergence rates of different models. Since the BTRPN network is not extensive, almost all BTRPN models converge at 250 K iterations.

Table 3. Time consumption of different BTRPN models for 50,000 iterations.

Model	**Scale**	**Iterations**	**Parameters**	**Training Time**
BTRPN10-64	4	50,000	670 K	40 min
BTRPN20-64	4	50,000	2600 K	45–50 min
BTRPN10-128	4	50,000	1040 K	50–55 min
BTRPN20-128	4	50,000	4070 K	1 h

Table 4. Test time of different BTRPN models.

Model	**Scale**	**Parameters**	**Test Time**
BTRPN10-64	4	670 K	0.016 s
BTRPN20-64	4	2600 K	0.036 s
BTRPN10-128	4	1040 K	0.027 s
BTRPN20-128	4	4070 K	0.066 s

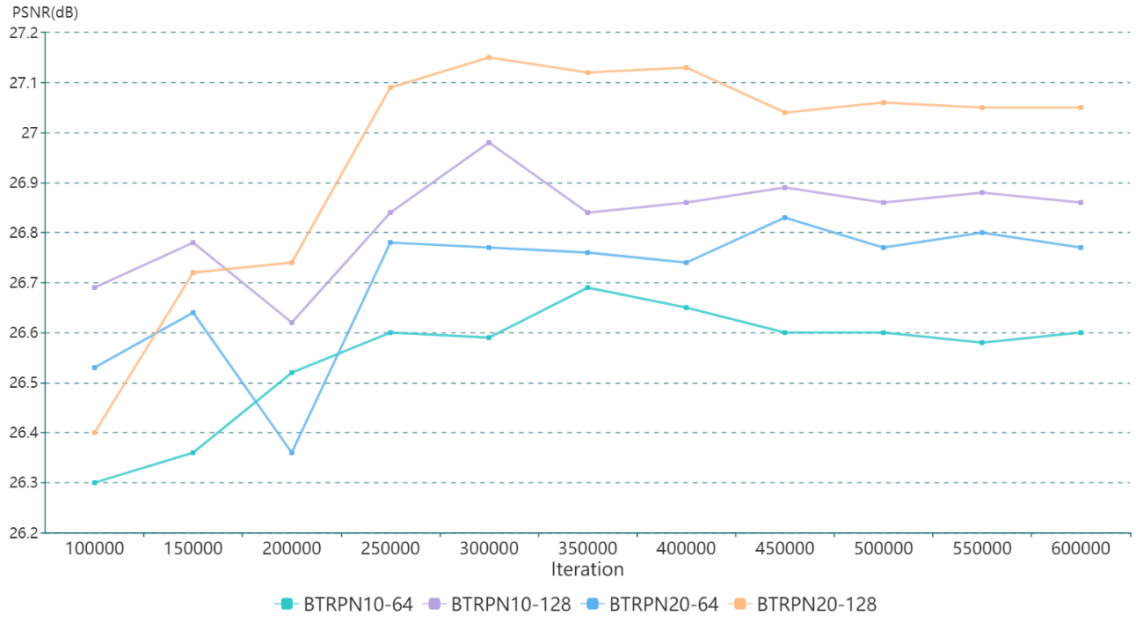

Figure 8. The convergence of different models.

4.2.2. Bidirectional Model Analysis

We test bidirectional and unidirectional models on positive and reverse VID4 dataset. To simplify the experiment and keep the same experimental conditions of the other control groups, we do not use progressive up-sampling TRP unit in Table 5. BTRPN-5L consists of 5 convolutional layers, and the fusion module is simplified to the convolutional layers concatenation. Due to the lack of a fusion module in TRPN, we use seven convolutional layers represented as TRPN-7L to guarantee the parameters consistent with BTRPN-5L. The parameters of TRPN-7L and BTRPN-5L are all around 1070 K. Experiments in Table 5 show that there is a big difference in the VSR results of positive and reverse video sequences for unidirectional TRPN-7L. The average PSNR difference of the VID4 dataset can reach 0.1db, and the maximum PSNR difference of a single video can reach 0.37db. However, the VSR results of the positive and reverse video sequences for bidirectional BTRPN-5L are almost identical. Furthermore, the results of BTRPN-5L on the VID4 dataset are better than those of TRPN-7L in both positive and reverse sequence. These results indicate that the bidirectional temporal network makes more use of temporal information and can reach better SR reconstruction.

