applied sciences

MDPI

Article

A Deep Multi-Frame Super-Resolution Network for Dynamic Scenes

Ze Pan [1,2,3], Zheng Tan [1,3] and Qunbo Lv [1,3,*]

[1] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China; panze14@mails.ucas.edu.cn (Z.P.); tanzheng@aircas.ac.cn (Z.T.)
[2] School of Optoelectronics, University of Chinese Academy of Sciences, Beijing 100049, China
[3] Key Laboratory of Computation Optical Imaging Technology, Chinese Academy of Sciences, Beijing 100094, China
* Correspondence: lvqb@aircas.ac.cn

Abstract: The multi-frame super-resolution techniques have been prosperous over the past two decades. However, little attention has been paid to the combination of deep learning and multi-frame super-resolution. One reason is that most deep learning-based super-resolution methods cannot handle variant numbers of input frames. Another reason is that it is hard to capture accurate temporal and spatial information because of the misalignment of input images. To solve these problems, we propose an optical-flow-based multi-frame super-resolution framework, which is capable of dealing with various numbers of input frames. This framework enables to make full use of the input frames, allowing it to obtain better performance. In addition, we use a spatial subpixel alignment module for more accurate subpixel-wise spatial alignment and introduce a dual weighting module to generate weights for temporal fusion. Both two modules lead to more effective and accurate temporal fusion. We compare our method with other state-of-the-art methods and conduct ablation studies on our method. The results of qualitative and quantitative analyses show that our method achieves state-of-the-art performances, demonstrating the advantage of the designed framework and the necessity of proposed modules.

Keywords: multi-frame super-resolution; subpixel alignment; dual weighting

Citation: Pan, Z.; Tan, Z.; Lv, Q. A Deep Multi-Frame Super-Resolution Network for Dynamic Scenes. *Appl. Sci.* **2021**, *11*, 3285. https://doi.org/10.3390/app11073285

Academic Editors: Antonio Fernández and Changhoon Yim

Received: 5 March 2021
Accepted: 2 April 2021
Published: 6 April 2021

1. Introduction

The super-resolution techniques, as one type of fundamental enhancement methods in the computer vision area, have received more attention over the past two decades. The purpose of these techniques is to recover high-frequency information from degraded low-resolution images to generate high-resolution images with rich details. Now, the super-resolution techniques have been widely applied to different areas, such as post-processing of optical imaging, image manipulation and video enhancement for Internet resources.

The super-resolution techniques can be divided into three categories according to the number of input and output images: single-frame super-resolution (SFSR), multi-frame super-resolution (MFSR) and video super-resolution (VSR). SFSR aims to recover one high-resolution image with only one input image, while the other two kinds of super-resolution techniques make use of a series of low-resolution images for restoration. The main difference between VSR and MFSR is that MFSR does not require the constant number of input images, while VSR only utilizes a fixed number of frames within a fixed sliding window to generate one high-resolution frame. Therefore, MFSR can be seemed as a generalized VSR.

Recently, deep learning-based methods have been applied to super-resolution tasks, leading to substantial performance gain, especially in SFSR [1–6] and VSR [7–20]. However, few works have explored the combination of deep learning and MFSR because of the following two difficulties.

One difficulty is that MFSR should make full use of fluctuating numbers of inputs. Theoretically, because the super-resolution restoration is an ill-posed problem, the restoration performs better if we use more input frames to fuse. Therefore, MFSR methods need to consider all the input frames to obtain better results. However, the numbers of input frames are variant in real-world situations, which is hard to process by current deep learning-based super-resolution methods. Specifically, a convolutional layer, one type of the essential basic layers in deep learning, naturally has difficulties processing variant numbers of inputs because the convolutional weights in a convolutional layer are fixed after training. Hence, the fixed shape of weights means the fixed input shape of a convolutional layer, resulting in a fixed number of input images. VSR methods do not solve this problem and use only part of the inputs. In other words, the low-resolution target frame together with its adjacent frames within a sliding window are sent to the VSR models to generate high-resolution target frames. In such conditions, once the size of the sliding window is determined, it cannot be changed during the inference procedure, which means the frames out of the window will not participate in the restoration. Thus, VSR methods are not applicable to MFSR.

Another difficulty is how to utilize the temporal and spatial correlation for restoration. There are two kinds of solutions to extract temporal and spatial information using deep networks, but none of them are perfect. One solution is to use optical flow as a motion vector [9,12,13]. Frames are warped to align with the target frame under the guidance of their optical flows. This kind of solution is suitable for dynamic scenes. However, both forward warping and backward warping will introduce extra error, which influences the restored results significantly. Moreover, sometimes it is hard to estimate optical flow accurately, which also results in the failure of the restoration. Another solution is to implicitly compensate temporal motion by the convolutional network, including 3D convolution [17] and recurrent structure [21]. However, when motions between frames become large, they fail to capture the correlation because of the limit of the receptive field.

