



Article

Position-Guided Multi-Head Alignment and Fusion for Video Super-Resolution

Yanbo Gao ^{1,2} , Xun Cai ^{1,2,*}, Shuai Li ³ , Jiajing Chai ¹ and Chuankun Li ⁴

¹ School of Software, Shandong University, Jinan 250100, China; ybgao@sdu.edu.cn (Y.G.); jiajing_chai@163.com (J.C.)

² Shandong University-Weihai Research Institute of Industrial Technology, Weihai 264209, China

³ School of Control Science and Engineering, Shandong University, Jinan 250100, China; shuaili@sdu.edu.cn

⁴ State Key Laboratory of Dynamic Testing Technology and School of Information and Communication Engineering, North University of China, Taiyuan 030051, China; chuankun@nuc.edu.cn

* Correspondence: caixunzh@sdu.edu.cn

Abstract: Video super-resolution (VSR), which takes advantage of multiple low-resolution (LR) video frames to reconstruct corresponding high-resolution (HR) frames in a video, has raised increasing interest. To upsample an LR frame (denoted by a reference frame), VSR methods usually align multiple neighboring frames (denoted by supporting frames) to the reference frame first in order to provide more relevant information. The existing VSR methods usually employ deformable convolution to conduct the frame alignment, where the whole supporting frame is aligned to the reference frame without a specific target and without supervision. Thus, the aligned features are not explicitly learned to provide the HR frame information and cannot fully explore the supporting frames. To address this problem, in this work, we propose a novel video super-resolution framework with Position-Guided Multi-Head Alignment, termed as PGMH-A, to explicitly align the supporting frames to different spatial positions of the HR frame (denoted by different heads). It injects explicit position information to obtain multi-head-aligned features of supporting frames to better formulate the HR frame. PGMH-A can be trained individually or end-to-end with the ground-truth HR frames. Moreover, a Position-Guided Multi-Head Fusion, termed as PGMH-F, is developed based on the attention mechanism to further fuse the spatial-temporal information across temporal supporting frames, across multiple heads corresponding to the different spatial positions of an HR frame, and across multiple channels. Together, the proposed Position-Guided Multi-Head Alignment and Fusion (PGMH-AF) can provide VSR with better local details and temporal coherence. The experimental results demonstrate that the proposed method outperforms the state-of-the-art VSR networks. Ablation studies have also been conducted to verify the effectiveness of the proposed modules.

Keywords: video super-resolution; multi-head alignment; multi-head fusion



Citation: Gao, Y.; Cai, X.; Li, S.; Chai, J.; Li, C. Position-Guided Multi-Head Alignment and Fusion for Video Super-Resolution. *Electronics* **2024**, *13*, 4372. <https://doi.org/10.3390/electronics13224372>

Academic Editor: Silvia Liberata Ullo

Received: 24 October 2024

Revised: 4 November 2024

Accepted: 5 November 2024

Published: 7 November 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Super-resolution is one of the fundamental tasks in image/video processing [1–17] and has several applications in areas such as general media, vision-based autonomous driving, and hyperspectral images [18–20]. It addresses the problem of how to restore high-resolution (HR) images with their corresponding low-resolution (LR) ones, either from a single image or from a video. For single-image super-resolution (SISR), an HR image is estimated by exploring the self-similarity within the image and the natural image priors for compensating missing details [21–25]. Meanwhile, for video super-resolution (VSR), both the spatial information within each image/frame and temporal information across frames can be used to produce an HR video frame and finally form the whole HR video [26–34]. Since HR video frames contain more details and can be applied in many applications, including video surveillance and high-resolution television, VSR has been

very popular in both the research and industrial communities. Moreover, with the rise of deep neural networks, deep-learning-based VSR is gaining more interest.

To best generate an HR frame corresponding to an LR reference frame, VSR utilizes information from both the LR reference frame and supporting frames (temporal neighboring frames) [10,26,27,30,31]. However, the supporting frames and the reference frame may not be well aligned due to the motion among temporal frames. Therefore, one key step in VSR is to align the supporting frames to the reference frame. To be specific, it is to align the supporting frames to the targeted HR frame corresponding to the reference frame. However, both the reference frame and the supporting frames are of LR frames and thus cannot be directly aligned to the HR frame. Aligning the supporting frames to the LR reference frame and then upsampling to the HR frame certainly results in information loss, which has not been investigated yet.

For temporal alignment, some methods [26–28,30] use explicit motion information such as optical flow to perform the alignment. First, the optical flow for motion estimation is computed and then each supporting frame is warped utilizing the corresponding motion field. Thus, the performance of VSR depends heavily on the optical flow result. However, predicting optical flow itself is challenging and time-consuming. Furthermore, inaccurate flow may introduce extra artifacts around the structures of the generated HR video frames. To address the above issues, some recent VSR methods [1,6,8,10,31,35,36] have explored the motion information in an implicit way using techniques such as dynamic filters, recurrent neural networks, and deformable convolution. The dynamic upsampling filters [6] are designed to utilize motion information among LR frames and directly construct the HR frame by filtering the LR reference frame. TDAN [31] was proposed by introducing deformable convolution into the VSR framework, which aligns the features of the reference frame and supporting frames using learned deformable offsets. EDVR [10] further proposed to perform alignment with deformable convolution on different spatial scales, which is then fused with temporal and spatial attention among the temporal frames and the spatial frame. Deformable-convolution-based methods [10,31,37] have then been widely applied in the VSR task, which is used to align the frames in an adaptive way, and they are suitable for implicit motion compensation.

However, the existing methods, including both the explicit- and implicit-alignment-based approaches, do not explicitly consider the relationship between the aligned features and the HR frame. The objective is to produce the HR frame, and thus the aligned features from the supporting frames are better to directly contribute to the reconstruction of the final HR frame. Accordingly, the existing alignment methods, aligned to the reference frame instead, lead to an inefficient alignment process with suboptimal aligned features.

In this paper, a novel video super-resolution framework with Position-Guided Multi-Head Alignment is proposed, where the temporal neighboring frame features are explicitly aligned to different positions of the HR frame, denoted by different heads. Position-Guided Multi-Head Fusion is then developed to fuse the aligned features in a position-explicit way in order to enhance the features to the corresponding position of an HR frame. The contributions of this paper can be summarized as follows:

- A novel video super-resolution framework with Position-Guided Multi-Head Alignment (PGMH-A) is proposed, which explicitly aligns reference frame features to different heads/positions of an HR frame. PGMH-A can be trained both end-to-end and individually utilizing the ground-truth HR frames.
- A Position-Guided Multi-Head Temporal–Spatial Fusion (PGMH-F) is developed to fuse the multi-head temporal features, and then the fused multi-head temporal features are further aggregated among the heads to construct a spatial feature volume in order to facilitate the extraction of the spatial correlation among different spatial heads.