Table 5. Results of the bidirectional and non-bidirectional model for 4× VSR on the VID4 dataset.

		Calendar		City		Foliage		Walk		Average	
Model	**Time Axis**	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**
TRPN-7L	positive	23.01	0.766	27.13	0.778	25.63	0.733	29.38	0.891	26.29	0.792
TRPN-7L	reverse	22.92	0.760	27.14	0.779	25.26	0.718	39.41	0.892	26.18	0.787
BTRPN-5L	positive	22.95	0.761	27.23	0.785	25.54	0.728	29.74	0.897	26.36	0.793
BTRPN-5L	reverse	22.95	0.761	27.23	0.785	25.54	0.728	29.73	0.897	26.36	0.793

4.2.3. Attention Mechanism

To demonstrate the effect of the attention mechanism, we use concatenation and channel attention fusion module to deal with the output of the two subnetworks, respectively. Experiments in Table 6 show that channel attention boosts the PSNR from 26.36 db to 26.78 db. This indicates that channel attention can direct the network to focus on more informative features and improve network performance.

Table 6. The influence of the attention mechanism for 4× VSR.

	Calendar		City		Foliage		Walk		Average	
Attention Mechanism	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**
not used	22.95	0.761	27.23	0.785	25.54	0.728	29.74	0.897	26.36	0.793
used	23.34	0.784	27.62	0.796	25.95	0.746	30.20	0.899	26.78	0.807

4.2.4. Progressive Up-Sampling

We test the BTRPN networks using one-step up-sampling and progressive up-sampling. The models in Table 7 both contain ten layers of convolution and maintain the same level of model size. Experiments show that the result of progressive up-sampling is better than that of one-step up-sampling. This shows that progressive up-sampling can indeed help the network achieve better SR performance than one-step up-sampling.

Table 7. The influence of the progressive up-sampling mechanism for 4× VSR.

	Calendar		City		Foliage		Walk		Average	
	PSNR	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**
one-step up-sampling	23.34	0.784	27.62	0.796	25.95	0.746	30.20	0.899	26.78	0.807
progressive up-sampling	23.56	0.794	27.78	0.804	26.15	0.754	30.44	0.904	26.98	0.814

4.3. Comparison with State-of-the-Art Algorithms

4.3.1. Quantitive and Qualitative Comparison

We compare the proposed BTRPN20-128 (referred to as BTRPN in the later section) with several the-state-of-the-art VSR algorithms on the VID4 dataset: VSRNet [10], VESPCN [31], DRVSR [11], Bayesian [17], B1,2,3+T [13], BRCN [37], SOF-VSR [38], FRVSR [14], DUF-16L [19], RBPN [18], RCAN [25]. Table 8 shows that our BTRPN network has the best average PSNR and the best average SSIM on the VID4 dataset. The qualitative result in Figures 9 and 10 also validates the superiority of the proposed method. In the short video of the city, BTRPN restores the clearest building edge lines, which reflects that the BTRPN network with progressive up-sampling has a strong reconstruction ability of regular patterns. In the foliage video, compared with other methods, BTRPN, accurately captures the motion trajectory of the white car and achieves a good motion compensation effect, which again proves the effectiveness of BTRPN's temporal propagation mechanism.

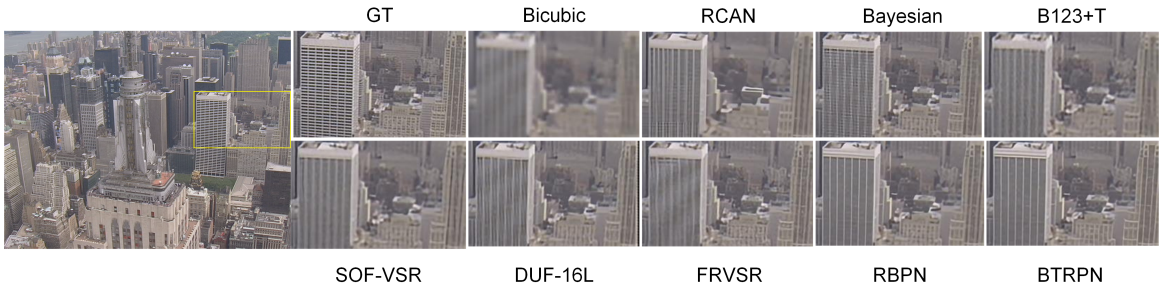

Figure 9. Qualitative comparison on the city clip for 4× video SR. GT is the abbreviation of ground truth.