To solve the first problem, we propose an optical-flow-based framework with a flexible number of inputs. We believe the key for a flexible number of inputs relies on the framework design, instead of proposing new deep learning layers, considering the fact that current deep learning layers do not meet the basic requirements in the MFSR task because the shape of weights in deep networks keeps unchanged once the model is determined. Therefore, it is necessary to build a framework, which replaces the deep learning layers with flexible temporal fusion operations. As a result, we fuse the features of different input frames together through a weighted sum operation, which is a non-deep-learning approach. The features for fusion are extracted from input frames by a deep network, and the fused features are input into another deep network to recover high-resolution images. In a word, the main difference of this framework is that it does not use deep learning layers to fuse, which allows variant input numbers in MFSR.

To solve the second problem, we propose a spatial subpixel alignment (SSA) module and a dual-weighting (DW) module. In the MFSR task, we should align the input frames to one of the input frames, a basic frame, before fusion. The accuracy of this alignment operation is essential for temporal fusion. Therefore, we propose a new alignment module, the SSA module, to conduct a subpixel-wise alignment in a high-resolution grid, while other methods perform a pixelwise alignment. As for the DW module, we design it to cooperate with the weighted sum fusion strategy. It generates a series of weights for all low-resolution input frame to increase the accuracy and eliminate the potential errors for fusion. It deploys a distance weighting network to eliminate the warping errors and uses a content-aware weighting network to generate weights according to the content in the frame. As shown in Figure 1, our method can generate sharper images with richer details than state-of-the-art methods, Recurrent Residual network (RRN) [21].

Figure 1. Visual comparison with one of the state-of-the-art methods Recurrent Residual network (RRN). The numbers and words of our methods in zoomed area are more clear (The Korean on the license plate in the figure are symobles for vehicle classification, having no explicit meanings).

The main contributions of this paper can be summarized as follows:

1. We propose an optical-flow-based multi-frame super-resolution framework, which is capable of handling any number of input frames to generate the high-resolution target frame with only one network;
2. We propose an SSA module for spatial alignment and a DW module for temporal fusion. The SSA module offers a subpixel-wise alignment, and the DW module generates proper weights to decrease the errors caused by warping and optical flow. The experimental results show that both modules can lead to performance gain.

2. Related Work

In the SFSR task, the deep learning-based methods totally surpass the traditional methods [1–3]. Ref. [4] proposed a lightweight network by recursive learning. Refs. [5,6] developed deeper and more complicated networks, allowing them to obtain better results.

VSR methods experienced huge performance improvement thanks to the deep learning-based technology [7–10]. There are two main categories for temporal and spatial information extraction: the flow-based methods and implicit temporal motion compensation methods. Refs. [9,12] used optical flow for the motion estimation and performed warping for alignment. Ref. [13] used the recurrent structure to enhance the temporal information. These optical-flow-based methods have limited performance because of the inaccuracy in optical flow estimation and the alignment operation errors. To solve this problem, ref. [14] refined the optical flow in the image restoration task. Refs. [15,16] introduced the high-resolution optical flow for video super-resolution (SOFVSR). For other methods, ref. [17] used 3D convolution to estimate temporal information in dynamic upsampling filters (DUF) networks. Ref. [18] used deformable convolutional layers to get temporal information by their offsets. Refs. [19,20] explored the 2D convolutional layer. RRN [21] revealed that recurrent, residual structure performs better than 2D convolution and 3D convolution. These implicit motion compensations are purely based on convolutional networks and removed the alignment step. Therefore, when the range of motion is larger than the receptive field of the convolutional network, there will be no temporal information available for restoration.

Multi-frame super-resolution methods mainly focused on the design of modeling super-resolution tasks and a regularization term. Farsiu was the first to introduce maximum a posteriori into super-resolution task and proposed an edge-preserving regulation term, total bilateral variation [22]. Then [23] used an M-estimator to deal with the outliers.

Ref. [24] came up with a Bayesian approach to model the registration parameters and restored images in one framework jointly. Zeng chose half-quadratic estimation to model [25]. He achieved the adaptive regularization of weight tuning by modeling noise levels [26]. Following [26], ref. [27] proposed a spatially adaptive Bayesian and reweighted the weight of regularization in optimizing. Ref. [28] took the blur effect into consideration. Batz regarded the super-resolution task as an interpolation based on Voronoi tessellation [29,30]. Ref. [31] also combined a hybrid super-resolution technique for SFSR with MFSR. Although various models have been proposed, MFSR still lacks effective frameworks to incorporate deep learning.