A high-frequency enhanced block is also used to improve the feature extraction with the high-frequency information of the frames. Extensive experiments have been conducted, and the proposed PGMH-A and PGMH-F framework achieves state-of-the-art performance on the Vid4 and Vimeo-90K-T benchmark datasets.

The rest of this paper is organized as follows. The related work including the SISR and VSR methods is described in Sections 2 and 3, presenting the proposed method with details regarding all the proposed modules. Section 4 provides the experimental results on the proposed method and ablation study on the modules. Finally, Section 5 comprises the conclusion.

2. Related Work

Some deep-learning-based image/video super-resolution methods related to this work are reviewed in this section. First, single-image super-resolution methods [3,4,18,19,38–46] are briefly introduced, and then video super-resolution methods [6,8,10,26–31,37,47–52] are explained.

2.1. Single-Image Super-Resolution

With the rise of deep learning, deep-learning-based SISR methods have been widely explored and outperform most traditional methods, including interpolation-based and dictionary-learning-based methods. SRCNN [3] first used a deep CNN for image super-resolution and achieved good results, showing great potential for deep learning in super-resolving LR images. Since then, many new network architectures [4,18,19,38–43] have been developed to take advantage of deep learning for SISR. Kim et al. [18] proposed to learn the residuals between an HR image and a conventional interpolated HR image with a deep CNN, which achieves significant improvements in accuracy. DRCN [38] proposed to use a deeply recursive CNN, which reuses the weight parameters of the convolutional layers while increasing the receptive field. Shi et al. [39] designed an efficient sub-pixel convolution layer to obtain an HR image/feature from the low-resolution ones, which reduces the computational complexity by processing in a small resolution. With the large dataset DIV2K [53] available, more networks have been developed with better performance, such as RCAN [19], EDSR [40], and RDN [41].

On the other hand, high-frequency information that represents the edges and textures of an image is important to the SR task. To estimate the high-frequency information, frequency decomposition is applied in SISR tasks. Frequency-decomposition-based SISR methods enable a lightweight model to function while preserving comparable performance. Based on the lattice filter bank, LatticeNet [44] designed a lattice block to simulate the Fast Fourier Transformation with the butterfly structure. ESRT [45] designed a high-frequency filtering module based on transformer to extract the high-frequency information at the feature level.

2.2. Video Super-Resolution

Due to the rapidly increasing number of videos, video super-resolution is attracting more and more interest. Although the SISR methods can solve the VSR task by upsampling each video frame independently, they cannot take advantage of the rich temporal information in a video sequence. Compared to SISR, the key to VSR is to explore the temporal complementary information from the neighboring supporting frames to assist the upsampling of the reference frame. To fully explore the useful information in the supporting frames while avoiding introducing irrelevant noise, a crucial issue is that the neighboring supporting frames need be aligned to the reference frame (the desired HR frame, to be specific) accurately. The existing VSR methods can be categorized into two classes based on the way they perform temporal alignment, with explicit or implicit motion compensation.

Explicit-Motion-Compensation-based VSR Methods: Methods with explicit motion compensation [26–30,47] usually adopt a two-stage process using optical flow. The motion among the frames is first estimated with optical flow and then spatial warping is performed with the estimated flow for the alignment. For example, VESPCN [26] applied an efficient spatial transformer network to encode optical flow. FRVSR [47] proposed a framework that combined an optical flow estimation network and super-resolution network to tackle VSR

tasks. TTVSR [54] proposed a trajectory-aware transformer using pre-aligned trajectories. However, this category of methods views VSR as two separate tasks without joint optimization. Moreover, optical flow estimation, as a dense prediction task, is difficult, and using the inaccurate flow for alignment may damage the image structures, especially texture-rich regions, in the generated HR frames, leading to poor alignment quality.

Implicit-Motion-Compensation-based VSR Methods: To avoid the explicit motion compensation process, some methods [6,8,48–51,55] directly explore the spatio-temporal information for feature extraction. In [6], 3D convolutional layers were used to extract the features from the reference frame and supporting frames in order to generate dynamic filters. The dynamic filters were then used to process the frames as implicit motion compensation. TGA [8] proposed to divide the video frames into groups and aggregate them hierarchically within a group and among groups. With an attention module and 3D dense blocks, the groups of information were deeply fused. MuCAN [48] proposed a temporal multi-correspondence aggregation module and a cross-scale non-local correspondence aggregation module to explore the temporal and spatial information. Non-local attention was also adopted for aggregating the spatio-temporal information by [56,57]. BasicVSR and BasicVSR++ [1,35] explored the whole sequence information in a recurrent way with enhanced propagation and alignment using optical flow and deformable convolution.

Deformable convolution was introduced in [10,31,37,52] to accomplish frame alignment. TDAN [31] performed a one-stage temporal-alignment-based deformable convolution, which can align the supporting frames to the reference frame. Inspired by TDAN, EDVR [10] extended the deformable alignment by introducing multi-scale information, which performed the alignment in a spatial pyramid manner in order to better handle large motions. Due to the effectiveness of the enhanced deformable convolution proposed in EDVR, plenty of works [10,31,37] have adopted it for implicit motion compensation. Deformable-convolution-based methods have achieved great performance in VSR and been widely studied. However, the alignment is performed in the latent feature space without any supervision and direct connection to the to-be-reconstructed HR frame. This paper further investigates the deformable-convolution-based methods and solves the above problem to improve the alignment efficiency.

3. Proposed Method

3.1. Overview

Given a low-resolution video with $2N + 1$ consecutive frames $I_{[t-N:t+N]}^{LR}$, the middle frame I_t^{LR} is called the reference frame, and the other temporal neighboring frames are called supporting frames. The goal of VSR is to reconstruct a high-resolution frame \hat{I}_t^{HR} corresponding to the reference frame. With the proposed VSR network f and corresponding parameters θ , the VSR problem can be defined as

$$\hat{I}_t^{HR} = f_{\theta}(I_{[t-N:t+N]}^{LR}) \quad (1)$$

The input $I_{[t-N:t+N]}^{LR}$ is of shape $T \times H \times W \times C$, where $T = 2N + 1$; H and W are the height and the width of the input LR frames, and C is the number of color channels. The HR output \hat{I}_t^{HR} is of shape $rH \times rW \times C$, corresponding to the LR reference frame I_t^{LR} , where r is the upscaling factor.

This paper proposes a Position-Guided Multi-Head Alignment and Fusion (PGMH-AF)-based video super-resolution framework. The overall framework is shown in Figure 1. It is composed of four sub-networks: an HFER module, a PGMH-A module, a PGMH-F module, and a feature enhancement module. The HFER module consists of High-Frequency Enhanced Residual (HFER) blocks [45] to extract high-frequency enhanced features in order to keep the details and rich-texture information in the image. The Position-Guided Multi-Head Alignment (PGMH-A) module is developed to align each supporting frame to the different positions of the HR frame, and the Position-Guided Multi-Head Temporal-Spatial Fusion (PGMH-F) module is developed to perform feature fusion across temporal frames

and heads. The final feature enhancement module processes the aligned and fused features obtained from the above modules to predict the final HR frame. In the following, the PGMH-A and PGMH-F modules are explained.

