applied sciences

MDPI

Article

Video Super-Resolution with Regional Focus for Recurrent Network

Yanghui Li [1], Hong Zhu [1,*], Lixin He [1], Dong Wang [1], Jing Shi [1] and Jing Wang [2]

[1] School of Automation and Information Engineering, Xi'an University of Technology, Xi'an 710048, China
[2] School of Printing, Packaging and Digital Media, Xi'an University of Technology, Xi'an 710054, China
* Correspondence: zhuhong@xaut.edu.cn

Abstract: Video super-resolution reconstruction is the process of reconstructing low-resolution video frames into high-resolution video frames. Most of the current methods use motion estimation and motion compensation to extract temporal series information, but the inaccuracy of motion estimation will lead to the degradation of the quality of video super-resolution results. Additionally, when using convolution network to extract feature information, the number of feature information is limited by the number of feature channels, resulting in poor reconstruction results. In this paper, we propose a recurrent structure of regional focus network for video super-resolution, which can avoid the influence of inaccurate motion compensation on super-resolution results. Meanwhile, regional focus blocks in the network can focus on different areas of video frames, extract different features from shallow to deep layers, and skip-connect to the last layer of the network through feature aggregation to improve the richness of features participating in the reconstruction. The experimental results show that our method has higher computational efficiency and better video super-resolution results than other temporal modeling methods.

Keywords: recurrent neural network; regional focus; feature aggregation; video super-resolution.

Citation: Li, Y.; Zhu, H.; He, L.; Wang, D.; Shi, J.; Wang, J. Video Super-Resolution with Regional Focus for Recurrent Network. *Appl. Sci.* **2023**, *13*, 526. https://doi.org/10.3390/app13010526

Academic Editor: Andrea Prati

Received: 24 November 2022
Revised: 23 December 2022
Accepted: 26 December 2022
Published: 30 December 2022

1. Introduction

Super-resolution is a classical problem in the computer vision field. Its main objective is to predict high-resolution images with rich texture details from low-resolution images, which is essentially the solution of ill-posed problems. In recent years, with the increasing demand for high-definition display, this problem has received wide attention. Single image super-resolution mainly uses the method based on image prior information and self-similarity, and takes a single low-resolution image as an input to predict the high-resolution image. Method [1] uses the combination of adaptive reconstruction and learning to complete the super-resolution reconstruction of a single image. However, the reconstruction results with these methods are very limited. Compared with the single image super-resolution, video super-resolution takes multiple consecutive low-resolution video frames as input and utilizes the temporal information between adjacent frames to further improve the recovery quality of video frame details.

Deep learning is one of the popular methods for video super-resolution. Video super-resolution methods can be divided into video frames alignment methods and video frames misalignment methods because of the differences in the processing of video super-resolution reconstruction. For the video frames alignment, explicit motion compensation methods [2–6] and implicit motion compensation methods [7–12] are the main methods. Explicit motion compensation methods usually obtain inter-frame motion information through optical flow calculation, and perform warping operations to align adjacent frames and target frames based on inter-frame motion information. The first end-to-end video super-resolution reconstruction methods were proposed in Reference [2], which combines optical flow estimation and spatio-temporal network training to achieve video super-resolution. In order to improve the reconstruction quality of video super-resolution, Wang et al. [13] calculated the high-resolution optical flow image to make the video frame

alignment more accurate, so as to improve the reconstruction quality. Song et al. [14] learned the gradient mapping function between high-resolution and low-resolution to regularize multi-frame fusion, and used forward and backward motion field priori to regularize the estimation of interframe motion flow, which can restore better details with less information.

Deformable convolution is the most common implicit motion compensation method. As its typical example, Wang et al. [10] uses a deformable alignment module and a spatio-temporal attention fusion module to solve the large motion problem and efficient fusion of multiple frames in video super-resolution, respectively. Different from alignment methods, non-aligned methods [15–17] do not perform frame alignment before video super-resolution reconstruction, and mainly utilize spatial information or spatio-temporal information to extract features, such as 3D convolution, recurrent neural network, and non-local method. The non-local operation of method [18] is to calculate the similarity between other pixels in the image and the pixel at the current position, where the response value of the current position is equal to the sum of the weights of all possible positions in the input feature map. Isobe et al. [19] adopts the structure of cascaded residual blocks in the recurrent network, which reconstructs rich texture details while avoiding the vanishing gradient problem. The cascaded residual structure [20] has achieved a good effect in processing single-image super-resolution. However, when applied to the video super-resolution, this structure does not take into account the unbalanced proportion between simple and complex areas of the video frame. Usually, most of the video frames are simple regions, which makes the optimization direction of the network dominated by simple regions during the training process, resulting in the deterioration of the reconstruction results quality and the blurring of local regions. In addition, because the convolutional network has the characteristics of local connection and global sharing, the super-resolution task based on the convolutional network needs to restore the detailed information of the reconstruction results point-by-point. The features extracted from different residual blocks in the residual cascade network have the characteristics of progressive refinement. The output features of different residual blocks are similar, and the feature diversity is limited by the number of feature channels, which affects the recovery of high-frequency details in complex regions.

Motivated by the above discussion, a recursive regional focus network for video super-resolution will be proposed in this paper. In this recursive network, the regional focus block is designed to change the coupling relationship between the former and the latter layers in the cascading residual structure, and to give different degrees of attention to the regions with different reconstruction difficulties in the video frame. Meanwhile, we designed a feature aggregation structure that can send different features from shallow layer to deep layer extracted from each regional focus block in parallel to the last layer of the network for feature fusion, so as to increase the feature richness involved in video frame reconstruction.

The innovation of this paper mainly includes the following two points:

(1) The regional focus block proposed in this paper changes the coupling relationship between adjacent layers in the cascading residual structure, and pays different degrees of attention to different texture regions in the video frame. The feature information extracted from the network of different layers can participate in the reconstruction of the video frame, and will not lead the optimization direction of the network because the video frame contains more simple areas.
(2) The feature aggregation structure in the network sends the feature information extracted from different layers of networks from the shallow layer to the deep layer into the last layer of the network in a parallel manner, and carries out the feature fusion with the feature information extracted from the series structure in the network, which increases the feature information involved in the reconstruction of video frames and improves the quality of the reconstruction results.

The above structure can effectively improve the reconstruction quality of the video frame while guaranteeing the network computing speed. As shown in Figure 1, it is a comparison of the results of our method and other different methods. From the figure, we can see that our method has achieved better results in both contour reconstruction and the

detailed recovery of images. Finally, we carry out experiments in different test datasets, and compare them with the current mainstream video super-resolution work, the experimental results of which show that the reconstruction results obtained by our method are better than those obtained using other methods.