3. Methods

3.1. Framework

Most VSR methods fuse features from different inputs by convolutional networks, which are only suitable for a constant number of fusion features. In our model, however, we used a normalized weighted sum feature fusion strategy because this operation has no limit to the number of fusion features. As a result, our framework could handle different numbers of inputs.

The framework is shown in Figure 2. The target frame I_0, along with other N-1 frames, $I_i(i = 1 : N-1)$ are sent into the body part to get series of deep features $[F_0, F_1 \dots F_i \dots F_{N-1}]$. Each F_i consists of M features, i.e., $F_i = \{f_{i,j} | 1 \le j \le M\}$. Then these features are aligned to F_0 by the SSA module to generate the aligned features F_i^a. Each feature $f_{i,j}^a$ in F_i^a are input to generate the fused feature $\hat{f}_j$, the j-th feature from the fused features $\hat{F}$, by a weighted sum operation, as shown in Equation (1):

$$\hat{f}_j = \frac{f_{0,j} + \sum_{i=1}^{N-1} w_i \times f_{i,j}^a}{1 + \sum_{i=1}^{N-1} w_i} \tag{1}$$

where w_i is the weight of features F_i^a from each input frame I_i and is generated by the DW module. Then, the fused features $\hat{F}$ are fed into the enhancement part and the tail part to generate the final output.

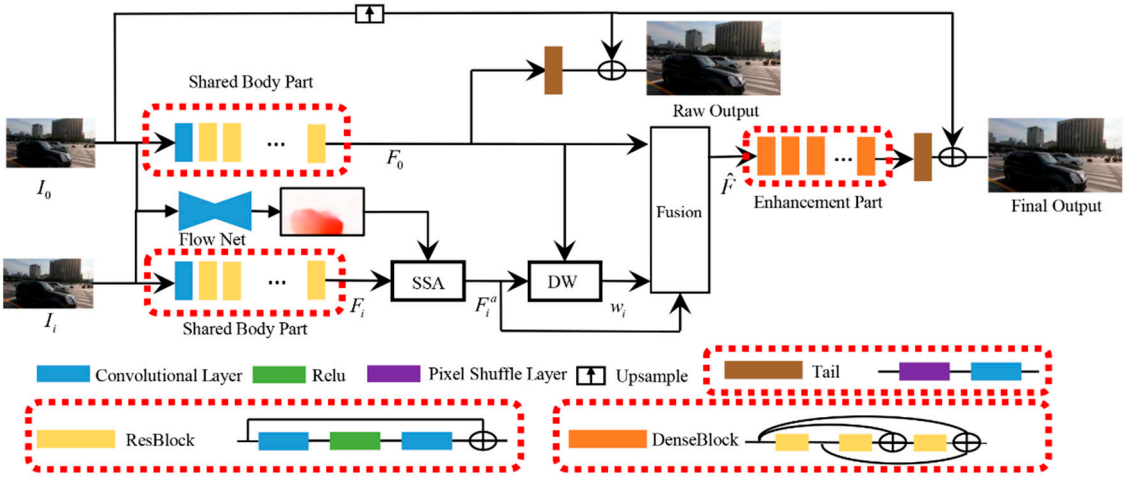

Figure 2. The framework of our method.

As for the body part, it consists of one convolutional layer and some resblocks. Each resblock has a basic form of conv-relu-conv, and all the resblocks has the same number of channels and the same resolution. The enhancement part consists of several denseblocks. The denseblock has dense connections between its three resblocks. The tail part consists

of a pixel shuffle layer and a convolutional layer. In addition, the tail part can be directly added at the end of the body part during the training phase to generate a raw output, which accelerated training.

According to Equation (1), the fused feature $\hat{F}$ equals F_0 when only one target frame I_0 is fed into the model, which serves as the SFSR. When the number of input frames N is fixed, this model turns to VSR. Therefore, our framework is more flexible and was suitable for various situations.

3.2. Spatial Subpixel Alignment Module

Commonly, most learning-based methods align the input frames to the target frame on a low-resolution grid. This kind of operation may result in a loss of subpixel information, which eventually affects the restoration. To solve this problem, we propose an SSA module to get more accurate alignment. The SSA module provides subpixel-wise alignment, which is processed with the aid of optical flow.

