sensors

MDPI

Article

Adapting Single-Image Super-Resolution Models to Video Super-Resolution: A Plug-and-Play Approach

Wenhao Wang, Zhenbing Liu *, Haoxiang Lu, Rushi Lan and Yingxin Huang

School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin 541004, China

* Correspondence: zbliu@guet.edu.cn

Abstract: The quality of videos varies due to the different capabilities of sensors. Video super-resolution (VSR) is a technology that improves the quality of captured video. However, the development of a VSR model is very costly. In this paper, we present a novel approach for adapting single-image super-resolution (SISR) models to the VSR task. To achieve this, we first summarize a common architecture of SISR models and perform a formal analysis of adaptation. Then, we propose an adaptation method that incorporates a plug-and-play temporal feature extraction module into existing SISR models. The proposed temporal feature extraction module consists of three submodules: offset estimation, spatial aggregation, and temporal aggregation. In the spatial aggregation submodule, the features obtained from the SISR model are aligned to the center frame based on the offset estimation results. The aligned features are fused in the temporal aggregation submodule. Finally, the fused temporal feature is fed to the SISR model for reconstruction. To evaluate the effectiveness of our method, we adapt five representative SISR models and evaluate these models on two popular benchmarks. The experiment results show the proposed method is effective on different SISR models. In particular, on the Vid4 benchmark, the VSR-adapted models achieve at least 1.26 dB and 0.067 improvement over the original SISR models in terms of PSNR and SSIM metrics, respectively. Additionally, these VSR-adapted models achieve better performance than the state-of-the-art VSR models.

Keywords: video super-resolution; single-image super-resolution; plug-and-play; deformable convolution

Citation: Wang, W.; Liu, Z.; Lu, H.; Lan, R.; Huang, Y. Adapting Single-Image Super-Resolution Models to Video Super-Resolution: A Plug-and-Play Approach. *Sensors* **2023**, *23*, 5030. https://doi.org/10.3390/s23115030

Academic Editor: Sylvain Girard

Received: 14 April 2023
Revised: 17 May 2023
Accepted: 22 May 2023
Published: 24 May 2023

1. Introduction

Numerous videos are captured every day; however, due to the different capabilities of sensors, the quality of captured videos can vary greatly, which affects the subsequent analysis and applications [1–4]. Recently, computer technologies have been applied to many fields [5–8]. In particular, video super-resolution (VSR) is a technology for improving the quality of captured video. It produces high-resolution (HR) video frames from their low-resolution (LR) counterparts. The VSR problem is challenging due to its ill-posed nature, but its applications include video display, video surveillance, video conferencing, and entertainment [9].

VSR models take consecutive frames as input. Single-image super-resolution (SISR) methods process only one image at a time. So, VSR models take both spatial information and temporal information into account, while SISR models only exploit spatial information for super-resolution (SR) reconstruction. Thus, many VSR methods adapt SISR models for spatial information extraction. For example, Haris et al. [10] introduced RBPN, which employs blocks from DBPN [11] in a recurrent encoder–decoder module to utilize spatial and temporal information. Tian et al. [12] adapted EDSR [13] as the main design for the SR reconstruction network in TDAN. Liang et al. [14] utilized residual Swin Transformer blocks from SwinIR [15] in their proposed RVRT. Although these works have adapted SISR models, each method utilizes only one SISR model. Applying SISR techniques to the

VSR models would require considerable effort and they may not perform as effectively as specialized VSR models.

Meanwhile, several VSR methods do not rely on SISR models. For instance, Xue et al. [16] proposed TOF, which estimates task-oriented flow to recover details in SR frames. Wang et al. [17] proposed SOF-VSR, which estimates HR optical flow from LR frames. SWRN [18] can be utilized in real time on a mobile device. However, the development of a VSR model without adapting SISR methods is very costly, as the model needs to capture both temporal and spatial information. Moreover, compared with SISR methods, they may be less effective in utilizing spatial information.

To alleviate the above issues, we propose a plug-and-play approach for adapting existing SISR models to the VSR task. Firstly, we summarize a common architecture of SISR models and provide a formal analysis of adaptation to achieve better effectiveness of different SISR models. Then, we present an adaptation method, which inserts a plug-and-play temporal feature extraction module into SISR models. Specifically, the temporal feature extraction module consists of three submodules. The spatial aggregation submodule aligns features extracted by the original SISR model. The alignment is performed based on the result of the offset estimation submodule. Then, the temporal aggregation submodule is applied to aggregate information extracted from all neighboring frames.

To evaluate the effectiveness of the proposed method, we adapt five representative SISR models, i.e., SRResNet [19], EDSR [13], RCAN [20], RDN [21], and SwinIR [15], and the evaluations are conducted on two popular benchmarks, i.e., Vid4 and SPMC-11. On the Vid4 benchmark, the VSR-adapted models achieve at least 1.26 dB and 0.067 improvements over original SISR models in terms of peak signal-to-noise ratio (PSNR) [22] and structural similarity index (SSIM) [23], respectively. On the SPMC benchmark, the VSR-adapted models achieve at least 1.16 dB and 0.036 gain over original SISR models in terms of PSNR and SSIM, respectively. Moreover, the VSR-adapted models surpassed the performance of state-of-the-art VSR models.

For this paper, the main contributions are as follows: (1) We propose a plug-and-play approach for adapting SISR models to the VSR task. Instead of adapting one SISR model, the proposed method is based on a common architecture of SISR models. (2) A plug-and-play temporal feature extraction module is introduced. Thus, the adapted model gains the capability to exploit temporal information. (3) Extensive experiments are conducted to evaluate its effectiveness.