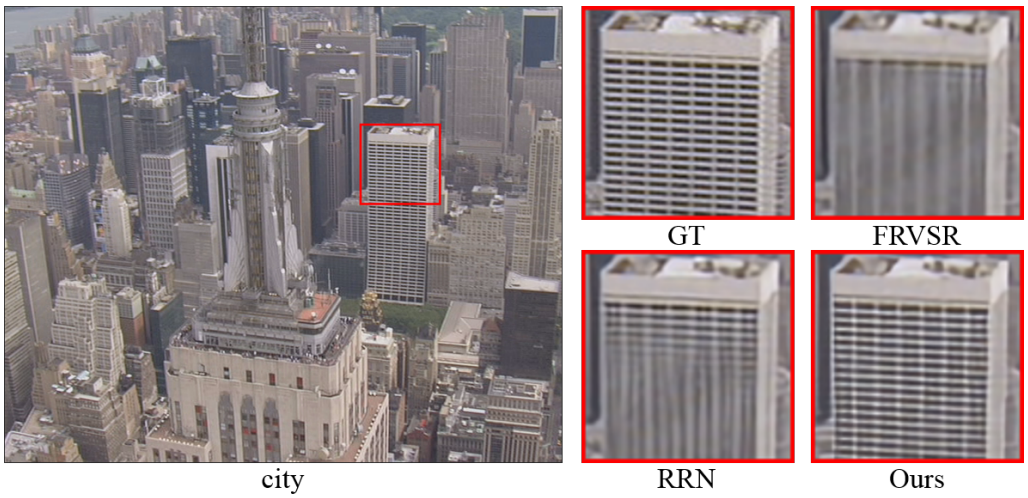

Figure 1. Comparison of the ×4 reconstruction results of our method and other methods in city video sequence.

2. Related Work

In the video super-resolution task, some methods explore the timing information between video frames with explicit motion estimation and compensation. Kim et al. [21] proposed a method of calculating optical flow using multiple frames of video simultaneously, which alleviates the occlusion problem when only adjacent frames are used to calculate optical flow. Xiao et al. [22] improved the performance of video super-resolution network via information extraction without changing the original network structure. Wang et al. [23] proposed a multi-memory convolutional neural network cascading optical flow network and image reconstruction network, which can make full use of the spatio-temporal complementary information of low-resolution video frames. Bao et al. [24] proposed an adaptive warping layer that combines optical flow and interpolation to synthesize target frame pixels. In order to fully utilize the temporal information, Li et al. [25] adopted motion compensation, depth full recursive convolution, and later feature fusion in the network to obtain more accurate video super-resolution results. The video super-resolution method proposed by Bare et al. [26] solved the inaccurate problem of motion estimation and the influence of the unreasonable design of partial convolutional network structure on reconstruction results. Kalarot et al. [27] proposed a full convolution neural network with unknown scene and category, which can take full advantage of previous estimates and share redundant information across continuous video frames. A time self-monitoring method was proposed by Chu et al. [28], and the method where temporal confrontation learning can achieve temporal consistency without sacrificing spatial details was proven. Li et al. [29] not only used the similarity between frames to overcome the error prone problem of optical flow estimation, but also designed a non-local aggregation scheme to discuss the self-similarity of cross-scale images. Li et al. [30] adopted the fusion of non-local modules and multi-scale features to further mine the effective information of video frames while calculating the optical flow of adjacent video frames. To better extract the complementary information of video frames, Isobe et al. [31] divided the pixels with different differences into two different subsets by calculating the temporal difference between video frames, and handed them with the receptive fields of different branches. Chan et al. [32] proposed a concise network structure by redesigning some basic functions of video super-resolution. After that, Chan et al. [33] redesigned the proposed network structure via second-order network

propagation and deformable alignment, and effectively exploited the spatio-temporal information of exaggerated misplaced video frames. In the alignment method, the purpose of motion estimation is to extract the motion information between video frames, and the motion compensation is to perform a warping operation to align the video frames according to the motion information between the video frames. Most of the motion estimation is achieved using optical flow methods, which calculates the motion between adjacent frames through the correlation and variation of adjacent frames in the temporal domain. However, the estimation of dense optical flow is a very time-consuming work. Meanwhile, errors are prone to occur in the optical flow estimation process, and the resulting inaccurate optical flow will cause artifacts in the results of video super-resolution based on optical flow.

In addition to the methods mentioned above, another class of methods are to use motion information between video frames to perform video super-resolution reconstruction in an implicit manner. Ying et al. [34] proposed a deformable 3D convolutional network to integrate spatio-temporal information in video super-resolution, which has excellent spatio-temporal modeling and flexible motion perception ability. Chen et al. [35] designed a separate non-local module to explore the relationship between video frames, effectively fuse video frames, and capture the relationship between feature maps via channel attention residual blocks. Dai et al. [36] introduced deformable convolution and deformable pooling in the model, used an additional offset to increase the spatial sampling location in the module, and learned the target task offset without additional supervision. The temporal variant alignment network proposed by Tian et al. [37] aligns the reference frame and each corresponding frame adaptively at the feature level without calculating the optical flow. Compared with the explicit motion compensation method, these implicit alignment methods reduce the computational cost, but in these methods, the redundant computation of multiple adjacent frames still exists in a temporal window, which requires the caching of multiple frames in advance to achieve video super-resolution.

Compared with the alignment methods, there are another class of methods that do not use frame alignment to achieve video super-resolution. Such methods mainly use spatial or spatio-temporal information for feature extraction, where the 2D convolution method extracts the feature information of video frames in space, and the 3D convolution method extracts the temporal information in the spatio-temporal domain to consider the correlation between video frames. Here, recurrent neural network methods are very suitable for processing video sequences with temporal information because of their recursive characteristics. Yan et al. [38] proposed an end-to-end trainable video frame and video super-resolution network about feature context correlation that is composed of local and context networks. The fast spatio-temporal residual network proposed by Li et al. [39] uses 3D convolution for video super-resolution tasks, which improves performance while maintaining a low computational load. Liu et al. [40] proposed a deep neural network with dual subnets and multistage communication up-sampling for the super-resolution of large motion video. Zhu et al. [41] proposed an end-to-end reversible residual spatio-temporal network, which takes full advantage of spatio-temporal information from low-resolution to high-resolution, and effectively modeled the temporal uniformity of continuous video frames.