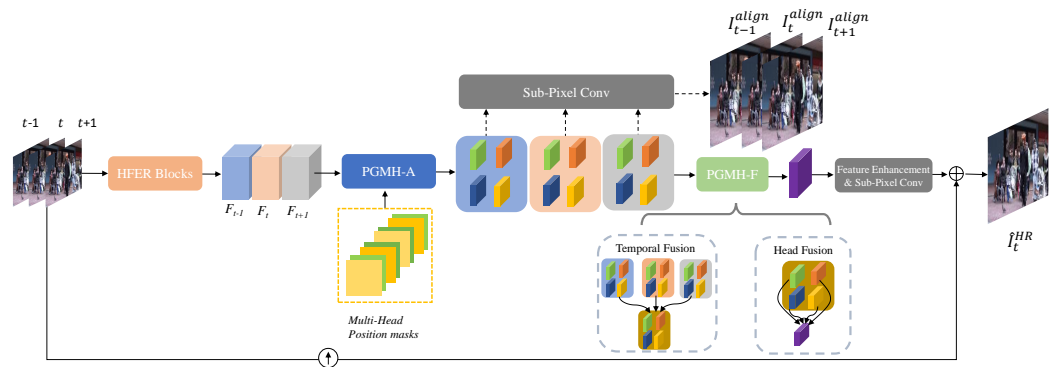


Figure 1. Framework of the proposed PGMH-AF method. Three input frames are used as an illustrative example.

3.2. Position-Guided Multi-Head Alignment

With features extracted for each frame, temporal alignment is then required to align the features of the temporal supporting frames to the current frame or corresponding timestep to explore the temporal information to reconstruct the current HR frame. While the objective is to reconstruct the HR frame, the existing methods only align the temporal features to the reference LR frame or as complete latent features, leading to suboptimal alignment results and thus lowering the overall super-resolution quality. Considering the HR frame is usually generated with a pixel shuffling layer, the HR frame is actually composed of LR frames from different positions of the HR frame, noted as heads in this paper, as shown in Figure 2. Therefore, to address the above problem of ambiguous alignment, in this paper, we propose to explicitly align the temporal features to LR frames corresponding to different heads/positions of the HR frame. The decomposition of the HR frames into LR frames can be obtained by colorredpixel-unshuffling operation, which is the inverse of the pixel-shuffling. It performs downscaling by rearranging an HR tensor T^{HR} of shape $rH \times rW \times C$ into $H \times W \times C \cdot r^2$ tensor T^{LR} . The operation can be expressed as follows:

$$T^{LR}_{\left(\frac{x}{r}, \frac{y}{r}, c \cdot r \cdot \text{mod}(y,r) + c \cdot \text{mod}(x,r)\right)} = T^{HR}_{(x,y,c)} \tag{2}$$

where x, y, c are the output coordinates in HR space. mod here means the modular arithmetic to calculate the remainder.

It is obvious that, for different heads of the HR frame, the temporal supporting LR frames contribute different information. For example, in a video with a slowly moving car from left to right, frame I_{t-1}^{LR} would contribute more to the left head of the HR frame reconstruction and frame I_{t+1}^{LR} for the right head. As shown in Figure 2, simply aligning all the supporting frames to the reference LR frame and then upsampling them to the HR frame (which can also be regarded as an alignment operation, but spatially to the different heads of the HR frame) is inefficient and introduces extra noise. Therefore, a Position-Guided Multi-Head Alignment is developed for each head of the HR frame. Specifically, multi-head features are obtained by aligning the features of supporting frames to different heads separately, shown in the upper part of Figure 2. In this paper, deformable convolution is adopted as a basic module to perform the temporal alignment. One naïve way to conduct the multi-head alignment is to use different networks (one architecture with different learned parameters) to extract different features, generate different deformable offsets, and thus produce different aligned features to different heads. However, for a large upscaling factor, the number of heads corresponding to the LR images is also very large, leading to a large network. On the other hand, the operations and functionality of the different networks are the same, only with different targeted positions in the HR frame. Therefore,

to reduce the complexity and improve the generality of the network, one Position-Guided Multi-Head Alignment (PGMH-A) network is developed. Considering that the offsets of the multi-head features relative to each position of the HR frame are different, explicit position encoding is injected into the generation of the offsets. The proposed PGMH-A is illustrated in Figure 3.

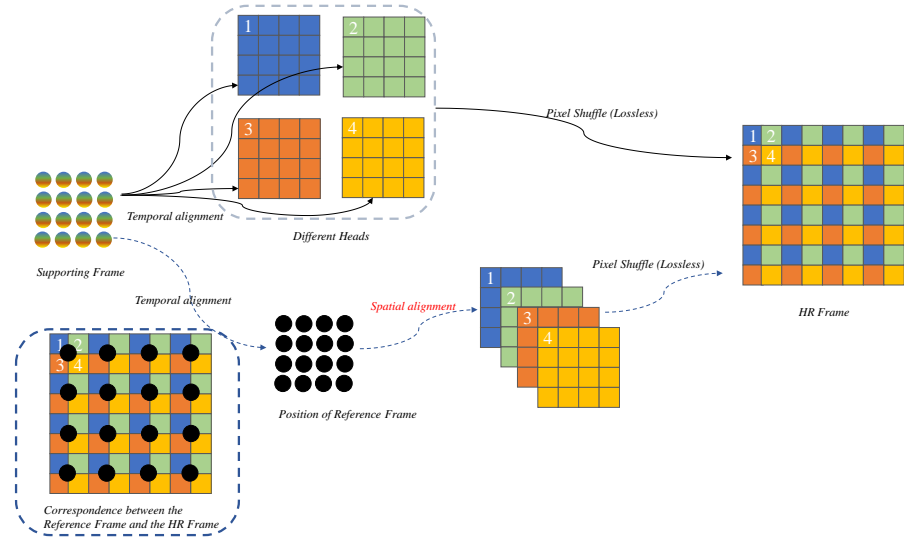


Figure 2. Illustration of the different procedures between the proposed Position-Guided Multi-Head Alignment and the conventional temporal-alignment-based methods.

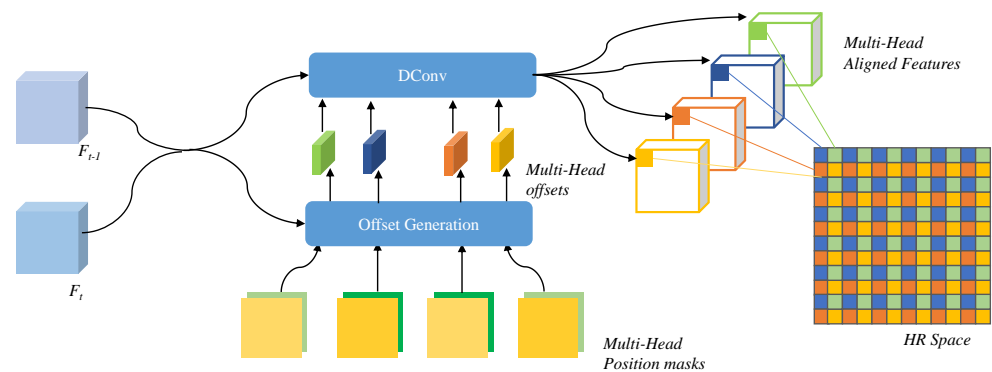


Figure 3. The Position-Guided Multi-Head Alignment (PGMH-A) module. Take $\times 2$ VSR task as an example: four multi-head features are first generated to perform alignment via a shared network with different position masks.