2. Related Work

2.1. Single-Image Super-Resolution

The SISR problem is an ill-posed problem, and learning-based methods have significantly improved the performance in terms of accuracy [13,15,19–21,24,25] and speed [26–29]. In 2014, Dong et al. [30] introduced a learning-based model, namely SRCNN, into the SISR field. Inspired by ResNet [31], Ledig et al. [19] proposed SRResNet in 2017. SRResNet [19] accepts LR images directly and achieves high performance and increased efficiency. Kim et al. [13] improved the SRResNet by removing unnecessary batch normalization in residual blocks and expanding the number of parameters. In 2018, Zhang et al. [21] employed a densely connected architecture. All extracted features are fused to utilize hierarchical information. Subsequently, Zhang et al. [20] introduced the channel attention mechanism that adaptively weights features channel-wisely. In 2021, Liang et al. [15] proposed SwinIR by making use of the Transformer [32]. Additionally, SwinIR uses the Swin Transformer [33] variation, which is more appropriate for computer vision tasks. By appropriately employing convolution layers and Swin Transformer modules, SwinIR can capture local and global dependencies at the same time, resulting in SOTA performance.

2.2. Video Super-Resolution

In recent years, deep-learning-based models have been used to solve the VSR problem, and have become increasingly popular [9]. We roughly divide VSR models into two categories:

(1) Models adapting SISR models: Sajjadi et al. [34] proposed FRVSR, which takes EnhanceNet [35] as the subnetwork for SR reconstruction. Haris et al. [10] applied the iterative up- and downsampling technique [11] in RBPN. The representative deep learning SISR model, EDSR [13], is utilized by many VSR models. Tian et al. [12] applied a shallow version of EDSR [13] in TDAN. EDVR [36] and WAEN [37] both employed the residual block and upsampling module from EDSR [13] in the reconstruction module. Inspired by [12], Xu et al. [38] adapted EDSR as the reconstruction module. EGVSR [39] applied ESPCN [26] as the backbone for the SR net. The recently proposed RVRT [14] utilized the residual Swin Transformer block, which is proposed in SwinIR [15].

(2) Models without adapting SISR models: DUF [40] reconstructs SR frames by estimating upsampling filters and a residual image for high-frequency details. Kim et al. [41] employed 3D convolution to capture spatial–temporal nonlinear characteristics between LR and HR frames. Xue et al. [16] proposed a method, namely TOF. It learns a task-specific representation of motion. Wang et al. [17] proposed SOF-VSR, which estimates HR optical flow from LR frames. To better leverage the temporal information, TGA [42] introduced a hierarchical architecture. Recently, Chan et al. [43] proposed BasicVSR by investigating the essential components of VSR models. Liu et al. [44] applied spatial convolution packing to jointly exploit spatial–temporal features. For better fusing information from neighboring frames, Lee et al. [45] utilized both attention-based alignment and dilation-based alignment. Lian et al. [18] proposed SWRN to achieve real-time inference while producing superior performance.

Because VSR models have to capture both temporal and spatial information, proposing a VSR method requires more effort. Thus, many researchers turn to adapting SISR models. Based on SISR models, proposing a VSR method can focus on capturing temporal information. However, these models either utilize a SISR model as a subnet or adapt modules from a SISR model to extract features. Additionally, they may be less effective than those methods that do not adapt SISR methods. Our work proposed a plug-and-play approach to adapt SISR models to the VSR task. The proposed method works on different SISR models as it follows the common architecture of SISR models we have summarized. The spatial information and temporal information are both extracted in the proposed method.

3. Methodology

In this section, we first summarize the common architecture of SISR models. Then, we provide a formal analysis of adaptation. Following that, a general VSR adaptation method is proposed. Finally, we present a plug-and-play temporal feature extraction module.

3.1. Revisit of Single-Image Super-Resolution Models

For the effectiveness on different SISR models [13,15,19–21,46], we first summarize a common architecture, as shown in Figure 1. For simplicity, some operations such as element-wise addition and concatenation are omitted. As shown in Figure 1a, the common architecture of SISR models can be divided into three modules: shallow feature extraction (FE) module, deep FE module, and reconstruction module. Figure 1b–e illustrate the details of four SISR models. As one can see, the shallow FE module takes one LR image as input and extracts features by a few convolution layers. The deep FE module consists of several submodules or blocks, where advanced techniques, such as dense connection [21], channel attention [20], and self-attention [15], are applied. Thus, the deep FE module is where the key novelty of SISR models lies. Finally, the features from the deep FE module are fed to the reconstruction module to produce the SR image.