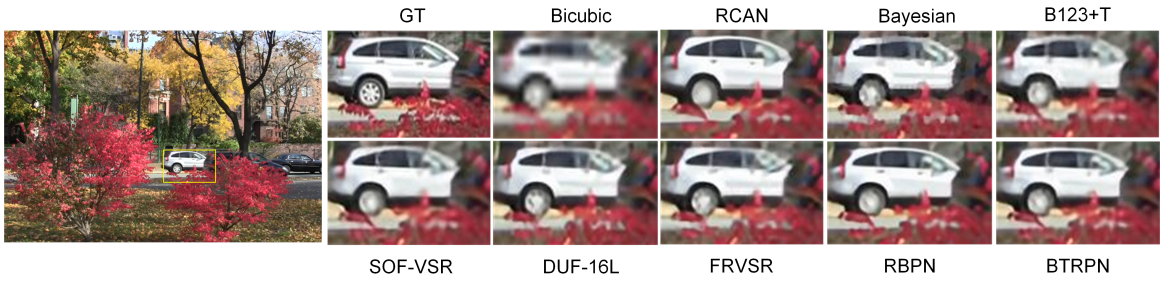

Figure 10. Qualitative comparison on the foliage clip for 4× video SR.

Table 8. Quantitative comparison on the VID4 dataset for 4× video SR. Red and blue indicate the best and second-best performance, respectively.

	Calendar		City		Foliage		Walk		Average	
Algorithm	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**
Bicubic	20.39	0.572	25.16	0.603	23.47	0.567	26.10	0.797	23.78	0.635
RCAN	22.33	0.725	26.10	0.696	24.74	0.665	28.65	0.872	25.46	0.740
VSRNet	-	-	-	-	-	-	-	-	24.84	0.705
VESPCN	-	-	-	-	-	-	-	-	25.35	0.756
DRVSR	22.16	0.747	27.00	0.757	25.43	0.721	28.91	0.876	25.88	0.775
Bayesian	-	-	-	-	-	-	-	-	26.16	0.815
$B_{1,2,3} + T$	21.66	0.704	26.45	0.720	24.98	0.698	28.26	0.859	25.34	0.745
BRCN	-	-	-	-	-	-	-	-	24.43	0.662
SOF-VSR	22.64	0.745	26.93	0.752	25.45	0.718	29.19	0.881	26.05	0.767
FRVSR	-	-	-	-	-	-	-	-	26.69	0.822
DUF-16L	-	-	-	-	-	-	-	-	26.81	0.815
RBPN	23.99	0.807	27.73	0.803	26.22	0.757	30.70	0.909	27.12	0.808
BTRPN	23.69	0.804	27.84	0.811	26.37	0.766	30.72	0.909	27.15	0.822

4.3.2. Parameters and Test Time Comparison

We compare the parameters and test times of BTRPN and other networks. We take the calendar clip in the VID4 dataset with 180 × 135 images input and 720 × 540 HR images output to record the test time for 4× enlargement. Figure 11 shows that BTRPN has achieved a good trade-off between model size and reconstruction effect. BTRPN makes an excellent reconstruction effect at only one-third of the size of the RBPN model. Compared with the same frame-recurrent type of FRVSR, BTRPN also obtains better video reconstruction quality with smaller network capacity. Table 9 shows that BTRPN has a distinct speed advantage compared with other methods.

Table 9. Test time compared to other models for a frame on the calendar clip.