For each head corresponding to a spatial position of an HR frame, a position encoding $M^{(j)}$ is generated first. The $M^{(j)}$ consists of two masks. Take $r \times r$ VSR task as an example: $r \times r$ groups of masks are obtained and each mask is of same element. For the $r \times r$ multi-head features, the corresponding masks are filled with elements $(-\frac{r-1}{2}, -\frac{r-1}{2})$, $(-\frac{r-1}{2} + 1, -\frac{r-1}{2})$, $(-\frac{r-1}{2}, -\frac{r-1}{2} + 1)$, \dots , and $(\frac{r-1}{2}, \frac{r-1}{2})$, respectively, corresponding to the horizontal and vertical relative positions of each head to the reference frame. Elements of mask are then normalized to the range of $[-\frac{r-1}{2 \times r}, \frac{r-1}{2 \times r}]$, which corresponds to the position change in terms of the LR frame pixel distance.

With guidance of the position encoding $M^{(j)}$, the offsets of multi-head features are generated together with features of the supporting frame and reference frame as follows:

$$\Delta P_{t+i}^{(j)} = f_o(F_{t+i}, F_t, M^{(j)}) \quad (3)$$

where three convolutional layers are used as f_o to generate the offsets. Using the offsets, the aligned multi-head features $F_{t+i}^{(j)}$ are then obtained by deformable convolution:

$$F_{t+i}^{(j)} = DConv(F_{t+i}, \Delta P_{t+i}^{(j)}) \tag{4}$$

Collecting all the multi-head-aligned features, the features $F_{t+i}^{align} = [F_{t+i}^{(1)}, F_{t+i}^{(2)}, \dots, F_{t+i}^{(r \times r)}]$ for the supporting frame I_{t+i}^{LR} can be obtained. In this way, the information from the supporting frame can be fully explored for the whole HR frame.

With the features explicitly aligned to different positions of the HR frame, the PGMH-A can be trained with direct supervision, in addition to being trained by the overall VSR objective. To supervise the alignment of each head feature, the sub-pixel convolution is used on the aligned features, which projects the multi-head features into the HR space.

$$\hat{I}_{t+i}^{align} = f_{sub-pixel}(F_{t+i}^{align}) \tag{5}$$

Accordingly, the alignment can be supervised by the ground-truth HR frame. By processing the $2N$ supporting frames separately, the corresponding aligned features of frames $\{F_{t+i}^{align} | t \in [-N, N], i \neq 0\}$ can be obtained for the following processing, and the corresponding generated HR frames $\{\hat{I}_{t+i}^{align} | t \in [-N, N], i \neq 0\}$ can be obtained for the supervision of the multi-head alignment.

3.3. Position-Guided Multi-Head Fusion

The multi-head-aligned temporal features, which correspond to the different positions in the HR space from different supporting frames, are further processed to reconstruct the final HR frame. In order to explore the spatio-temporal information across temporal supporting frames, across multiple heads corresponding to the spatial positions of an HR frame, and across the multiple channels, a Position-Guided Multi-Head Fusion (PGMH-F) module is developed based on the attention mechanism. PGMH-F consists of two sequential processing modules: head-wise temporal fusion and head fusion, as shown in Figure 4.

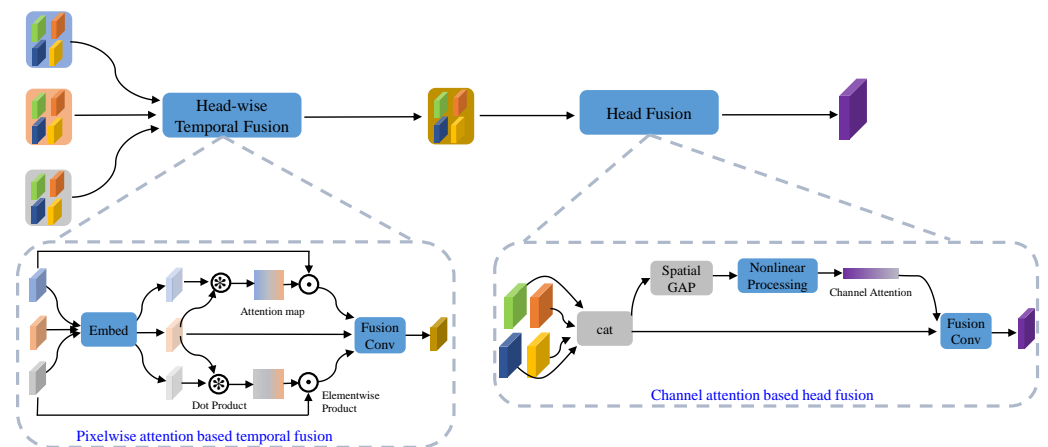


Figure 4. Illustration of the Position-Guided Multi-Head Fusion containing the pixel-wise-attention-based temporal fusion and the channel-attention-based head fusion.

First, features from different supporting frames are fused for each head to combine the temporal information. To reduce the network complexity, the multiple branches for processing the multi-head temporal feature fusion are also shared. However, while the multi-head features are aligned with the deformable offset obtained with the explicit position encoding, the resulting multi-head features are from the features of the supporting frames without position information. Using one shared network cannot obtain the multi-head features appropriate for the different positions of the final HR frame. Thus, to make

the network be informed of the position differences of the multi-head features, the position encoding is further injected into the multi-head feature processing module as well.

For simplification and without loss of generality, taking the processing of one supporting frame as an example, the aligned multi-head features are denoted by $F^{(1)}, F^{(2)}, \dots, F^{(r \times r)}$. Each head feature with the position mask is first processed with a shared convolutional layer, so the position information can be integrated into multi-head features:

$$F^{p(j)} = f_{P-fus}(F^{(j)}, M^{(j)}) \quad (6)$$

By combining the position masks, the network is equipped with the ability to know the position of each head feature and adaptively process the feature close to the divided ground-truth of the HR frame.