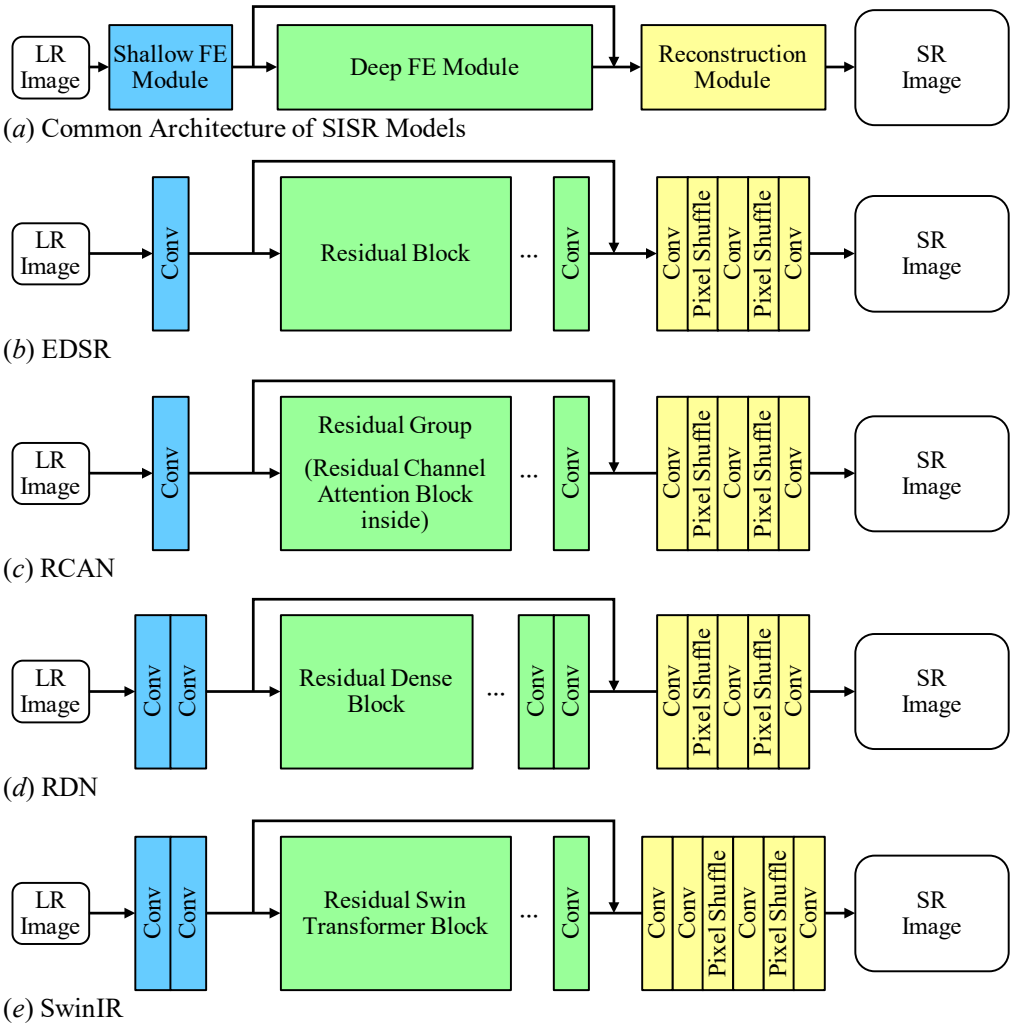

Figure 1. The architectures of typical SISR models.

Thus, given an LR image $\mathbf{y} \in \mathbb{R}^{H \times W \times 3}$, these SISR models can be generalized using the following representation:

$$\mathbf{x} = Method_{SISR}(\mathbf{y}), \tag{1}$$

where $Method_{SISR}(\cdot)$ is the SISR model. $\mathbf{x} \in \mathbb{R}^{sH \times sW \times 3}$ represents the SR result with upscale factor s. H and W denote the height and width of LR image, respectively. According to the common architecture of SISR models, Equation (1) can be expanded as

$$\mathbf{x} = Recons(FE_{deep}(FE_{shallow}(\mathbf{y})) + FE_{shallow}(\mathbf{y})), \tag{2}$$

where the shallow and deep FE modules are noted as $FE_{shallow}(\cdot)$ and $FE_{deep}(\cdot)$, respectively. The reconstruction module is denoted as $Recons(\cdot)$.

Different from the SISR problem, the VSR methods have to exploit both spatial and temporal information. Thus, we make use of sliding window framework [12] to capture temporal dependency. Given consecutive $2n+1$ LR frames $\mathbf{Y} = \{\mathbf{y}_{t-n}, \cdots \mathbf{y}_{t-1}, \mathbf{y}_t, \mathbf{y}_{t+1}, \cdots \mathbf{y}_{t+n}\}$, the representation of VSR models is formulated as

$$\mathbf{x}_t = Method_{VSR}(\mathbf{Y}), \tag{3}$$

where the VSR method is $Method_{VSR}(\cdot)$. $\mathbf{x}_t$ represents the reconstructed SR frame, the frame index of which is t.

Note that the main difference between Equations (1) and (3) is the input, and Equation (2) is an expanded representation of Equation (1). In order to adapt existing SISR models to

the VSR task, a straightforward method is to modify the shallow FE module. Then, the adapted model can be represented as

$$\mathbf{x}_t = Recons(FE_{deep}(FE'_{shallow}(\mathbf{Y})) + FE'_{shallow}(\mathbf{Y})), \tag{4}$$

where $FE'_{shallow}(\cdot)$ is the modified shallow FE module.

3.2. Proposed Video Super-Resolution Adaptation Method

According to the analysis in Section 3.1, we propose a general method to easily adapt SISR models to the VSR task. As shown in Figure 2, the architecture of the proposed VSR-adapted models consists of 4 modules. Firstly, the VSR-adapted model applies the shallow FE module $FE_{shallow}(\cdot)$ to obtain low-level features $\mathbf{F}_{s,i} \in \mathbb{R}^{H\times W\times C}$ for each LR frame $\mathbf{y}_i$. The subscript $_i$ represents the relative index of the center frame. The center frame is denoted as $_0$, and C stands for the number of channels in a feature. The shallow feature of center frame $\mathbf{F}_{s,0}$ is skip-connected to the output of the deep FE module with element-wise addition for global residual leaning. Secondly, the temporal FE module $FE_{temporal}(\cdot)$ is employed to exploit spatial–temporal information. It takes LR frames to estimate the offsets of pixels. It also takes shallow features which will be spatially aggregated based on the offsets. In order to enable the deep FE module to leverage information from all LR frames, spatial-aggregated features are temporally aggregated in the temporal FE module. Thirdly, the deep FE module $FE_{deep}(\cdot)$ is responsible for estimating accurate residual features with advanced techniques. Finally, the reconstruction module $Recons(\cdot)$ upsamples features with specific scale factors and produces SR frames. The architecture can be represented as