We aim to align features in high-resolution for more accurate alignment. First, we upsample the dense optical flow and the features into a high-resolution grid. The optical flow with a resolution of (H, W) is resized to the resolution of (H × r, W × r), where H, W and r represent the image height, the image width and the image upscaling factor, respectively. Similarly, the features with the size of (m × r × r, H, W) is resampled to the size of (m, H × r, W × r) through pixel shuffle operation, where m indicats the number of feature groups and each group has r*r features. Different from resizing, pixel shuffle rearranges the values along the dimension of different features to a high-resolution grid, keeping the total number of elements unchanged. Then a backward warping operation is applied with both optical flow and features to generate aligned features with the size of (m, H × r, W × r). Finally, aligned features on the HR grid are converted into low-resolution features with the size of (m × r × r, H, W) by a reverse pixel shuffle operation. In this paper, we set m to 8. A simple example is shown in Figure 3, where m = 1, r = 2.

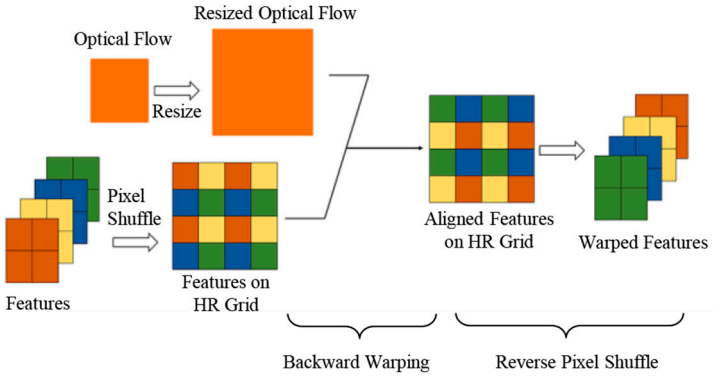

Figure 3. The architecture of the spatial subpixel alignment (SSA) module.

The basic reason for such a design is that we choose pixel shuffle operation in our tail part, which is a common choice. It is clear that low-resolution features in each channel had a subpixel relationship with other channels because they are adjacent to each other after pixel shuffle. As a result, any subpixel operation would modify all the channels of features, and low-resolution features in our model should be determined by all the channels of the former features. Under this premise, if we directly warp on a low-resolution grid for alignment, each new channel is determined by one channel, which will cause the loss of subpixel information.

Figure 3 shows a simple case that how the subpixel operation affects all channels. For simplicity, four channels of features in red, yellow, blue and green are planned to move 0.5 pixels in both horizontal and vertical direction, which is exactly 1 pixel in both directions on ×2 resolution. After applying the SSA module, the color order turns to green, blue, yellow and red. This proves the cross-channel effect caused by optical flow.

3.3. Temporal Dual-Weighted Module

To eliminate the errors caused by optical flow and warping operation, we propose a DW module. Because we apply a weighted sum operation in the final fusion, as shown in Equation (1), DW gives the weight w_i of each frame according to its content and its difference to the target frame, as shown in Figure 4. It is worth mentioning that although one frame only computes one weight, this weight is applied to multiply each feature from this frame for fusion. The w_i is the multiplication of different weight w_i^d and the content weight w_i^c as shown in Equation (2):

$$w_i = w_i^d \times w_i^c \tag{2}$$

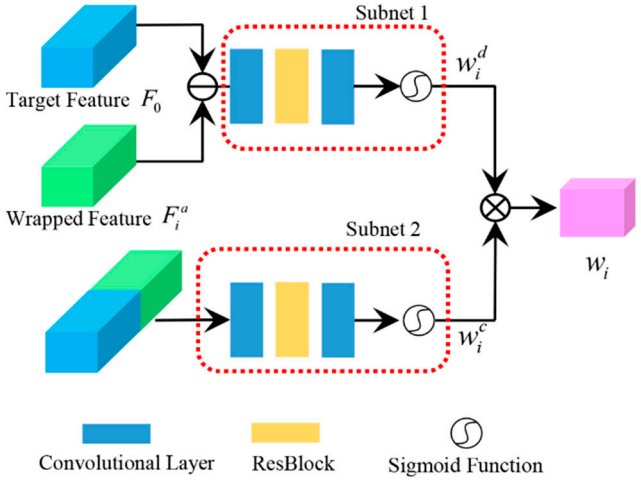

Figure 4. The architecture of the dual-weighting (DW) module.

The warped features or images often suffer from distortion, which is attributed to the inaccuracy of optical flow and the warping itself. Therefore, we compute a weight w_i^d for each warped features, which measures the difference between the warped features and the target features. By doing so, the unreliable areas are not taken into consideration in the later fusion process. Difference weight w_i^d is generated by a small common convolutional network with five layers, as shown in Equation (3):

$$w_i^d = \sigma(Net_d(F_0 - F_i^a)) \tag{3}$$

where $Net_d(*)$ denotes the inference process of the small network. It takes the difference between F_0 and F_i^a as the input, and the output is followed by a sigmoid function $\sigma(*)$, so that the range of w_i^d is limited in the range of 0–1.