There are also some methods to consider video super-resolution problems from another perspective. Li et al. [42] expanded the super-resolution reconstruction magnification from the traditional $\times 4$ to the high magnification of $\times 16$ and $\times 32$. Lee et al. [43] completed the final video super-resolution task by taking different video frames taken by the telephoto, wide-angle, and ultra wide cameras as mutual references. The head pose estimation methods [44,45] proposed by Liu et al. also have some enlightenment and reference significances.

To sum up, the non-alignment methods mainly rely on the nonlinear ability of the neural network to learn the motion correlation between video frames, thereby achieving the super-resolution reconstruction of video frames. The reconstruction result of the network is not affected by whether the video frames are aligned or not, and the design of the network structure determines the learning ability of the network. Therefore, the reasonable

design of the network structure is helpful for improving the performance of video super-resolution. The method proposed in this paper increases the feature information involved in the reconstruction through the feature extraction from shallow layer to deep layer, and the attention of different network layers to different feature regions. By using the feature aggregation structure to fuse the feature information, more feature information can be obtained to participate in the final super-resolution reconstruction of the network. The quality of network reconstruction has not been improved with the increase in the number of network parameters.. This issue will be further discussed in the subsequent ablation experiments.

3. Proposed Model

The main contribution of this paper is to design a module that combines residual blocks and regional focus blocks on the basis of recurrent network, and to combine feature aggregation in the network structure. This design maintains the diversity of network features, thus enhancing the quality of video super-resolution.

In the internal design of the network, a base module (BM) is composed of two residual blocks and a regional focus block, in which residual blocks transfer rich image details from the previous layer to the latter in a cascading manner while avoiding the problem of gradient disappearance. The regional focus block carries out regional focus to some feature information while retaining reconstruction details. After that, the output results of each BM module are spliced by skip connection to maintain the diversity of features in the reconstruction process of the network. Finally, the aggregated features are used for network output reconstruction results and hidden state updates.

3.1. Network Structure

The network structure designed in this paper mainly consists of three parts, which are the positive feedback part as the input, the basic module part for feature regional focus and multi-feature fusion, and the output part for updating and reconstructing the hidden state. Their structure relationship is shown in Figure 2.

The leftmost part of Figure 2 is the input part of the network; in the input part of the network, the adjacent LR frames x_{t-1} and x_t, the reconstruction result y_{t-1} of the previous frame along the channel axis and the hidden state h_{t-1} generated in the reconstruction process are spliced, and then sent to the network. For the video super-resolution task, if the temporal continuity of the network model is limited, the reconstruction results are prone to flicker distortion. The hidden state is spliced with the adjacent LR frames and fed to the network, which helps to improve the continuity between frames and reduce the occurrence of flicker and distortion. The reconstruction result y_{t-1} of the previous frame is not directly spliced, but is obtained at the same size as the input frame after the space-to-depth transformation, and then spliced as the network input. Due to the high correlation between adjacent frames, when the reconstruction result y_t is generated, the reconstruction result of the previous frame contains the additional texture information that has been processed at the previous time, which can supplement the information of the similar part of the current frame. In addition, the preceding LR frame x_{t-1} is input into the network along with the co rresponding rebuild output y_{t-1}. This input–output mapping relationship can guide the content that needs to be learned for the current frame reconstruction.

As shown in Figure 2, then 3×3 convolution is used to extract features from the input network concatenation information, and the extracted features are sent into the BM module. The BM module consists of two residual blocks and one regional focus block. The residual block here can extract high-frequency information contained in the feature information, and the regional focus block can decouple from the next BM module and focus on areas that are difficult to recover in the video frame. Through the feature aggregation structure, the high-frequency feature information extracted from different BM modules and the regional feature information of focus are sent to the last layer of the network for concatenation. After the feature fusion of 1×1 convolution, the feature information is sent to the two

outputs of the network, respectively. One channel is directly output to hidden state h_t after 3×3 convolutions. In other channel, the result after a depth-to-space transformation is added with the up-sampling result of the LR frame x_t after bicubic interpolation as the super-resolution output of the current frame y_t. The LR frame x_t contains the profile information of the frame to be reconstructed. The interpolation and up-sampling of this frame is to reduce the network's estimation of the profile information, put more computing resources into the recovery of details, and reduce the network's computation.

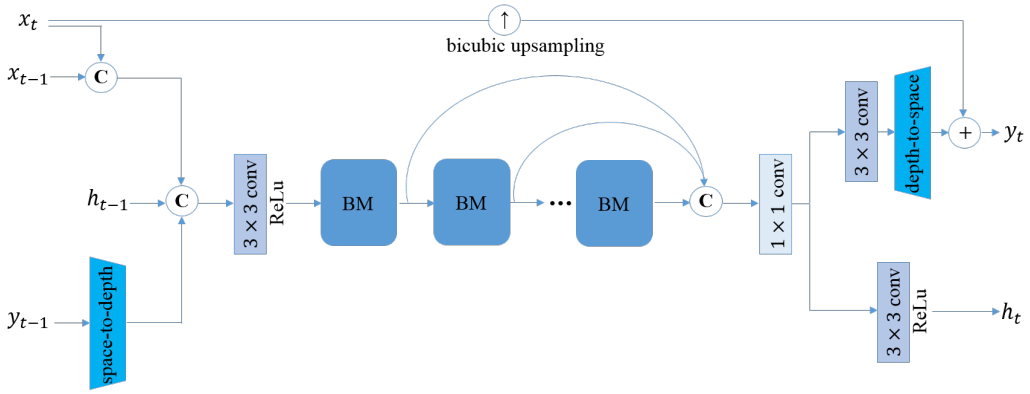

Figure 2. The video super-resolution recurrent network structure proposed in this paper.

When the reconstruction result y_{t-1} of the previous frame is sent to the network as positive feedback to participate in the reconstruction of the next frame, a space-to-depth transformation needs to be performed to match the size of the input low-resolution frame. The output part of the network is transformed from depth to space. After transformation, all pixels contained in the low-resolution feature are rearranged in the corresponding high-resolution feature area, preserving local integrity. The depth-to-space and space-to-depth transformations are inverse to each other. The depth-space transformation can reduce the complexity of the model without losing local information, as shown in Figure 3.

Equation (1) represents the process:

$$[y_{LR}]^{H \times W \times 2s^2} \leftrightarrow [y_{HR}]^{sH \times sW \times 2} \tag{1}$$

where $H \times W$ represent the size of the low-resolution frame, s represents the reconstruction factor, and y_{LR} and y_{HR} represent the low-resolution feature and high-resolution features, respectively.