$$\mathbf{F}_{s,i} = FE_{shallow}(\mathbf{y}_i), \tag{5}$$

$$\mathbf{F}_T = FE_{temporal}(\mathbf{F}_{s,-n}, \cdots, \mathbf{F}_{s,0}, \cdots, \mathbf{F}_{s,n}, \mathbf{y}_{-n}, \cdots, \mathbf{y}_0, \cdots, \mathbf{y}_n), \tag{6}$$

$$\mathbf{x}_0 = Recons(FE_{deep}(\mathbf{F}_T) + \mathbf{F}_{s,0}), \tag{7}$$

where i denotes the relative index of the target frame, ranging from $-n$ to n. The temporal feature $\mathbf{F}_T \in \mathbb{R}^{H\times W\times C}$ is the output of temporal FE module.

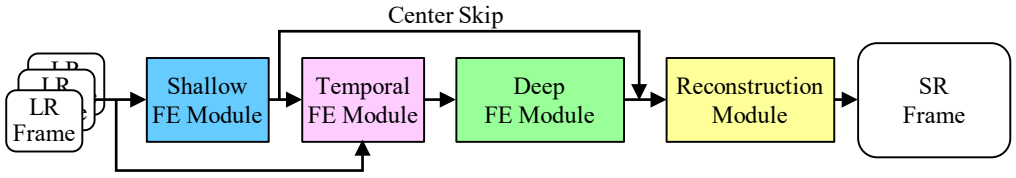

Figure 2. The Architecture of Proposed General VSR-Adapted Models.

For adapting different SISR models, the proposed method maintains the shallow FE module, deep FE module, and reconstruction module unmodified. Furthermore, we employ the temporal feature extraction module between the shallow FE module and the deep FE module in accordance with accuracy and latency concerns.

From an accuracy perspective, the main difference between an input LR frame and its ground truth HR frame is the high-frequency content. Thus, the better the residual feature that is extracted, the better the achieved performance. The proposed architecture takes advantage of the deep FE module, where the key novelties of SISR models lie [46]. Further, with the information from neighboring frames, the deep FE module is able to extract more accurate features for reconstruction. Thus, the temporal FE module is employed before deep FE module.

From a latency perspective, the temporal FE module aggregates the features extracted from all input frames. It requires previous modules to complete their processing for each frame. To minimize the overall computation time, the proposed temporal FE module is

employed after shallow FE module because its relatively small number of layers has a negligible impact on inference latency.

3.3. Plug-and-Play Temporal Feature Extraction Module

In order to exploit spatial–temporal information, the temporal FE module is proposed. The detailed architecture is illustrated in Figure 3, which consists of three submodules, i.e., offset estimation, spatial aggregation, and temporal aggregation.

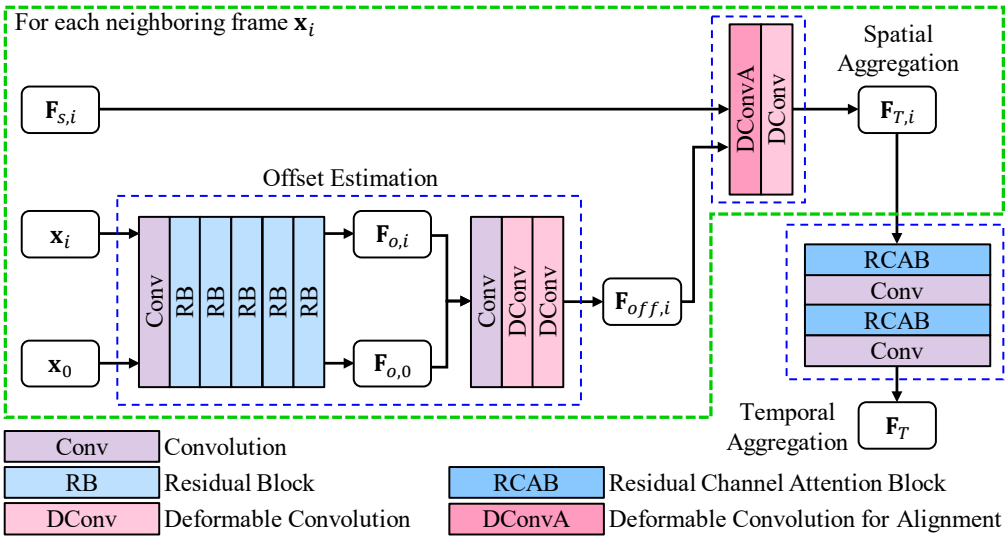

Figure 3. The Temporal Feature Extraction Module.

The offset estimation submodule takes the center LR frame $\mathbf{y}_0$ and each neighboring frame $\mathbf{y}_i$ as inputs. The intermediate feature extraction is performed by a convolution layer and five residual blocks, and the parameters are shared across all input LR frames. The intermediate features are noted as $\mathbf{F}_{o,i} \in \mathbb{R}^{H\times W\times C}$. The offset feature $\mathbf{F}_{off,i} \in \mathbb{R}^{H\times W\times C}$ is estimated from the intermediate feature $\mathbf{F}_{o,0}$ and $\mathbf{F}_{o,i}$ using a convolution layer and two deformable convolution layers. The offset estimation submodule can be formulated as

$$\mathbf{F}_{o,i} = RB_5(\cdots RB_1(Conv_1(\mathbf{y}_i))\cdots), \tag{8}$$

$$\mathbf{F}_{off,i} = DConv_2(DConv_1(Conv_2(CAT(\mathbf{F}_{o,i}, \mathbf{F}_{o,0})))), \tag{9}$$

where $RB(\cdot)$ is residual block. $Conv(\cdot)$ and $DConv(\cdot)$ are convolution and deformable convolution, respectively. The concatenation is denoted as $CAT(\cdot)$.