A temporal fusion module is then developed to fuse the multi-head-aligned temporal features from the supporting frames together with the features of the reference frame. The temporal fusion layer aims at adaptively aggregating the features in temporal neighboring frames at the pixel level. Thus, a pixel-wise-attention-based temporal fusion scheme is used, as shown in the lower left part of Figure 4. First, a temporal attention mask, which corresponds to the similarity between the aligned features from the supporting frame and the reference frame in an embedding space, is generated. This similarity indicates the feature distance between the supporting frame (temporal neighbors) and the reference frame (at the temporal location of the HR frame) and thus can be used to enhance the features of different supporting frames. The attention mask is then multiplied to the original aligned features of each head in a pixel-wise manner. It can be expressed as

$$\tilde{F}_{t+i}^{p(j)} = F_{t+i}^{p(j)} \odot h_t(F_{t+i}^{p(j)}, F_t^{p(j)}) \quad (7)$$

where $h(F_{t+i}^{p(j)}, F_t^{p(j)})$ denotes the pixel-wise similarity distance between $F_{t+i}^{p(j)}$ and $F_t^{p(j)}$, i.e., the aligned features of the supporting frame and reference frame, respectively. The size of $h_t(F_{t+i}^{p(j)}, F_t^{p(j)})$ is the same as that of $F_{t+i}^{p(j)}$. Two convolutional layers are used to calculate the attention, and Sigmoid is applied as the nonlinear activation function for the last layer. It is worth noting that the temporal fusion is used for each head feature separately since the similarity of each head to the corresponding position in the HR frame can be different. Since the attention is calculated in a manner of self-attention without parameters, this does not increase the number of network parameters.

With the temporal features of each head properly enhanced using the attention, a convolutional layer is further used to perform the fusion operation over the different supporting frames, which is shared over all the heads.

$$F_{t-fus}^{p(j)} = Conv([\tilde{F}_{t-N}^{p(j)}, \dots, \tilde{F}_{t-1}^{p(j)}, F_t^{p(j)}, \tilde{F}_{t+1}^{p(j)}, \dots, \tilde{F}_{t+N}^{p(j)}]) \quad (8)$$

After the temporal fusion, $r \times r$ temporal fused features are obtained from the different heads. While such features can be used to directly construct an HR frame using pixel shuffle, the information among the heads cannot be fully explored. Therefore, multi-head fusion is developed to further fuse the features over the heads, corresponding to the different positions of the HR frame. Different from the temporal fusion that fuses the features from the supporting frame to the reference frame, the multiple heads are fused to construct a spatial volume rather equally corresponding to the final HR frame. Thus, the above temporal attention cannot be simply applied for the head fusion. Moreover, each head consists of different channels of features, and the importance of each channel among the heads to the final reconstruction can also be different. Therefore, we enhance the head features through channels and then perform the final fusion. Particularly, the channel

attention using the squeeze and excitation model [58] is adopted as shown in the lower right part of Figure 4. It can be represented by

$$F_{s-fus} = f_{s-fus} \left(f_{gap}(F_{t-fus}) \odot F_{t-fus} \right) \quad (9)$$

where $F_{t-fus} = (F_{t-fus}^{p(1)}, F_{t-fus}^{p(2)}, \dots, F_{t-fus}^{p(r \times r)})$ represents the head features, f_{gap} represents the squeeze and excitation operation to obtain the global information embedding and perform adaptive recalibration. It contains a global average pooling operation and a nonlinear processing operation. f_{s-fus} is the final convolution to fuse the multiple channels.

Via the proposed PGMH-F with the above two fusion modules, multi-head spatio-temporal features are fused, which can be used to produce the final HR frame. In this paper, the fused features are further processed via cascaded Resblocks as in [10] to enhance the features spatially. Finally, a sub-pixel convolution layer is added to produce the residual information of the HR video frame, which then produces the final HR frame together with a bilinearly interpolated HR base frame (\bar{I}_t^{HR}).

$$\hat{I}_t^{HR} = f_{recon} \left(F_{s-fus} \right) + \bar{I}_t^{HR} \quad (10)$$

3.4. Loss Functions

The proposed PGMH-AF method can be trained in an end-to-end manner. Two loss functions, i.e., L_{multi} and L_{sr} , based on the Charbonnier penalty function [59] explained in the following, are used for supervision. Specifically, L_{multi} is used to optimize the PGMH-A module and L_{sr} is used for the whole network. For PGMH-A, the ground-truth HR frame can be used as the label to supervise each aligned supporting frame:

$$L_{multi} = \frac{1}{2N+1} \sum_{i=-N}^N \sqrt{\|\hat{I}_{t+i}^{align} - I_t^{HR}\|^2 + \epsilon^2} \quad (11)$$

The Charbonnier penalty of the final reconstructed frame is used as the objective function of the whole network:

$$L_{sr} = \sqrt{\|\hat{I}_t^{HR} - I_t^{HR}\|^2 + \epsilon^2} \quad (12)$$

Combining two loss functions, overall loss function for our VSR network training is finally expressed as

$$L = L_{sr} + \lambda L_{multi} \quad (13)$$

where λ is used to balance the two losses and set to 0.01 in the experiments.

4. Experiments

4.1. Datasets and Evaluation Metrics

The most commonly used VSR datasets are Vimeo-90K and Vid4.

Vimeo-90K [30] is a large-scale video dataset used for various video-related tasks, which covers diverse scenes and motions. The VSR subset has 72,436 sequences of 7 frames with 448×256 resolution. Training and testing datasets contain 64,612 and 7824 sequences, respectively.

Vid4 [60] contains long sequences of frames with diverse scenes. It consists of 4 video sequences: City, Walk, Calendar, and Foliage, and the lengths of the sequences are all over 30 frames in the resolution of 720×480 .

Vimeo-90K dataset is used for training along with Vimeo-90K-T and Vid4 for evaluating the performance of the network, similar to [10]. Peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) are adopted as evaluation metrics in our experiments.

4.2. Implementation Details

The VSR on an upsampling scale of $\times 4$ is used for training and evaluation, and same process applies for other scales. Similar to [8,10,30], RGB patches of 256×256 cropped from high-resolution video clips are used as ground-truth frames. Patches of 64×64 are produced with $4\times$ downsampling from the groundtruth patch as the low-resolution input for the network training. A Gaussian blur with a standard deviation of $\sigma = 1.6$ is used, similar to [8,30]. The training data are augmented with horizontal flips, 90° rotations, and frame reversing.

The network is optimized with the Adam optimizer in which $\beta_1 = 0.9$ and $\beta_2 = 0.999$ during training. The initial learning rate is set to 2×10^{-4} and reduced by a factor of 0.1 when validation accuracy no longer improved. The network is supervised by L_{multi} and L_{sr} , and the balance factor λ is set to 1. Batch size is set to 16. Seven consecutive frames (i.e., $N = 3$) are used as inputs. Five HFEblocks are used for feature extraction. The basic channel numbers are set to 128, while the number of multi-head feature channels is set to 48. All experiments are conducted on the PyTorch 1.2.0 platform and Nvidia Tesla V100 GPU.