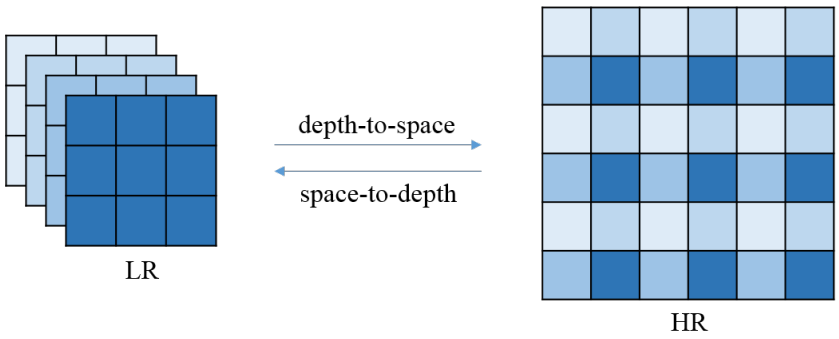

Figure 3. Depth–space transformation.

In the designed network structure of this paper, the current frame x_t to be reconstructed can be spliced with the hidden state of the positive feedback network, and the reconstruction result can be obtained from previous frame, and then input into the network. In addition, we also designed a skip connection structure for the current frame to be reconstructed, which adds the input network frame x_t to be reconstructed with the output reconstruction

results via bicubic interpolation up-sampling. The high-frequency information of the frame to be reconstructed is seriously lost, but it contains a lot of low-frequency information. The skip connection after interpolation and up-sampling can guide the network to pay maximum attention to the high-frequency components of the reconstructed frame, avoid the complex transformation of the network learning from the LR frame to the HR frame, and greatly simplify the training difficulty of the network.

3.2. Residual Calculation and Regional Focus in the BM Module

After the different information input to the network is spliced, the convolution operation is first performed to extract different feature information, and then the information is sent to the BM basic module of the network. Each BM base module in the designed network structure contains two residual blocks and one regional focus block, as shown in Figure 4. The cascade of residual blocks can transfer rich image information from the previous layer to the next layer. The regional focus block (RF-Block) can adaptively focus on the features of specific regions, and extract the feature information that needs to be focused on in different regions.

The calculation process of the first BM block in Figure 4 can be described by mathematical Formula (2); the calculation of other BM blocks is the same as this process:

$$F_1 = RF\{Res\{Res\{F_0\}\}\} \tag{2}$$

where F_0 is the input feature, $Res\{\cdot\}$ is the residual feature calculation, $RF\{\cdot\}$ is the regional focus feature calculation, and F_1 is the output feature of the first BM block. After all BM blocks are calculated, the output feature F_i is obtained to participate in the subsequent work of the network.

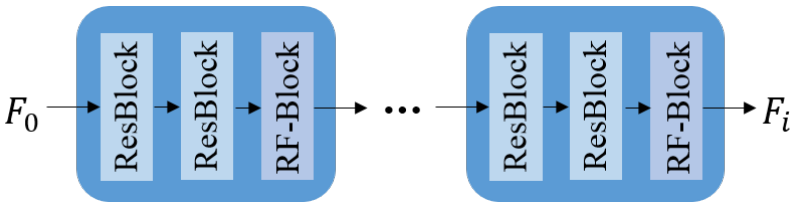

Figure 4. The structure of the base module (BM).

The cascade of residual blocks can extract the feature information contained in the image. With the increase in network layers, the deep feature information of the input information can be further extracted, and the skip connections of residual blocks can alleviate the gradient disappearance problem with the increase in network depth. The task of video super-resolution is to restore the missing information to the complete information. In the traditional network structure, the information will be lost in the transmission process, which will affect the reconstruction quality. The skip connection structure can ensure the integrity of information transmission in the network and avoid information loss.

The main task of video super-resolution reconstruction is to minimize the pixel-by-pixel error between the reconstruction result and the high-resolution label. Video frames contain smooth regions with single details, and complex regions with rich textures. In the reconstruction process of the video frame, the loss values of smooth regions and label frames are small, but the loss values of complex regions and label frames are quite different. Compared with complex regions, smooth regions account for a larger proportion in video frames. This data imbalance leads to the optimization direction of the network being dominated by smooth regions, resulting in the unsatisfactory reconstruction results of complex regions. Therefore, we add regional focus blocks to the BM. During the network training process, the regional focus module is used to focus on different regions in the process of video frame reconstruction, and not just simple regions.

The structure of the regional focus block is shown in Figure 5. After the feature F_n is input into the regional focus block, step convolution is first used to reduce the dimension of

space and channel for the feature F_n. The down-samplings of space and channels ensures the lightweight nature of the module. Meanwhile, not all features require a regional focus. The dimensionality reduction for feature channels can adaptively select the feature channels that need to be focused. Then, the large receptive field is obtained by pooling the large-size max, and the significant features in the space are obtained by connecting the convolution layer. The significant features are restored to the input size using bilinear interpolation and 1×1 convolution, and the regional focus weights are obtained by the sigmoid function. The result of multiplying the weight by the input feature F_n is spliced with the input feature F_n in the channel dimension, and the new feature F_m after regional focus is obtained through 1×1 convolution fusion. The regional focus block not only retains the details learned currently, but also performs regional focus on some features.

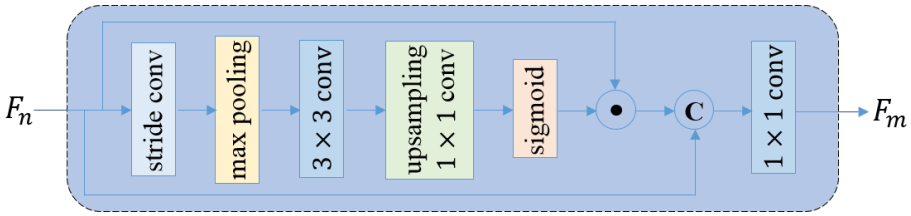

Figure 5. The structure of the regional focus block (RF-Block).

The calculation process in Figure 5 can be expressed using Formula (3) as follows:

$$F_m = Conv_{1\times1}\{Concat\{F_n, F_n \cdot Mask\}\} \tag{3}$$

where F_n represents the feature of the input regional focus block (RF-Block). *Mask* represents the focus weight calculated by the pooling, up-sampling, and sigmoid functions. $Concat\{\cdot\}$ represents the splicing operation, $Conv_{1\times1}\{\cdot\}$ represents fusing the features with 1×1 convolution kernels, and F_m represents the output features.