The shallow feature $\mathbf{F}_{s,i}$ and the estimated offset $\mathbf{F}_{off,i}$ are then fed into the spatial aggregation submodule. Here, a variation of deformable convolution is used to extract features $\mathbf{F}_{s,i}$, which takes $\mathbf{F}_{off,i}$ for offset. This allows the offset feature $\mathbf{F}_{off,i}$ to guide the alignment in the spatial aggregation submodule. Another deformable convolution is applied for refinement, resulting in output feature $\mathbf{F}_{T,i} \in \mathbb{R}^{H\times W\times C}$. The spatial aggregation submodule can be given by

$$\mathbf{F}_{T,i} = DConv_3(DConvA(\mathbf{F}_{s,i}, \mathbf{F}_{off,i})), \tag{10}$$

where $DConvA(\cdot,\cdot)$ is the variation of deformable convolution. The variation of deformable convolution $DConvA(\cdot,\cdot)$ takes the first input for feature extraction and the second input for offset.

After spatial aggregation, the temporal aggregation submodule fuses these spatial-aggregated features $\mathbf{F}_{T,-n}\cdots\mathbf{F}_{T,n}$. For fusing a feature with $(2n+1)\times C$ channels, a simple convolution layer is not sufficient. Therefore, a residual channel attention block [20]

is employed to adaptively weight these features channel-wise. A convolution layer for channel reduction is then applied. The channel shrinkage is performed in two steps to minimize information loss: first reducing to twice the SISR features' channels and then reducing to once. The temporal aggregation submodule can be represented as

$$\mathbf{F}_T = Conv_4(RCAB_2(Conv_3(RCAB_1(CAT(\mathbf{F}_{T,-n}, \cdots, \mathbf{F}_{T,n})))), \tag{11}$$

where $RCAB_1(\cdot)$ and $RCAB_2(\cdot)$ are residual channel attention blocks. The number of channels of the features output by $Conv_3(\cdot)$ and $Conv_4(\cdot)$ is $2 \times C$ and C, respectively. The temporal-aggregated feature is $\mathbf{F}_T \in \mathbb{R}^{H \times W \times C}$.

Overall, the spatial aggregation aligns neighboring features based on the result of the offset estimation submodule. Then, the temporal aggregation submodule fuses the spatial-aggregated features, resulting in an output containing information from all input LR frames. Finally, the plug-and-play module extracts feature $\mathbf{F}_T$, which contains spatial–temporal information from all input frames. Further, we summarize the detailed algorithm of the VSR-adapted method with plug-and-play temporal feature extraction module in Algorithm 1. For easy understanding, we divided the loop into multiple ones.

Algorithm 1: Video Super-Resolution with SISR Model and Plug-and-Play Temporal Feature Extraction Module.

Input : Consecutive low-resolution frames $\mathbf{y}_i$. i is relative index to the center frame ranging from $-n$ to n.
Output: Super-resolution center frame $\mathbf{x}_0$.
// Shallow FE module from SISR model
for $i = -n, -n+1, \cdots, n$ **do**
  $\mathbf{F}_{s,i} = FE_{shallow}(\mathbf{y}_i)$;
end
// Offset estimation submodule of temporal FE module
for $i = -n, -n+1, \cdots, n$ **do**
  $\mathbf{F}_{o,i} = RB_5(\cdots RB_1(Conv_1(\mathbf{y}_i))\cdots)$;
  $\mathbf{F}_{off,i} = DConv_2(DConv_1(Conv_2(CAT(\mathbf{F}_{o,i}, \mathbf{F}_{o,0}))))$;
end
// Spatial aggregation submodule of temporal FE module
for $i = -n, -n+1, \cdots, n$ **do**
  $\mathbf{F}_{T,i} = DConv_3(DConvA(\mathbf{F}_{s,i}, \mathbf{F}_{off,i}))$;
end
// Temporal aggregation submodule of temporal FE module
$\mathbf{F}_T = Conv_4(RCAB_2(Conv_3(RCAB_1(CAT(\mathbf{F}_{T,-n}, \mathbf{F}_{T,-n+1}, \cdots, \mathbf{F}_{T,n})))))$;
// Deep FE module and reconstruction module from SISR model
$\mathbf{x}_0 = Recons(FE_{deep}(\mathbf{F}_T) + \mathbf{F}_{s,0})$;

4. Experiment

4.1. Datasets

Following previous studies [12,16,47], we utilized the widely used Vimeo90K dataset for training. This dataset includes videos with different scenarios, such as moving objects, camera motion, and complex scene structures. It consists of 90,000 video clips with a resolution of 448 × 256. As per the official split, we use 64,612 video clips for training. The HR frames of these videos were used as the ground truth. For training, we randomly cropped these HR frames to patches with the size of 256 × 256, and these patches were bicubically downsampled to the size of 64 × 64 using the Matlab function *imresize*. We randomly flipped and rotated the data during training.

For testing, we evaluated the effectiveness of our proposed model on two public benchmarks, i.e., the Vid4 [48] and SPMC-11 [47]. The quantitative metrics were PSNR [22]

and SSIM [23], computed in the luminance (Y) channel. We also cropped 8 pixels near the image boundary, similar to the previous approach [12].