Moreover, we introduce content weight w_i^c to adjust the weights according to the content. The computation of content weight w_i^c can be formulated in Equation (4):

$$w_i^c = \sigma(Net_c(cat(F_0, F_i^a))) \tag{4}$$

where $Net_c(*)$ is the inference process of another small network with a similar architecture to $Net_d(*)$. Differently, this small network processes a concatenation operation $cat(*)$ in the beginning.

The two subnets for computing the two weights have the same structure, as shown in Figure 4. They are simply a cascade of two convolutional layers and one resblock placed in the middle. Both subnets end with a sigmoid function. Differently, the channel number for content weight is twice that of distance weight.

4. Results

4.1. Dataset and Implementation Details

Currently, ref. [32] is the only dataset for MSFR. However, this dataset focuses on a remoted multi-band vegetation images and is not suitable for general usage. Therefore, we choose the realistic and dynamic scenes (REDS) dataset [33], one of the benchmark datasets, as our training and evaluation set. One reason is that the REDS contains various realistic and dynamic scenes. Another reason is that the motion range between frames in the REDS dataset is larger and more complex than other datasets, which is in line with the real-world situation. There are 270 clips in the REDS dataset, and each clip contains 100 frames. We chose 240 clips for training and left the rest 30 clips for evaluation.

The number of resblocks in the body part was set to 10, and the number of denseblocks in the enhancement part was set to 2. The channel number was 128. We selected the pretrained pyramid, warping, and cost volume network (PWCNet [34]) as the optical flow estimator. During the training phase, we divided 100 frames in each clip into 20 groups and randomly selected the target frame and input numbers. We randomly selected crop positions for different groups. The patch size was set to 256 pixels of height and width. We applied data augmentation techniques, including flipping and rotation. We first used Adam optimizer [35] to train our model for 100 epochs with a learning rate of 10^{-4}. During this stage, we computed the L1 distance to the labels for both the raw and final outputs. Moreover, we froze the weights of PWCNet. After this, we trained our network for another 50 epochs. During this stage, we only computed the L1 distance between the final outputs and the labels. The weights of PWCNet were added to the training list, and the initial learning rate was $10^{-5,}$ and it decreased to 10^{-6} after 30 epochs. We implemented our method in PyTorch with NVidia 1080Ti GPU (manufacted by Gigabyte Technology in Taiwan, China).

4.2. Comparisons with Other State-of-the-Art Methods

We compared our method with the following three state-of-the-art algorithms, DUF [17], SOFVSR [16] and RRN [21]. DUF and RRN are two typical implicit motion compensation algorithms, which use 3D convolutional and the recurrent neural network to capture the motion information between frames, respectively. SOFVSR is an optical-flow-based algorithm, which surpasses other optical-flow-based methods. DUF and SOFVSR use 7 and 3 frames to generate one high-resolution frame, respectively. RRN is only capable of handling 2 frames, so it uses 2 frames for evaluation. All the experiments with these state-of-the-art methods were conducted using the corresponding code officially released by their authors. To ensure a fair comparison, we used 2 and 3 frames in our method, which means 1 high-resolution output was computed by its corresponding low-resolution frame together with 1 and 2 adjacent frames, respectively. It should be mentioned that, as for DUF, we did not use the official low-resolution input frames to evaluate because we found huge misalignments between the restoration and the ground truth. Therefore, we used the downsampling function in their code to generate low-resolution frames for evaluation.

The quantitative comparison results are shown in Table 1, where Ours-2 and Ours-3 represent our method with 2 and 3 input frames, respectively. It was obvious that our method could process different numbers of input frames, and it outperformed other methods in terms of both the PSNR and runtime metrics on the REDS dataset.

Table 1. Quantitative comparison on the realistic and dynamic scenes (REDS) dataset.

	Bicubic	**DUF [17]**	**SOFVSR [16]**	**RRN [21]**	**Ours-2**	**Ours-3**
PSNR (db)	25.76	26.13	28.17	28.28	28.39	28.45
Runtime (s)	-	1.26	0.21	0.37	0.09	0.19

Compared to SOFVSR, an optical-flow-based method using 3 input frames, our method with the same inputs (Ours-3) had an improvement of 0.28 db on PSNR. This

proves that the performance gain of our method comes from the fusion strategy proposed in this paper instead of the optical flow approach itself. It may be because that SOFVSR only concentrates on the accuracy of optical flow, while our method focuses on the temporal fusion approach, leading to better results.

RRN had better performance than SOFVSR and DUF, which indicated that the 3D convolutional layer in RRN learns meaningful temporal information in the super-resolution task. Even so, the value of RRN's PSNR was still a bit lower than that of our method with 2 input frames. This means our fusion approach could obtain promising results.