3.3. Feature Aggregation

The results of the regional focus block are sent to the lower BM module as the input features. This serial structure determines that the features participating in the final reconstruction results are deep-seated features, while the shallow features can not effectively participate in the network reconstruction, thus reducing the learning efficiency of the network. Therefore, we have designed a skip connection structure in the network, which connects the shallow features of the regional focus block output to the last layer of the network. In addition to the serial structure, there is also a parallel structure between the features of each layer, which sends the features of different regional focus blocks to the last layer of the network for splicing with the deep features. The dimension of splicing results is reduced through 1×1 convolutions, and the aggregation features obtained can be used to reconstruct the final result and update the hidden state. We call this structure Feature Aggregation (FA), as shown in Figure 6.

We use the calculation Formula (4) to express the feature aggregation calculation process in Figure 6, as follows:

$$F_s = Conv_{1\times1}\{Concat\{F_1, F_2, F_3, F_4, F_5\} \tag{4}$$

where F_1, F_2, F_3, F_4, F_5 represents the output characteristics of different BMs. $Concat\{\cdot\}$ represents the splicing operations. $Conv_{1\times1}\{\cdot\}$ represents the fuse features with 1×1 convolution kernels. F_s represents the output features, which participate in the final result reconstruction and the hidden state update.

By splicing the shallow features and deep features via skip connection, the number of features involved in video frame reconstruction can be increased without increasing the network parameters. There are n BMs in the network, and the number of features is expanded by N times. The number of BMs is set to 5 in our network. The feature

aggregation structure takes into account the diversity of features while retaining the shallow features to participate in network reconstruction. In the process of network reconstruction, the shallow network mainly focuses on the simple areas and smooth areas of the video frame to be reconstructed, and the deep network mainly focuses on the complex areas and texture areas of the video frame to be reconstructed.

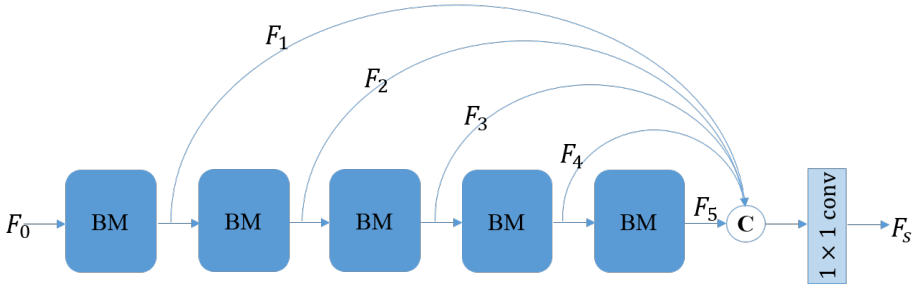

Figure 6. Feature aggregation structure.

Feature aggregation is essentially a feature selection process. Under the function of a regional focus block, the features of different layers focus on the regions with different reconstruction difficulties of the video frame. The skip connection structure sends the different features output from different BMs to the last layer of the network, and then splices them for convolution and fusion, so as to improve the availability of parameters. During network training, the loss between the reconstruction results and the label frames directly affects each regional focus block through the skip connection structure of feature aggregation. The regions with poor reconstruction results can be directly fed back to the regional focus block, which can be optimized in the subsequent network, thereby obtaining better reconstruction results.

Finally, the aggregation features that contain rich feature information are divided into two ways. One is the convolution operation, depth-to-space transformation, and the interpolated up-sampled frame to be reconstructed, which is added as the reconstruction result y_t of the current frame output network. The others is the result with convolution and nonlinear operation as the hidden state h_t output network. Then, the reconstruction result y_t, the hidden state h_t, the next frame to be reconstructedm and the preamble frame of this frame are sent to the network together to start the reconstruction of the next frame, and the process is repeated until the reconstruction of all video frames is completed.

4. Experiments

4.1. Experimental Datasets

The training dataset in this paper adopts the Vimeo-90K dataset [6], which consists of 91,701 short video sequences. Each video sequence contains seven video frames, and the resolution of each video frame is 448 × 256. Vimeo-90K datasets can be used for video denoising, deblocking, and super-resolution tasks. The 7K videos were selected as the validation dataset in this dataset. The task of video super-resolution reconstruction is to reconstruct the detailed information of the video frame while enlarging the low-resolution video frame by r times. The low-resolution video frames in the training data are obtained by using Gaussian fuzzy kernel with standard deviation $\sigma = 1.6$ to down-sample the existing high-resolution video frame labels by r times. The test data adopts the Vid4 [46] and UDM10 [11] datasets; the Vid4 dataset contains four different video sequences, and each video sequence contains different scene information and motion patterns. The UDM10 dataset contains video sequences of 10 different scenes, and each video sequence consists of high-definition video frames with a resolution of 2K.

4.2. Experimental Details

Label video frames are first down-sampled using a Gaussian fuzzy kernel of $\sigma = 1.6$ before being sent to the network. Since we mainly verify the results of the ×4 times video

frame reconstruction in this paper, we obtain low-resolution video frames via the $\times 4$ times down-sampling of label frames. During the training process, a video frame is randomly extracted from a low-resolution video sequence, and then a 64×64 patch is randomly cropped into the network, and the corresponding region is cropped in the labeled video frame for the input network. We use random flips and mirrors to increase the amount of training data, which can improve the generalization ability of the network.

Because our network is a recurrent structure, we initialize both the previous frame reconstruction result and the hidden state to zero. The network model is trained for a total of 130 epochs, and the learning rate of each 60 epochs decays to 0.1 times of the original. Here, we set the learning rate to 1×10^{-4}. The Adam optimizer is used in the training process, and we set $\beta_1 = 0.9$, $\beta_2 = 0.999$, and the weight decay rate is 5×10^{-4}. The experiment was performed on NIVIDIA GTX 1080Ti GPU in Python 3.6.4 and Pytorch 1.1. The Peak Signal to Noise Ratio (PSNR) and Structural Similarity (SSIM) were used as the evaluation indexes. All the evaluation indexes were performed on the brightness channel of the video frame YCbCr space.

4.3. Ablation Experiment

In order to further explore the influences of different modules of the recurrent network structure proposed in this paper on the final video super-resolution reconstruction results, we designed a network structure containing different modules for ablation experiments to verify the role of each module in the network and its impact on the final reconstruction results.

We conducted the ablation experiment on the Vid4 dataset. Firstly, the regional concern block and feature aggregation structure in the network structure were removed, and a recursive network structure containing only 10 residual blocks with series structure was adopted for the super-resolution reconstruction of datasets. Secondly, the feature aggregation structure in the network structure was removed, and the regional focus blocks were reserved. Thirdly, all regional focus blocks in the network structure were removed, and the feature aggregation structure was retained. Finally, the dataset was tested with a complete network structure, including regional focus blocks and feature aggregation structures. The experimental results are shown in Table 1.