4.2. Implementation Details

To evaluate the proposed method, we employed it on five representative SISR models: (1) SRResNet [19] is the generator model in SRGAN. (2) EDSR [13] is a representative SISR model. (3) RCAN [20] makes use of channel attention. (4) RDN [21] has the advantage of a dense connection. (5) SwinIR [15] introduces Swin Transformer [33]. For SISR models, we generated SR videos frame by frame.

In our implementation of SRResNet [19], we removed all batch norm layers. We used the EDSR baseline [13] with a feature channel count and block count of 64 and 16, respectively. For SwnIR [15], the LR patch size was 48×48, and the GT patch size was 192×192. We used a smaller patch size for SwinIR for lower memory consumption. The batch size for training all models was 16. We empirically set $n = 2$, indicating that a VSR-adapted model takes five frames as input. For SISR models, the number of input frames was one. Each SISR model and its VSR-adapted model were trained from scratch using the same setting except for the number of input frames.

We used the mean square error (MSE) as the loss function, defined as $Loss = \|HR - SR\|^2$. The parameters were updated using the Adam optimizer [49] with $\beta 1 = 0.9$ and $\beta 2 = 0.99$. The learning rate was initialized as 1×10^{-4} and halved for every 1×10^5 iterations. We trained the models for 3×10^5 iterations. All experiments were implemented in Pytorch and ran on a server with NVIDIA GPUs.

4.3. Effectiveness on Different Single-Image Super-Resolution Models

To evaluate the effectiveness of the proposed method, we conducted experiments on five representative SISR models. Table 1 displays the quantitative results on two popular benchmarks. The PSNR and SSIM metrics of VSR-adapted models improved by at least 1.16 dB and 0.036, respectively. It demonstrates that the proposed method works effectively on various SISR models. Moreover, the performance of the VSR-adapted models is positively correlated with the capacity of the original models. In the SISR task, EDSR [13] is better than SRResNet [19] but underperforms RCAN [20] and RDN [21]. The performance of RCAN and RDN is on par, and SwinIR [15] has the best performance. As shown in Table 1, the VSR-adapted models exhibit similar trends. We use the suffix "-VSR" to represent the VSR-adapted models. The performances of SRResNet-VSR and EDSR-VSR are weaker than those of RCAN-VSR and RDN-VSR, and SwinIR-VSR achieves the best results on both benchmarks. Moreover, we computed the PSNR metric on the Vid4 benchmark during training. As illustrated in Figure 4, the VSR-adapted models benefit from the information aggregated from neighboring frames, and they performed better in the early iterations during training. Thus, the proposed method is effective on different SISR models, and the plug-and-play temporal feature extraction module enables the VSR-adapted models to exploit spatial and temporal information.

Further, we visualized the results of the Vid4 and SPMC-11 benchmarks for qualitative comparison. Several processed frames are shown in Figures 5 and 6. We can observe that the VSR-adapted models provide visually appealing results. By contrast, the original SISR models produce blurry SR frames and incorrect textures. Overall, the VSR-adapted models reconstruct results with clearer text, richer textures, and fewer artifacts. Among the results of the VSR-adapted models, SRResNet-VSR and EDSR-VSR produce more artifacts than other VSR-adapted models. This is consistent with the capabilities of original SISR models.

Table 1. Quantitative Comparison of SISR Models and VSR-Adapted Models on Vid4 and SPMC-11. The best results are in bold.

Benchmark	Method	Original		VSR Adapted	
		PSNR	SSIM	PSNR	SSIM
Vid4	SRResNet [19]	25.30	0.728	26.56	0.797
	EDSR [13]	25.27	0.726	26.58	0.798
	RCAN [20]	**25.45**	0.737	26.74	0.804
	RDN [21]	25.40	0.734	26.75	0.806
	SwinIR [15]	25.41	**0.738**	**26.84**	**0.811**
SPMC-11	SRResNet [19]	27.92	0.815	29.16	0.853
	EDSR [13]	27.85	0.813	29.14	0.853
	RCAN [20]	28.32	0.823	29.48	0.859
	RDN [21]	28.24	0.821	29.55	0.862
	SwinIR [15]	**28.46**	**0.826**	**29.74**	**0.866**

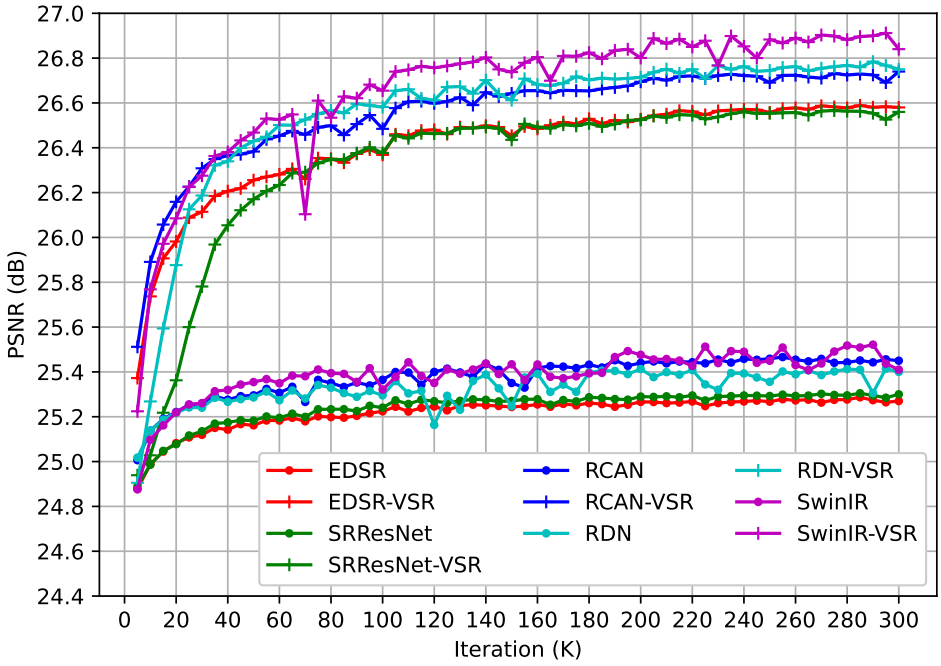

Figure 4. The PSNR Curve on Vid4 Benchmark During Training.