4.3. State-Of-The-Art VSR Methods Comparison

Our method is compared with one of the state-of-the-art SISR methods, RCAN [19], and several VSR methods using the same sliding-window framework including ToFlow [30], DUF [6], EDVR [10] RBPN [36], PFNL [61], and TGA [10]. Other methods such as BasicVSR [1,35] using a recurrent framework are not compared since different lengths of information are explored with different application scenarios. The quantitative results are shown in Table 1 and 2 for Vid4 and Vimeo-90K-T, respectively.

Table 1. Result comparison of different methods on Vid4 under upscale factor 4 in terms of both PSNR (dB) and SSIM. The best results are highlighted in bold.

Method	Frames	Calendar	City	Foliage	Walk	Average
Bicubic	1	20.39/0.572	25.16/0.602	23.47/0.566	26.10/0.797	23.78/0.634
RCAN [19]	1	22.33/0.725	26.10/0.696	24.74/0.664	28.65/0.871	25.46/0.739
ToFlow [30]	7	22.47/0.731	26.78/0.740	25.27/0.709	29.05/0.879	25.89/0.765
DUF [6]	7	24.04/0.811	28.27/0.831	26.41/0.770	30.60/0.914	27.33/0.831
EDVR [10]	7	24.05/0.814	28.00/0.812	26.34/0.763	31.02/0.915	27.35/0.826
EDVR * [10]	7	24.56/0.833	28.49/0.843	26.48/0.775	30.91/0.918	27.61/0.842
RBPN [36]	7	24.02/0.808	27.83/0.804	26.21/0.757	30.62/0.911	27.17/0.820
PFNL [61]	7	24.37/0.824	28.09/0.838	26.51/0.776	30.65/0.913	27.40/0.838
TGA [8]	7	24.47/0.828	28.37/0.841	26.59/0.779	30.96/0.918	27.59/0.841
Ours	7	24.64/0.837	28.77/0.853	26.66/0.784	31.09/0.921	27.79/0.848

EDVR [10] reported the results in [10], tested with a different setting from ours. EDVR * [10] shows the results with the same setting as ours.

Table 2. Result comparison of different methods on Vimeo-90K-T under upscale factor 4 in terms of both PSNR (dB) and SSIM. The best results are highlighted in bold.

Method	Frames	PSNR/SSIM
Bicubic	1	31.32/0.8684
ToFlow [30]	7	34.83/0.9220
DUF [6]	7	36.37/0.9387
EDVR [10]	7	37.61/0.9489
EDVR * [10]	7	37.41/0.9488
RBPN [36]	7	37.20/0.9458
TGA [8]	7	37.61/0.9489
Ours	7	37.75/0.9517

EDVR [10] and EDVR * [10] are the same as noted in Table 1.

It can be seen that the proposed method achieves the best performance in terms of both PSNR and SSIM compared to the existing methods in the same category. It is also worth noting that our method outperforms EDVR, which is most related to our work and uses the same framework. Some visual results are presented in Figure 5. Compared with EDVR, it can be seen that our method restores more details of video frames than others, demonstrating the effectiveness of the proposed method.

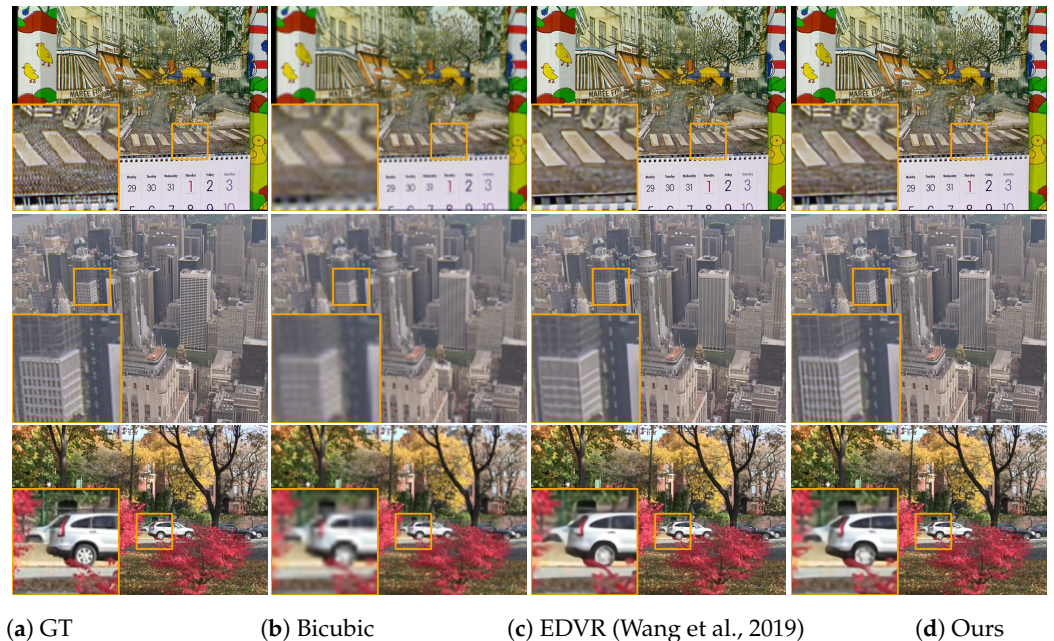


Figure 5. Qualitative comparison for VSR with upscale factor 4 on the Vid4 dataset in comparison with the results of Bicubic and EDVR [10]. Zoom in for best view.

4.4. Ablation Studies

In this section, several ablation studies are conducted to investigate the effectiveness and necessity of the proposed modules in our method. All the studies are conducted on the Vid4 dataset. The baseline is the EDVR network without the temporal–spatial attention (TSA) module using deformable convolution to align the supporting frames for video super-resolution.

Influence of HFER. HFER blocks are used in our network for feature extraction to enhance the high-frequency information by replacing the same number of Resblocks. HFER blocks have shown to be effective for image super-resolution but are not verified for video super-resolution yet, with the changes including moving objects among frames. The results of using HFER blocks are shown in Table 3, noted as Model 1. Compared with the baseline, it can be seen that HFER blocks are also effective for video super-resolution, proving the effectiveness of the high-frequency information for image/video super-resolution.

Table 3. Ablation results of different modules in terms of both PSNR (dB) and SSIM. The best results are highlighted in bold.

Method	HFER	PGMH-A	PGMH-F	PSNR/SSIM
Baseline				27.42/0.8366
Model 1	✓			27.64/0.8446
Model 2	✓	✓		27.71/0.8469
Full	✓	✓	✓	27.79/0.8486

Influence of PGMH-A. The proposed PGMH-A module is used to align the supporting frames to the different positions of the HR frame. To verify its effectiveness, it is further

added on top of the baseline and the HFER for experiments. To isolate its effect, after the Position-Guided Multi-Head Alignment, the temporal aligned features from different supporting frames are simply fused together with a convolution layer, and the features from different heads are also concatenated and processed with a convolution layer. The result is shown in Table 3, noted as Model 2. It can be seen that, with our PGMH-A module, the result is further improved with 0.07 dB, verifying its effectiveness.