Table 1. Ablation experiments of our method on the Vid4 dataset, by removing different modules in the network and comparing the functions of different modules in the network structure.

Method	PSNR	SSIM
Ours w/o FA + RF-blocks	27.69	0.848
Ours w/o feature aggregation	27.91	0.849
Ours w/o regional focus blocks	27.70	0.849
Ours	28.39	0.858

After removing the regional focus blocks and feature aggregation structure, the network structure in this paper degenerates into a recurrent network structure with only concatenated residual blocks. It can be seen from Table 1 that the reconstruction result index of this network is the lowest. In the network that removes the feature aggregation structure and retains the regional focus blocks, the reconstruction results of the video frames are greatly improved. It shows that the regional focus block does pay attention to multiple features of different regions in the video frame, which increases the features diversity involved in reconstruction and improves the quality of the network reconstruction results.

In the network that removes the regional focus blocks and retains the feature aggregation structure, the index of the reconstruction results is similar to that of the reconstruction results of that are completely degenerated into a recursive network with only serial residual blocks. This is because the skip connection method of feature aggregation does not completely change the direct relationship between the output of the previous layer and the input of the next layer. Although the shallow features are skip-connected to the last

layer to participate in the reconstruction, the features input to the deep network does not change, and the network can choose only the deep features of the last layer instead of the shallow features during the training process. The network does not use shallow features and degenerates into a series residual network, resulting in similar reconstruction results of the two networks.

The last line of Table 1 is the result of the method in this paper. Different features extracted from the regional focus block are skip-connected to the last layer of the network through the feature aggregation structure, and spliced with the deep features to participate in the network reconstruction, and to obtain better reconstruction results.

A Reasonable network model structure design can improve the quality of reconstruction results to different degrees. In addition, simply increasing the depth of the network can expand the receptive field of the convolution kernel, and extract deeper feature information to participate in the reconstruction of the final result, which is helpful to improve the quality of the reconstruction result, but meanwhile, the complexity of the model and the amount of parameters will also increase. In order to explore the influence of model parameters on the reconstruction results, we designed a comparative experiment in Table 2 using Vid4 as the experimental dataset.

Table 2. The influence of the parameters of the network model on the reconstruction results. The results show that the performance of the network model cannot increase indefinitely with the increase in the number of network model parameters.

Method	Numbers of Residual Blocks	Numbers of Regional Focus Blocks	Param	PSNR/SSIM
Ours w/o FA + RF-blocks	10	0	3.4M	27.69/0.848
Ours w/o FA + RF-blocks	12	0	3.9M	27.85/0.849
Ours w/o Feature Aggregation	10	5	3.8M	27.91/0.849

After removing the regional focus blocks and feature aggregation structures, our method degenerates to a recurrent network structure in which the number of residual blocks is 10, as shown in Table 2. Although the structure has fewer parameters, the reconstruction will result in the lower index. The method marked with '$*$' also removes the regional focus block and feature aggregation structure; the difference is that the number of residual blocks is set to 12. This structure increases the depth of the network and the amount of parameters, which brings a small improvement to the quality of the reconstruction results of the model. The last is the reconstruction result of our method that removed the feature aggregation structure. Because of the introduction of regional focus blocks, the model parameters of the network also increase, but they are still less than the parameters amount of the network structure with 12 residual blocks in series. At the same time, the quality of the network reconstruction results is higher than that of the network structure with 12 residual blocks in series. This experiment demonstrates that the proposed regional focus block in the network structure can achieve higher reconstruction benefits with a small increase in the number of model parameters. However, it is not that the larger the number of parameters contained in the network model, the higher the performance of the network model will be. In Table 2, the number of parameters in the second row and the third row differed by 0.1M, but the index of reconstruction results with a large number of parameters are lower. In the method of this paper, we need to balance the relationship between the quantity of network parameters and the quality of network reconstruction. Increasing the number of modules in the network blindly may lead to the opposite results.

The network structure of this paper adds a feature aggregation structure on the basis of the regional focus block. This feature aggregation structure does not increase additional parameters, which makes our method obtain great benefits with a small increase in model complexity. The interaction between the regional focus blocks and the feature aggregation structures enriches the diversity of features participating in the reconstruction, and improves the quality of the reconstruction results.

To further demonstrate the role of regional focus blocks in the video frame reconstruction process, we visualize the features extracted from the regional focus blocks. As shown in Figure 7, the green highlighted part is the attention mask learned by the regional focus block. We selected the 'walk' and 'foliage' video sequences from the Vid4 dataset, and the 'caffe' and 'photography' video sequences from the Udm10 dataset. The video sequences 'foliage' and 'caffe' were taken under a fixed lens, and the car and the raised coffee cup in the video frame are constantly moving and changing areas, which are difficult to recover. The regional focus mask captures these two regions very accurately; especially in the 'foliage' video sequence, for the same car region, the stationary white car is not captured by the regional focus block, but the moving black car is accurately captured. The video sequences 'walk' and 'photography' were taken with motion shots, and the moving character areas in the video sequences can also be accurately captured. Through the display of feature visualization, we can see that the regional focus block can focus on the regions that are difficult to recover in the video frame, which can guide the subsequent network to learn to different degrees for different reconstructed regions, thereby improving the reconstruction quality of these regions.

Figure 7. Feature visualization of the regional focus block; the green part shows the attention paid to different reconstruction areas of different video frames.

4.4. Comparative Experiment

We test our proposed method on the datasets Vid4 and Udm10, and compare them with the existing video super-resolution methods TOFLOW [6], FRVSR [4], DUF [9], RBPN [8], RLSP [7], PFNL [11], and RRN [19]. Here, low-resolution video frames are obtained by down-sampling a Gaussian fuzzy kernel with a standard deviation of $\sigma = 1.6$. The FRVSR method uses explicit motion estimation and compensation to reconstruct video frames. The networks designed using the DUF and PFNL methods transfer motion information in an implicit way. The RBPN method does not use explicit motion compensation, but it calculates the optical flow and inputs it to the network as additional information. The RLSP method adopts a cascaded convolution structure, and the RRN method adopts a cascaded residual block structure. Our method also adopts a cascade structure. The difference is that our method not only cascades residual blocks and regional focus blocks, but it also designs a feature aggregation structure to increase the feature richness of the video frames involved in reconstruction. The quantitative comparison results are shown in Table 3. We calculated PSNR and SSIM in the luminance channel of the YCbCr color space and the RGB color space, respectively. In addition, we also calculated the amount of the model parameters and the average time taken by the network to reconstruct a frame of video.