4.4. Comparisons with State-of-the-Art Methods

We compared these VSR-adapted models with 10 state-of-the-art VSR algorithms, i.e., STAN [50], EGVSR [39], TOFlow [16], STMN [51], SOF-VSR [17], ST-CNN [44], TDAN [12], D3Dnet [47], FRVSR [34], and WAEN [37]. Table 2 shows the quantitative metrics on the Vid4 and SPMC-11 benchmarks. The values with † are reported in [47]. As shown in Table 2, the VSR-adapted models achieve competitive performance on both Vid4 and SPMC-11 benchmarks. All VSR-adapted models perform better than D3Dnet. Compared with D3Dnet, the SRResNet-VSR and EDSR-VSR achieve comparative performance. The performances achieved by RCAN-VSR and RDN-VSR are between FRVSR and WAEN. Among them, the SwinIR-VSR outperforms all models in terms of PSNR metrics.

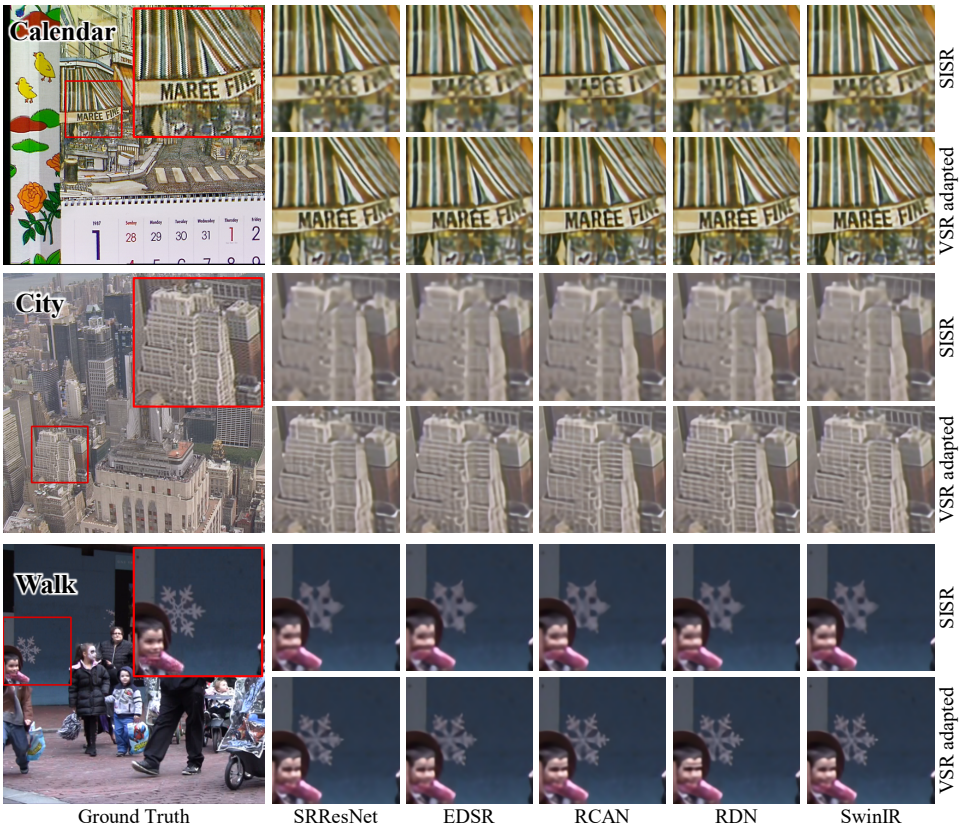

Figure 5. The Qualitative Comparison of SISR Models and Corresponding VSR Adaptations on Vid4 Benchmark.