Model	**Scale**	**Test Time**
BRCN	4	0.024 s
SOF-VSR	4	0.120 s
DUF-16L	4	0.420 s
DUF-28L	4	0.500 s
RBPN	4	0.50 s
BTRPN	4	**0.066 s**

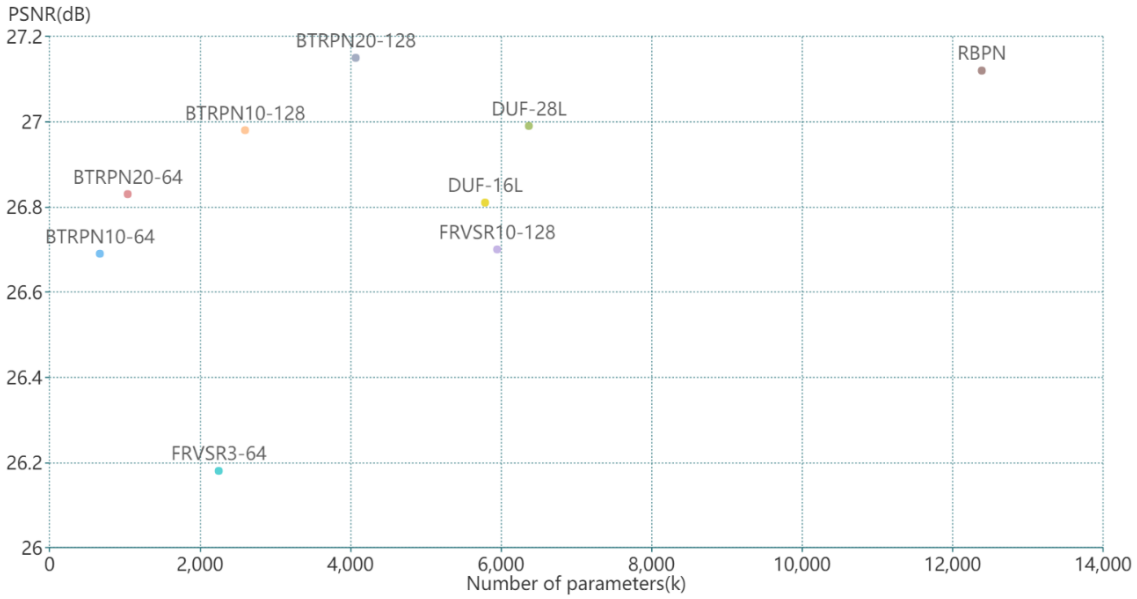

Figure 11. Parameters of different models for 4x VSR on the VID4 dataset.

5. Conclusions

In this paper, we propose a novel bidirectional neural network that can integrate temporal information between frames. To fuse the bidirectional neural network better, we use the channel attention. We also find that progressive up-sampling is better than one-step up-sampling. Extensive experiments on the VID4 dataset demonstrate the effectiveness of the proposed method.

Author Contributions: L.H. and L.Z. completed the main work, including proposing the idea, conducting experiments and writing the paper. C.F. revised the paper. C.F. and Y.Y. collected and analyzed the data. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China Enterprise Innovation and Development Joint Fund (Project No.U19B2004).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhang, L.; Zhang, H.; Shen, H.; Li, P. A super-resolution reconstruction algorithm for surveillance images. *Signal Process.* **2010**, *90*, 848–859. [CrossRef]
2. Greenspan, H. Super-resolution in medical imaging. *Comput. J.* **2009**, *52*, 43–63. [CrossRef]
3. Cao, L.; Ji, R.; Wang, C.; Li, J. Towards Domain Adaptive Vehicle Detection in Satellite Image by Supervised Super-Resolution Transfer. In Proceedings of the AAAI 2016, Phoenix, AZ, USA, 12–17 February 2016; Volume 35, p. 36.
4. 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 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4681–4690.
5. Tong, T.; Li, G.; Liu, X.; Gao, Q. Image super-resolution using dense skip connections. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4799–4807.
6. Tai, Y.; Yang, J.; Liu, X. Image Super-Resolution via Deep Recursive Residual Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
7. Lai, W.S.; Huang, J.B.; Ahuja, N.; Yang, M.H. Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.