It can be seen from Table 3 that the sliding window methods DUF, RBPN, and PFNL are more temporal-consuming than the recurrent structure methods FRVSR, RLSP, and RRN. Therefore, our method adopts a recursive structure, which takes about the same time as other recursive structure methods. Meanwhile, our method has a better performance in PSNR and SSIM than for other methods, which objectively proves the effectiveness of our proposed method.

Table 3. Comparison of results with different methods in Vid4 and Udm10 datasets.

Method	Param (M)	Runtime (ms)	Vid4 (Y)	Vid4 (RGB)	Udm10 (Y)	Udm10 (RGB)
TOFLOW [6]	1.4	1658	25.85/0.765	24.39/0.743	36.26/0.943	34.46/0.929
FRVSR [4]	5.1	129	26.68/0.810	25.01/0.791	37.09/0.952	35.39/0.940
DUF [9]	5.8	1393	27.38/0.832	25.91/0.816	38.48/0.960	36.78/0.951
RBPN [8]	12.8	3482	27.17/0.820	25.65/0.799	38.66/0.959	36.53/0.946
RLSP [7]	4.3	50	27.48/0.838	25.69/0.815	38.48/0.960	36.39/0.946
PFNL [11]	3.0	295	27.16/0.836	25.67/0.818	38.74/0.962	36.91/0.952
RRN [19]	3.4	45	27.69/0.848	26.16/0.820	38.96/0.964	37.03/0.953
Ours	3.9	58	28.39/0.858	26.81/0.839	39.69/0.967	37.42/0.955

We select representative 'calendar' and 'city' video sequences from the Vid4 dataset to visually demonstrate our method. As shown in Figure 8, for the 'calendar' video sequence, we intercepted complex texture regions and character regions. Our method can completely restore the texture regions, and the outlines of character regions are also clearer than other methods. For the 'city' video sequence, our method obtains clear building textures and outlines compared with other methods, and the reconstruction results of other methods have blurred or unclear building outlines.

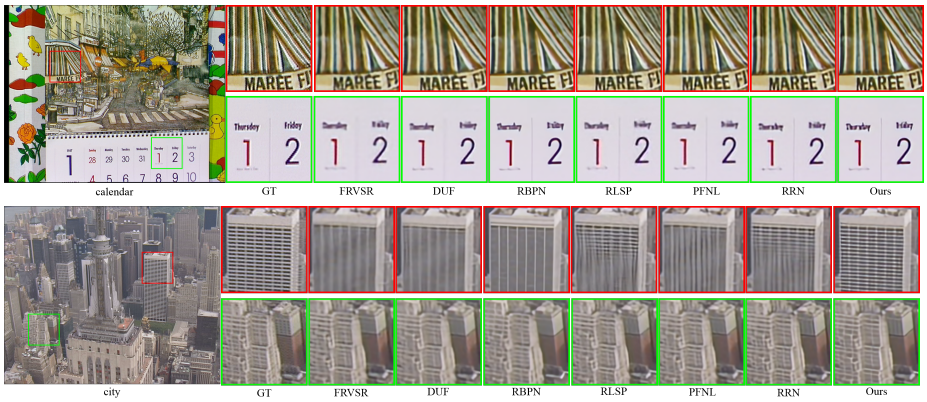

Figure 8. Comparison of reconstruction results with different methods in the Vid4 dataset.

We select representative 'archpeople' and 'auditorium' video sequences from the Udm10 dataset to visualize the reconstruction results of our method. As shown in Figure 9, for the 'archpeople' video sequence, our method is superior to other methods in reconstructing the details of human hair and the contour details of sunglasses. For the 'auditorium' video sequence, other methods have lost details in the reconstruction of the horizontal axis connection part in the shelf. Our method completely reconstructs this part, and the reconstruction result is similar to the label. The comparison proves that the results of this method are of higher reconstruction quality and give a better visual experience than those of other methods.

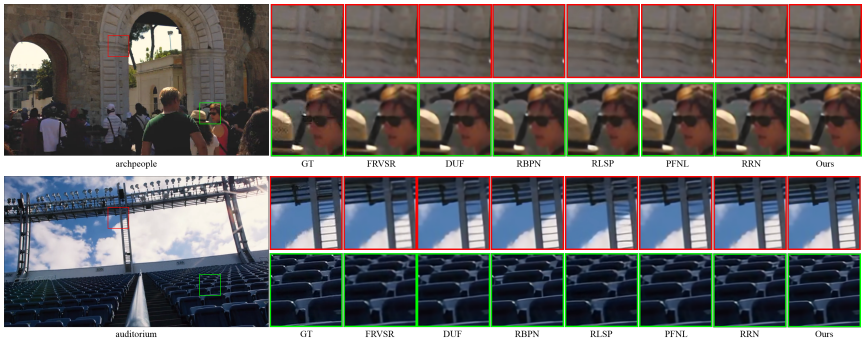

Figure 9. Comparison of reconstruction results with different methods in the Udm10 dataset.

5. Conclusions

A video super-resolution method with a recurrent network structure is proposed in this paper. In addition to the cascade of residual blocks in our network structure, we also designed a regional focus block to focus on the areas with different reconstruction difficulties in the video frames to be reconstructed. The feature extractions of these areas overcome the problem of feature limitation in traditional cascade residual network. Then, the different extracted features are fused through the feature aggregation structure, which increases the diversity of features involved in video frame reconstruction without excessively increasing the network parameters, and effectively improves the quality of the reconstruction results. Experiments on the test datasets show that the reconstruction results of our method obtain a higher index and a better visual experience, which verifies the validity of this method.

Author Contributions: Y.L. and L.H. conceived and designed the whole experiment. Y.L. designed and performed the experiment, and wrote the original draft. H.Z. contributed to the review of this paper. L.H. and D.W. participated in the design of the experiments and in the verification of the experimental results. J.S. participated in the review and revision of the paper, and provided funding support. J.W. contributed to the review of this paper and provided funding support. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported in part by the Research and Development of Manufacturing Information System Platform Supporting Product Life Cycle Management No. 2018GY-030, in part by the Natural Science Foundation of Shaanxi Province No. 2021JQ-487, and in part by the Scientific Research Program Funded of Shaanxi Education Department No. 20JK0788.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used to support the findings of this study are included within the paper.