Influence of PGMH-F. The proposed PGMH-F module is used to fuse the spatial-temporal features across frames and heads. It can be evaluated by comparing the full model with the above Model 2 using PGMH-A without the attention-enhanced fusion. The comparison is illustrated in Table 3. It can be seen that our PGMH-F also improves the overall performance of the model, validating its effectiveness.

5. Conclusions

This paper proposes a Position-Guided Multi-Head Alignment and Fusion framework for VSR, which effectively explores the information from the temporal supporting frames. PGMH-A reduces the ambiguity in aligning the temporal supporting frames to the target HR frame by explicitly using multiple heads corresponding to the different positions of the HR frame. PGMH-F then fuses the spatio-temporal information in the three dimensions, i.e., temporal, the heads corresponding to the positions of the LR frame, and the channels. In addition, the feature extraction takes advantage of cascaded high-frequency enhanced blocks to enhance the high-frequency information for alignment. Our method is compared with the state-of-the-art VSR methods, and the experiments on two popular benchmark datasets have shown that our PGMH-AF can achieve better performance. Ablation studies have further validated the effectiveness of each module.

6. Limitation and Future Research

While this paper provides an interesting PGMH-AF method, its complexity still needs to be further reduced in order to realize real-time implementation. In the future, more diverse datasets and real-time scenarios will be investigated, especially in terms of reducing and parallelizing the proposed PGMH-AF.

Author Contributions: Conceptualization, Y.G., X.C., and S.L.; methodology, Y.G., X.C., S.L., J.C., and C.L.; software, Y.G., J.C., and C.L.; validation, J.C. and C.L.; formal analysis, Y.G., X.C., S.L., J.C., and C.L.; writing—original draft preparation, Y.G. and J.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China (62271290, 62001092, 62101512, and 62271453), Fundamental Research Program of Shanxi Province (20210302124031 and 202203021212123), and Research Project by Shanxi Scholarship Council of China (2023-131).

Data Availability Statement: Dataset available on request from the authors.

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

Abbreviations

The following abbreviations are used in this manuscript:

PGMH-A	Position-Guided Multi-Head Alignment
PGMH-F	Position-Guided Multi-Head Fusion
PGMH-AF	Position-Guided Multi-Head Alignment and Fusion

References

1. Chan, K.C.K.; 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 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 4945–4954.
2. Zhang, Z.; Peng, B.; Lei, J.; Shen, H.; Huang, Q. Recurrent Interaction Network for Stereoscopic Image Super-Resolution. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 2048–2060. [[CrossRef](#)]

3. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image Super-Resolution Using Deep Convolutional Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 295–307. [[CrossRef](#)] [[PubMed](#)]
4. Haris, M.; Shakhnarovich, G.; Ukita, N. Deep Back-Project Networks for Single Image Super-Resolution. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 4323–4337. [[CrossRef](#)]
5. Zhang, Z.; Lei, J.; Peng, B.; Zhu, J.; Huang, Q. Self-Supervised Pretraining for Stereoscopic Image Super-Resolution with Parallax-Aware Masking. *IEEE Trans. Broadcast.* **2024**, *70*, 482–491. [[CrossRef](#)]
6. 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 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3224–3232.
7. Zhu, Q.; Chen, F.; Liu, Y.; Zhu, S.; Zeng, B. Deep Compressed Video Super-Resolution with Guidance of Coding Priors. *IEEE Trans. Broadcast.* **2024**, *70*, 505–515. [[CrossRef](#)]
8. Isobe, T.; Li, S.; Jia, X.; Yuan, S.; Slabaugh, G.G.; Xu, C.; Li, Y.; Wang, S.; Tian, Q. Video Super-Resolution with Temporal Group Attention. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 8005–8014.
9. Agrahari Baniya, A.; Lee, T.K.; Eklund, P.W.; Aryal, S. Omnidirectional Video Super-Resolution Using Deep Learning. *IEEE Trans. Multimed.* **2024**, *26*, 540–554. [[CrossRef](#)]
10. Wang, X.; Chan, K.C.K.; Yu, K.; Dong, C.; Loy, C.C. EDVR: Video Restoration with Enhanced Deformable Convolutional Networks. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, 16–17 June 2019; pp. 1954–1963.
11. Feng, Z.; Zhang, W.; Liang, S.; Yu, Q. Deep Video Super-Resolution Using Hybrid Imaging System. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 4855–4867. [[CrossRef](#)]
12. Lei, J.; Li, X.; Peng, B.; Fang, L.; Ling, N.; Huang, Q. Deep Spatial-Spectral Subspace Clustering for Hyperspectral Image. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 2686–2697. [[CrossRef](#)]
13. Peng, B.; Zhang, X.; Lei, J.; Zhang, Z.; Ling, N.; Huang, Q. LVE-S2D: Low-Light Video Enhancement from Static to Dynamic. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 8342–8352. [[CrossRef](#)]
14. Guo, C.; Li, C.; Guo, J.; Cong, R.; Fu, H.; Han, P. Hierarchical Features Driven Residual Learning for Depth Map Super-Resolution. *IEEE Trans. Image Process.* **2019**, *28*, 2545–2557. [[CrossRef](#)]
15. Chen, L.; Ye, M.; Ji, L.; Li, S.; Guo, H. Multi-Reference-Based Cross-Scale Feature Fusion for Compressed Video Super Resolution. *IEEE Trans. Broadcast.* **2024**, *70*, 895–908. [[CrossRef](#)]
16. Lei, J.; Zhang, Z.; Fan, X.; Yang, B.; Li, X.; Chen, Y.; Huang, Q. Deep Stereoscopic Image Super-Resolution via Interaction Module. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 3051–3061. [[CrossRef](#)]
17. Chen, P.; Yang, W.; Wang, M.; Sun, L.; Hu, K.; Wang, S. Compressed Domain Deep Video Super-Resolution. *IEEE Trans. Image Process.* **2021**, *30*, 7156–7169. [[CrossRef](#)] [[PubMed](#)]
18. Kim, J.; Lee, J.K.; Lee, K.M. Accurate Image Super-Resolution Using Very Deep Convolutional Networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1646–1654.
19. Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; Fu, Y.R. Image Super-Resolution Using Very Deep Residual Channel Attention Networks. In Proceedings of the ECCV, 15th European Conference, Munich, Germany, 8–14 September 2018.
20. Zhao, T.; Lin, Y.; Xu, Y.; Chen, W.; Wang, Z. Learning-Based Quality Assessment for Image Super-Resolution. *IEEE Trans. Multimed.* **2022**, *24*, 3570–3581. [[CrossRef](#)]
21. Yang, F.; Yang, H.; Fu, J.; Lu, H.; Guo, B. Learning Texture Transformer Network for Image Super-Resolution. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 5790–5799.
22. Zhang, H.; Xiao, J.; Jin, Z. Multi-Scale Image Super-Resolution Via a Single Extendable Deep Network. *IEEE J. Sel. Top. Signal Process.* **2021**, *15*, 253–263. [[CrossRef](#)]
23. He, Z.; Jin, Z.; Zhao, Y. SRDRL: A Blind Super-Resolution Framework With Degradation Reconstruction Loss. *IEEE Trans. Multimed.* **2022**, *24*, 2877–2889. [[CrossRef](#)]
24. Li, F.; Wu, Y.; Bai, H.; Lin, W.; Cong, R.; Zhang, C.; Zhao, Y. Learning Detail-Structure Alternative Optimization for Blind Super-Resolution. *IEEE Trans. Multimed.* **2022**, *25*, 2825–2838. [[CrossRef](#)]
25. Guo, B.; Zhang, X.; Wu, H.; Wang, Y.; Zhang, Y.; Wang, Y.F. LAR-SR: A Local Autoregressive Model for Image Super-Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 1909–1918.
26. Caballero, J.; Ledig, C.; Aitken, A.P.; Acosta, A.; Totz, J.; Wang, Z.; Shi, W. Real-Time Video Super-Resolution with Spatio-Temporal Networks and Motion Compensation. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2848–2857.
27. Liu, D.; Wang, Z.; Fan, Y.; Liu, X.; Wang, Z.; Chang, S.; Huang, T.S. Robust Video Super-Resolution with Learned Temporal Dynamics. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2526–2534.
28. Tao, X.; Gao, H.; Liao, R.; Wang, J.; Jia, J. Detail-Revealing Deep Video Super-Resolution. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 4482–4490.