As for the computational complexity, we computed an average runtime for 100 samples. As shown in Table 1, our method had the fastest runtime. We believe that the fast speed was attributed to the lightweight of the body part and the fusion part in our network.

In addition, we show the visual comparisons of the results on the REDS dataset, as shown in Figure 5. Our results were sharper and had fewer artifacts than other methods.

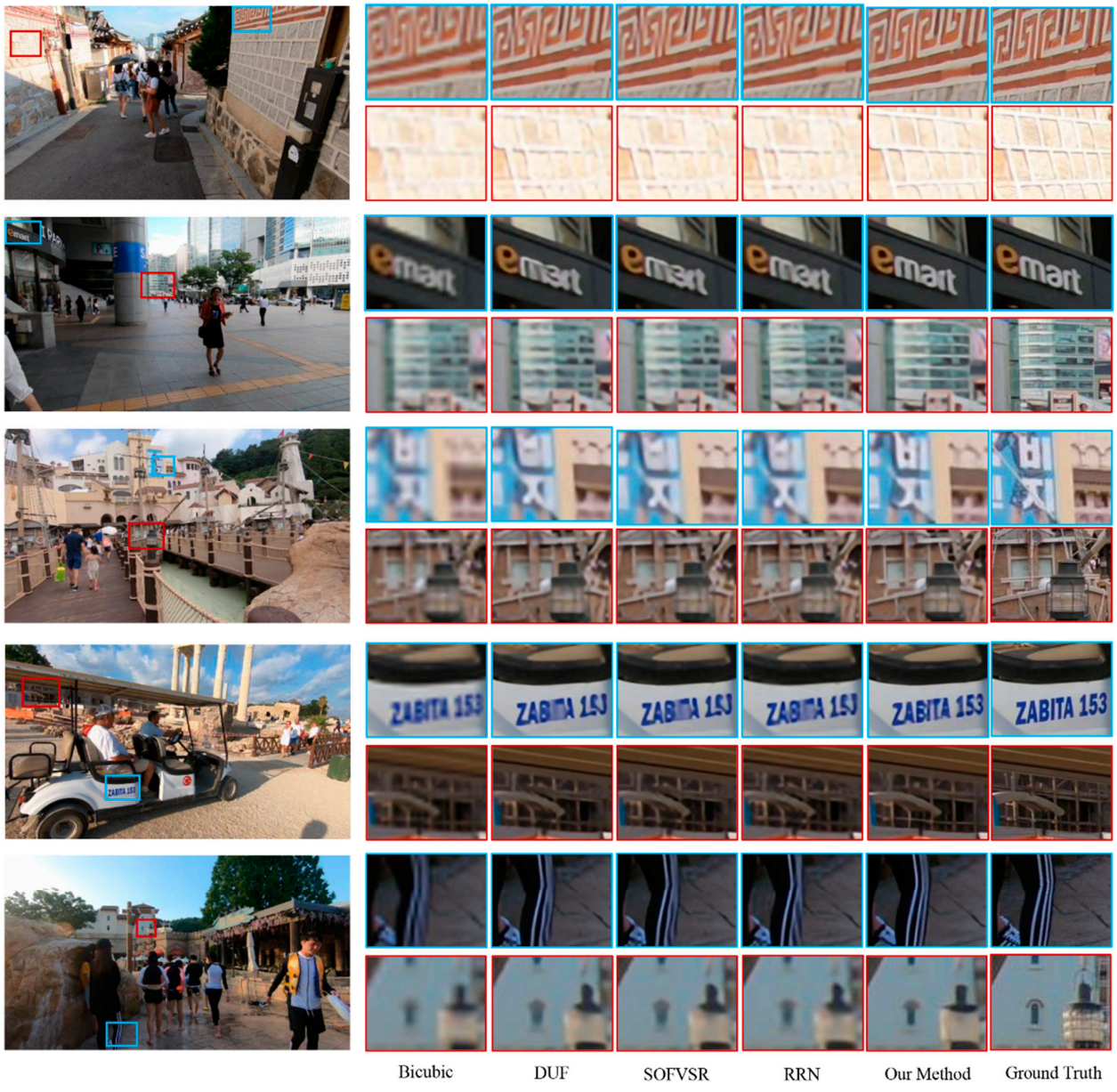

Figure 5. Visual comparison of the results on the REDS dataset. The images on the left are the results of our method. The zoomed images on the right side are the results of various methods for the area in the bounding box(The Korean in the third figure means 'busy').

4.3. Ablation Studies

We conducted three experiments with 5 input frames to demonstrate the effectiveness of the SSA module and the DW module. As shown in Table 2, model 1, equipped with both the SSA module and the DW module, had the best performance, whose PSNR reaches 28.50 db.