8. Zhang, K.; Zuo, W.; Chen, Y.; Meng, D.; Zhang, L. Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising. *IEEE Trans. Image Process.* **2017**, *26*, 3142–3155. [CrossRef] [PubMed]
9. Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A.P.; Bishop, R.; Rueckert, D.; Wang, Z. 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, Las Vegas, NV, USA, 27–30 June 2016; pp. 1874–1883.
10. Kappeler, A.; Yoo, S.; Dai, Q.; Katsaggelos, A.K. Video super-resolution with convolutional neural networks. *IEEE Trans. Comput. Imaging* **2016**, *2*, 109–122. [CrossRef]
11. Tao, X.; Gao, H.; Liao, R.; Wang, J.; Jia, J. Detail-Revealing Deep Video Super-Resolution. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
12. Xue, T.; Chen, B.; Wu, J.; Wei, D.; Freeman, W.T. Video enhancement with task-oriented flow. *Int. J. Comput. Vis.* **2019**, *127*, 1106–1125. [CrossRef]
13. Liu, D.; Wang, Z.; Fan, Y.; Liu, X.; Wang, Z.; Chang, S.; Huang, T. Robust Video Super-Resolution With Learned Temporal Dynamics. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
14. Sajjadi, M.S.M.; Vemulapalli, R.; Brown, M. Frame-Recurrent Video Super-Resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.
15. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.
16. Yang, Y.; Fan, S.; Tian, S.; Guo, Y.; Liu, L.; Wu, M. Progressive back-projection networks for large-scale super-resolution. *J. Electron. Imaging* **2019**, *28*, 033039. [CrossRef]
17. Liu, C.; Sun, D. On Bayesian Adaptive Video Super Resolution. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 346–360. [CrossRef] [PubMed]
18. Haris, M.; Shakhnarovich, G.; Ukita, N. Recurrent back-projection network for video super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3897–3906.
19. Jo, Y.; Oh, S.W.; Kang, J.; Kim, S.J. 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 (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.
20. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
21. Dong, C.; Loy, C.C.; He, K.; Tang, X. Learning a deep convolutional network for image super-resolution. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 184–199.
22. Kim, J.; Kwon Lee, J.; Mu Lee, K. Accurate image super-resolution using very deep convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1646–1654.
23. Kim, J.; Lee, J.K.; Lee, K.M. Deeply-Recursive Convolutional Network for Image Super-Resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
24. Haris, M.; Shakhnarovich, G.; Ukita, N. Deep back-projection networks for super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1664–1673.
25. Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; Fu, Y. Image Super-Resolution Using Very Deep Residual Channel Attention Networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
26. Lim, B.; Son, S.; Kim, H.; Nah, S.; Mu Lee, K. Enhanced deep residual networks for single image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 136–144.
27. Timofte, R.; Agustsson, E.; Gool, L.V.; Yang, M.H.; Guo, Q. NTIRE 2017 Challenge on Single Image Super-Resolution: Methods and Results. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017.

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 Trans. Multimed.* **2019**, *21*, 3106–3121. [CrossRef]
29. 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 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8008–8017.
30. Tian, Y.; Zhang, Y.; Fu, Y.; Xu, C. TDAN: Temporally-Deformable Alignment Network for Video Super-Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3360–3369.
31. 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 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
32. Li, S.; He, F.; Du, B.; Zhang, L.; Xu, Y.; Tao, D. Fast Spatio-Temporal Residual Network for Video Super-Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
33. Wang, X.; Chan, K.C.; Yu, K.; Dong, C.; Change Loy, C. EDVR: Video Restoration With Enhanced Deformable Convolutional Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Long Beach, CA, USA, 15–20 June 2019.
34. Nah, S.; Baik, S.; Hong, S.; Moon, G.; Son, S.; Timofte, R.; Mu Lee, K. NTIRE 2019 Challenge on Video Deblurring and Super-Resolution: Dataset and Study. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Long Beach, CA, USA, 15–20 June 2019.
35. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
36. Huber, P.J. Robust estimation of a location parameter. In *Breakthroughs in Statistics*; Springer: Berlin/Heidelberg, Germany, 1992; pp. 492–518.
37. Huang, Y.; Wang, W.; Wang, L. Bidirectional recurrent convolutional networks for multi-frame super-resolution. In Procedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 235–243.
38. Wang, L.; Guo, Y.; Lin, Z.; Deng, X.; An, W. Learning for video super-resolution through HR optical flow estimation. In Procedings of the Asian Conference on Computer Vision, Perth, Australia, 2 December 2018; pp. 514–529.