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

References

1. Liu, H.; Xiong, R.; Qiang, S.; Feng, W.; Wen, G. Image super-resolution based on adaptive joint distribution modeling. In Proceedings of the 2017 IEEE Visual Communications and Image Processing (VCIP), St. Petersburg, FL, USA, 10–13 December 2017.
2. 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.
3. 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]
4. 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.
5. 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; pp. 4472–4480.
6. 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]
7. Fuoli, D.; Gu, S.; Timofte, R. Efficient video super-resolution through recurrent latent space propagation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Republic of Korea, 27–28 October 2019; pp. 3476–3485.
8. Haris, M.; Shakhnarovich, G.; Ukita, N. Recurrent back-projection network for video super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 3897–3906.
9. 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, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3224–3232.
10. 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 Workshops, Long Beach, CA, USA, 16–17 June 2019.

11. 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, Republic of Korea, 27–28 October 2019; pp. 3106–3115.
12. Kim, S.Y.; Lim, J.; Na, T.; Kim, M. 3dsrnet: Video super-resolution using 3d convolutional neural networks. *arXiv* **2018**, arXiv:1812.09079.
13. 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]
14. Song, Q.; Liu, H. Deep Gradient Prior Regularized Robust Video Super-Resolution. *Electronics* **2021**, *10*, 1641. [CrossRef]
15. Huang, Y.; Wang, W.; Wang, L. Bidirectional recurrent convolutional networks for multi-frame super-resolution. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, USA, 7–12 December 2015; Volume 28.
16. Isobe, T.; Jia, X.; Gu, S.; Li, S.; Wang, S.; Tian, Q. Video super-resolution with recurrent structure-detail network. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 645–660.
17. 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.
18. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7794–7803.
19. Isobe, T.; Zhu, F.; Jia, X.; Wang, S. Revisiting temporal modeling for video super-resolution. *arXiv* **2020**, arXiv:2008.05765.
20. 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.
21. Kim, T.H.; Sajjadi, M.S.; Hirsch, M.; Scholkopf, B. Spatio-temporal transformer network for video restoration. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 106–122.
22. Xiao, Z.; Fu, X.; Huang, J.; Cheng, Z.; Xiong, Z. Space-time distillation for video super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2113–2122.
23. Wang, Z.; Yi, P.; Jiang, K.; Jiang, J.; Han, Z.; Lu, T.; Ma, J. Multi-memory convolutional neural network for video super-resolution. *IEEE Trans. Image Process.* **2018**, *28*, 2530–2544. [CrossRef] [PubMed]
24. Bao, W.; Lai, W.S.; Zhang, X.; Gao, Z.; Yang, M.H. Memc-net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 933–948. [CrossRef] [PubMed]
25. Li, D.; Liu, Y.; Wang, Z. Video super-resolution using non-simultaneous fully recurrent convolutional network. *IEEE Trans. Image Process.* **2018**, *28*, 1342–1355. [CrossRef] [PubMed]
26. Bare, B.; Yan, B.; Ma, C.; Li, K. Real-time video super-resolution via motion convolution kernel estimation. *Neurocomputing* **2019**, *367*, 236–245. [CrossRef]
27. Kalarot, R.; Porikli, F. Multiboot vsr: Multi-stage multi-reference bootstrapping for video super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019.
28. Chu, M.; Xie, Y.; Mayer, J.; Leal-Taixé, L.; Thuerey, N. Learning temporal coherence via self-supervision for GAN-based video generation. *ACM Trans. Graph.* **2020**, *39*, 75. [CrossRef]
29. Li, W.; Tao, X.; Guo, T.; Qi, L.; Lu, J.; Jia, J. Mucan: Multi-correspondence aggregation network for video super-resolution. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 335–351.
30. Li, Y.; Zhu, H.; Hou, Q.; Wang, J.; Wu, W. Video Super-Resolution Using Multi-Scale and Non-Local Feature Fusion. *Electronics* **2022**, *11*, 1499. [CrossRef]
31. Isobe, T.; Jia, X.; Tao, X.; Li, C.; Li, R.; Shi, Y.; Mu, J.; Lu, H.; Tai, Y.W. Look Back and Forth: Video Super-Resolution with Explicit Temporal Difference Modeling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 17411–17420.
32. Chan, K.C.; Wang, X.; Yu, K.; Dong, C.; Loy, C.C. BasicVSR: The search for essential components in video super-resolution and beyond. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 4947–4956.
33. Chan, K.C.; Zhou, S.; Xu, X.; Loy, C.C. BasicVSR++: Improving video super-resolution with enhanced propagation and alignment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 5972–5981.
34. Ying, X.; Wang, L.; Wang, Y.; Sheng, W.; An, W.; Guo, Y. Deformable 3d convolution for video super-resolution. *IEEE Signal Process. Lett.* **2020**, *27*, 1500–1504. [CrossRef]
35. Chen, J.; Tan, X.; Shan, C.; Liu, S.; Chen, Z. Vesr-net: The winning solution to youku video enhancement and super-resolution challenge. *arXiv* **2020**, arXiv:2003.02115.
36. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 764–773.
37. 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.

38. Yan, B.; Lin, C.; Tan, W. Frame and feature-context video super-resolution. In Proceedings of the AAAI Conference on Artificial Intelligence, Atlanta, GA, USA, 8–12 October 2019; Volume 33, pp. 5597–5604.
39. 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, Long Beach, CA, USA, 16–17 June 2019; pp. 10522–10531.
40. Liu, H.; Zhao, P.; Ruan, Z.; Shang, F.; Liu, Y. Large motion video super-resolution with dual subnet and multi-stage communicated upsampling. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 2127–2135.
41. Zhu, X.; Li, Z.; Zhang, X.Y.; Li, C.; Liu, Y.; Xue, Z. Residual invertible spatio-temporal network for video super-resolution. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 5981–5988.
42. Li, Y.; Zhu, H.; Yu, S. High-Magnification Super-Resolution Reconstruction of Image with Multi-Task Learning. *Electronics* **2022**, *11*, 1412. [CrossRef]
43. Lee, J.; Lee, M.; Cho, S.; Lee, S. Reference-based Video Super-Resolution Using Multi-Camera Video Triplets. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 17824–17833.
44. Liu, H.; Nie, H.; Zhang, Z.; Li, Y.F. Anisotropic angle distribution learning for head pose estimation and attention understanding in human-computer interaction. *Neurocomputing* **2021**, *433*, 310–322. [CrossRef]
45. Liu, H.; Fang, S.; Zhang, Z.; Li, D.; Lin, K.; Wang, J. MFDNet: Collaborative poses perception and matrix Fisher distribution for head pose estimation. *IEEE Trans. Multimed.* **2021**, *24*, 2449–2460. [CrossRef]
46. Liu, C.; Sun, D. On Bayesian adaptive video super resolution. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *36*, 346–360. [CrossRef] [PubMed]