Table 2. Ablation study results on the REDS dataset.

Models	Input Number	Alignment	Fusion Weight	PSNR
Model 1	5	SSA	DW	28.50
Model 2	5	Standard	DW	28.19
Model 3	5	SSA	Average	27.93
Model 4	1	SSA	DW	28.18
Model 5	3	SSA	DW	28.45
Model 6	9	SSA	DW	28.52

Compared with model 1, model 2 removes the SSA module and had a substantial drop of 0.31 db, which may have been caused by the loss of the subpixel information in standard low-resolution grid warping operation. This means subpixel-wise alignment was essential in MFSR tasks, especially in optical-flow-based methods. Therefore, we can confirm the effectiveness of the SSA module through this comparison.

In addition, model 1 used the DW module to decrease errors caused by the optical flow or the alignment operation before fusion operation. In contrast, model 3 replaces the DW module with the average operation and gets a performance drop of 0.57 db compared with model 1. It is not strange that, without the DW module, the fusion operation in our method introduced many errors into the restoration procedure. This comparison demonstrates that the DW module was indispensable in our method. Surprisingly, model 3 behaved even worse than model 4, which indicated that errors introduced by optical flow were harmful to the restoration if they were not carefully processed.

We also tested our method with different numbers of input frames to justify the advantages of our framework for being able to handle variant numbers of inputs. Comparing model 1 with models 4 and 5, we found that the more low-resolution frames participated in restoration, the better performance the model had. It shows that our model was capable of computing useful temporal information between different input frames. However, when the input number increased to 9, as shown in model 6, the performance gain became tiny. This may have been caused by the decrease of the overlapping area when the motion range between frames was large.

5. Conclusions

In this paper, we proposed an optical-flow-based MFSR network. On one hand, we design a framework, which enables us to handle variant numbers of input frames by replacing deep learning layers with a weighted sum operation for fusion. This design allows making use of more temporal information from input frames. On the other hand, our method contains an SSA module and a DW module. The SSA module offers more accurate subpixel-wise alignment by a backward warping operation on the high-resolution grid. Moreover, the DW module generates an error suppression weight to incorporate with the weighted sum fusion strategy, considering the distance on both values and contents of features from each frame. Moreover, the qualitative and quantitative analyses show that our method achieves the-state-of-the-art performances, which demonstrates the effectiveness of the framework and two modules. We believe our multi-frame fusion approach is applicable to other temporal computer vision tasks as well, such as multi-frame deblurring and multi-frame dehazing, which we will explore in future work.

Author Contributions: Conceptualization, Z.P.; methodology, Z.P.; validation, Z.P.; writing—original draft preparation, Z.P.; writing—review and editing, Z.P.; visualization, Z.P.; supervision, Z.T. and Q.L.; project administration, Z.T.; funding acquisition, Q.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by GFZX, grant number GFZX0403260306.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: This paper was supported by GFZX0403260306.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Haris, M.; Shakhnarovich, G.; Ukita, N. Deep back-projection networks for super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018. [CrossRef]
2. Kim, J.; Lee, J.K.; Lee, K.M. 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, 26 June–1 July 2016. [CrossRef]
3. 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, Honolulu, HI, USA, 21–26 July 2017. [CrossRef]
4. 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, Las Vegas, NV, USA, 26 June–1 July 2016. [CrossRef]
5. Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; Fu, Y. Image Super-Resolution Using Very Deep Residual Channel Attention Networks. Available online: https://arxiv.org/pdf/1807.02758.pdf (accessed on 15 February 2021).
6. Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; Fu, Y. Residual dense network for image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018. [CrossRef]
7. Liao, R.; Tao, X.; Li, R.; Ma, Z.; Jia, J. Video super-resolution via deep draft-ensemble learning. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 13–16 December 2015. [CrossRef]
8. 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, Venice, Italy, 22–29 October 2017. [CrossRef]
9. 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, Honolulu, HI, USA, 21–26 July 2017. [CrossRef]
10. Liu, C.; Sun, D. On Bayesian Adaptive Video Super Resolution. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 346–360. [CrossRef] [PubMed]
11. Yu, K.; Dong, C.; Lin, L.; Loy, C. Crafting a toolchain for image restoration by deep reinforcement learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018. [CrossRef]
12. 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, Venice, Italy, 22–29 October 2017. [CrossRef]
13. Sajjadi, M.S.M.; Vemulapalli, R.; Brown, M. Frame-recurrent video super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018. [CrossRef]
14. Xue, T.; Chen, B.; Wu, J.; Wei, D.; Freeman, W. Video enhancement with task-oriented flow. *Int. J. Comput. Vis.* **2019**, *127*, 1106–1125. [CrossRef]
15. Wang, L.; Guo, Y.; Lin, Z.; Deng, X.; An, W. Learning for video super-resolution through HR optical flow estimation. In Proceedings of the Asian Conference on Computer Vision, Perth, Australia, 2–6 December 2018.
16. Wang, L.; Guo, Y.; Liu, L.; Lin, Z.; Deng, X.; An, W. Deep Video Super-Resolution Using HR Optical Flow Estimation. *IEEE Trans. Image Process.* **2020**, *29*, 4323–4336. [CrossRef] [PubMed]
17. 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/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018. [CrossRef]
18. 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, 16–20 June 2020. [CrossRef]
19. Yi, P.; Wang, Z.; Jiang, K.; Jiang, J.; Ma, J. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019. [CrossRef]