29. Kim, T.H.; Sajjadi, M.S.M.; Hirsch, M.; Schölkopf, B. Spatio-Temporal Transformer Network for Video Restoration. In Proceedings of the ECCV, 15th European Conference, Munich, Germany, 8–14 September 2018.
30. Xue, T.; Chen, B.; Wu, J.; Wei, D.; Freeman, W.T. Video Enhancement with Task-Oriented Flow. *Int. J. Comput. Vis.* **2018**, *127*, 1106–1125. [[CrossRef](#)]
31. Tian, Y.; Zhang, Y.; Fu, Y.R.; Xu, C. TDAN: Temporally-Deformable Alignment Network for Video Super-Resolution. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 3357–3366.
32. Xu, G.; Xu, J.; Li, Z.; Wang, L.; Sun, X.; Cheng, M.M. Temporal Modulation Network for Controllable Space-Time Video Super-Resolution. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 6384–6393.
33. Chiche, B.N.; Woiselle, A.; Frontera-Pons, J.; Starck, J.L. Stable Long-Term Recurrent Video Super-Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 837–846.
34. 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 (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 17411–17420.
35. 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 (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 5972–5981.
36. Haris, M.; Shakhnarovich, G.; Ukita, N. Recurrent Back-Projection Network for Video Super-Resolution. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–17 June 2019; pp. 3892–3901.
37. Xiang, X.; Tian, Y.; Zhang, Y.; Fu, Y.R.; Allebach, J.P.; Xu, C. Zooming Slow-Mo: Fast and Accurate One-Stage Space-Time Video Super-Resolution. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 3367–3376.
38. Kim, J.; Lee, J.K.; Lee, K.M. Deeply-Recursive Convolutional Network for Image Super-Resolution. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1637–1645.
39. Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A.P.; Bishop, R.; Rueckert, D.; Wang, Z. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1874–1883.
40. Lim, B.; Son, S.; Kim, H.; Nah, S.; Lee, K.M. Enhanced Deep Residual Networks for Single Image Super-Resolution. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 1132–1140.
41. Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; Fu, Y.R. Residual Dense Network for Image Restoration. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 2480–2495. [[CrossRef](#)]
42. Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Aitken, A.P.; Tejani, A.; Totz, J.; Wang, Z.; Shi, W. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 105–114.
43. Li, Z.; Li, G.; Li, T.H.; Liu, S.; Gao, W. Information-Growth Attention Network for Image Super-Resolution. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual Event, 20–24 October 2021.
44. Luo, X.; Xie, Y.; Zhang, Y.; Qu, Y.; Li, C.; Fu, Y.R. LatticeNet: Towards Lightweight Image Super-Resolution with Lattice Block. In Proceedings of the ECCV, 16th European Conference, Glasgow, UK, 23–28 August 2020.
45. Lu, Z.; Liu, H.; Li, J.; Zhang, L. Efficient Transformer for Single Image Super-Resolution. *arXiv* **2021**, arXiv:2108.11084v3.
46. Liang, J.; Zeng, H.; Zhang, L. Details or Artifacts: A Locally Discriminative Learning Approach to Realistic Image Super-Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 5657–5666.
47. Sajjadi, M.S.M.; Vemulapalli, R.; Brown, M.A. Frame-Recurrent Video Super-Resolution. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6626–6634.
48. 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 Computer Vision—ECCV, 16th European Conference, Glasgow, UK, 23–28 August 2020*; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M., Eds.; Springer: Cham, Switzerland, 2020; pp. 335–351.
49. Luo, J.; Huang, S.; Yuan, Y. Video Super-Resolution using Multi-scale Pyramid 3D Convolutional Networks. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020.
50. Xiao, Z.; Xiong, Z.; Fu, X.; Liu, D.; Zha, Z. Space-Time Video Super-Resolution Using Temporal Profiles. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020.
51. Xu, Y.; Gao, L.; Tian, K.; Zhou, S.; Sun, H. Non-Local ConvLSTM for Video Compression Artifact Reduction. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 7042–7051.

52. Deng, J.; Wang, L.; Pu, S.; Zhuo, C. Spatio-Temporal Deformable Convolution for Compressed Video Quality Enhancement. In Proceedings of the AAAI, New York, NY, USA, 7–12 February 2020.
53. Timofte, R.; Agustsson, E.; Gool, L.V.; Yang, M.H.; Zhang, L.; Lim, B.; Son, S.; Kim, H.; Nah, S.; Lee, K.M.; et al. NTIRE 2017 Challenge on Single Image Super-Resolution: Methods and Results. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1110–1121.
54. Liu, C.; Yang, H.; Fu, J.; Qian, X. Learning Trajectory-Aware Transformer for Video Super-Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022.
55. Li, S.; Li, W.; Cook, C.; Zhu, C.; Gao, Y. Independently Recurrent Neural Network (IndRNN): Building A Longer and Deeper RNN. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5457–5466.
56. Yu, J.; Liu, J.; Bo, L.; Mei, T. Memory-Augmented Non-Local Attention for Video Super-Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 17834–17843.
57. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), Red Hook, NY, USA, 2017; pp. 6000–6010.
58. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [[CrossRef](#)] [[PubMed](#)]
59. Lai, W.S.; Huang, J.B.; Ahuja, N.; Yang, M.H. Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5835–5843.
60. Liu, C.; Sun, D. On Bayesian Adaptive Video Super Resolution. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 346–360. [[CrossRef](#)] [[PubMed](#)]
61. 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 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3106–3115.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.