20. Fuoli, D.; Gu, S.; Timofte, R. Efficient video super-resolution through recurrent latent space propagation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019. [CrossRef]
21. Isobe, T.; Zhu, F.; Wang, S. Revisiting Temporal Modeling for Video Super-Resolution. Available online: https://arxiv.org/abs/2008.05765v1 (accessed on 15 February 2021).
22. Farsiu, S.; Robinson, M.D.; Elad, M.; Milanfar, P. Fast and Robust Multiframe Super Resolution. *IEEE Trans. Image Process.* **2004**, *13*, 1327–1344. [CrossRef] [PubMed]
23. Patanavijit, V.; Jitapunkul, S. A Lorentzian Bayesian approach for robust iterative multiframe super-resolution reconstruction with Lorentzian-Tikhonov regularization. In Proceedings of the International Symposium on Communications and Information Technologies, Bangkok, Thailand, 20 September–18 October 2006. [CrossRef]
24. Liu, C.; Sun, D. A Bayesian approach to adaptive video super resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, USA, 21–25 June 2011. [CrossRef]
25. Zeng, X.; Yang, L. A robust multiframe super-resolution algorithm based on half-quadratic estimation with modified BTV regularization. *Digit. Signal Process.* **2013**, *23*, 98–109. [CrossRef]
26. He, H.; Kondi, L.P. An image super-resolution algorithm for different error levels per frame. *IEEE Trans. Image Process.* **2006**, *15*, 592–603. [CrossRef] [PubMed]
27. Kohler, T.; Huang, X.; Schebesch, F.; Aichert, A.; Maier, A.; Hornegger, J. Robust Multiframe Super-Resolution Employing Iteratively Re-Weighted Minimization. *IEEE Trans. Comput. Imaging* **2016**, *2*, 42–58. [CrossRef]
28. Ma, Z.; Liao, R.; Tao, X.; Xu, L.; Jia, J.; Wu, E. Handling motion blur in multi-frame super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015. [CrossRef]
29. Batz, M.; Eichenseer, A.; Kaup, A. Multi-image super-resolution using a dual weighting scheme based on Voronoi tessellation. In Proceedings of the IEEE International Conference on Image Processing, Phoenix, AZ, USA, 25–28 September 2016. [CrossRef]
30. Batz, M.; Koloda, J.; Eichenseer, A.; Kaup, A. Multi-image super-resolution using a locally adaptive denoising-based refinement. In Proceedings of the IEEE 18th International Workshop on Multimedia Signal Processing, Montreal, QC, Canada, 21–23 September 2016. [CrossRef]
31. Batz, M.; Eichenseer, A.; Seiler, J.; Jonscher, M.; Kaup, A. Hybrid super-resolution combining example-based single-image and interpolation-based multi-image reconstruction approaches. In Proceedings of the IEEE International Conference on Image Processing, Quebec City, QC, Canada, 27–30 September 2015. [CrossRef]
32. Deudon, M.; Kalaitzis, A.; Goytom, I.; Arefin, M.R.; Lin, Z.; Sankaran, K.; Michalski, V.E.; Kahou, S.; Cornebise, J.; Bengio, Y. HighRes-net: Recursive Fusion for Multi-Frame Super-Resolution of Satellite Imagery. Available online: https://arxiv.org/pdf/2002.06460.pdf (accessed on 15 February 2021).
33. Nah, S.; Baik, S.; Hong, S.; Moon, G.; Son, S.; Timofte, R.; Lee, K.M. 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 Workshops, Long Beach, CA, USA, 16–20 June 2019. [CrossRef]
34. Deqing, S.; Xiaodong, Y.; Mingyu, L.; Jan, K. PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018. [CrossRef]
35. Kingma, D.P.; Ba, J.L. Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference for Learning Representations, San Diego, CA, USA, 7–9 May 2015.