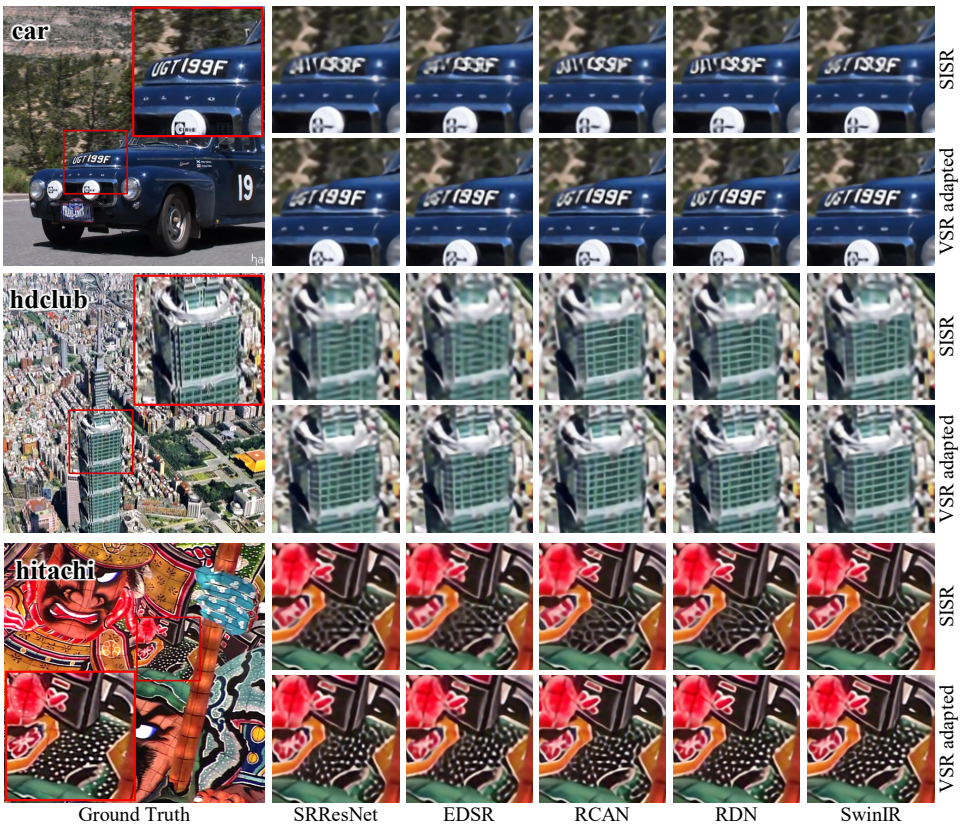

Figure 6. The Qualitative Comparison of SISR Models and Corresponding VSR Adaptations on SPMC-11 Benchmark.

Table 2. Quantitative comparison of Vid4 and SPMC-11. The best results are in bold. The values with † are reported in [47].

Method	Vid4		SPMC-11	
	PSNR (dB)	SSIM	PSNR (dB)	SSIM
STAN [50]	25.58	0.743	—	—
EGVSR [39]	25.88	0.800	—	—
TOFlow [16]	25.90	0.765	—	—
STMN [51]	25.90	0.788	—	—
SOF-VSR [17]	26.02	0.772	28.21 †	0.832 †
ST-CNN [44]	26.12	**0.823**	—	—
TDAN [12]	26.42	0.789	28.51 †	0.841 †
D3Dnet [47]	26.52	0.799	28.78	0.851
FRVSR [34]	26.69	0.822	—	—
WAEN [37]	26.79	—	—	—
SRResNet-VSR	26.56	0.797	29.16	0.853
EDSR-VSR	26.58	0.798	29.14	0.853
RCAN-VSR	26.74	0.804	29.48	0.859
RDN-VSR	26.75	0.806	29.55	0.862
SwinIR-VSR	**26.84**	0.811	**29.74**	**0.866**

For a finer quantitative comparison on the Vid4 benchmark, we illustrate the PSNR metric of each frame in Figure 7. For simplicity, we select four models, i.e., TDAN [12], FRVSR [34], EDSR-VSR, and SwinIR-VSR. Compared with TDAN, the EDSR-VSR achieves similar performance. Note that the first two and last two frames show a greater difference between TDAN and EDSR-VSR. Because there is less neighboring information for VSR models to exploit, the VSR models exhibit poor performance at the beginning and end of a video. Compared with FRVSR, the SwinIR-VSR achieved better performance on the *Calendar* and *Walk*. As the frame index increases on the *Calendar*, the gap between SwinIR-VSR and FRVSR becomes smaller. Additionally, the performance of SwinIR-VSR is lower than that of FRVSR after the first five frames on the *City*. This is because the SwinIR-VSR makes use of neighboring frames in a sliding window scheme while the FRVSR utilizes them in a recurrent scheme.

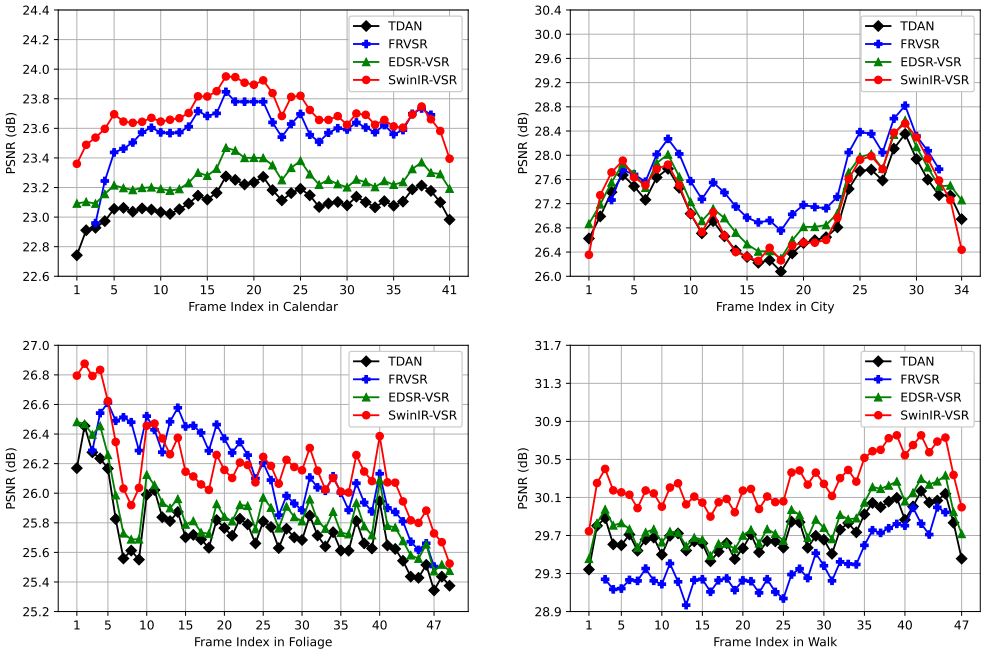

Figure 7. The PSNR curve of VSR models on Vid4 benchmark.

For a qualitative comparison, we compared the VSR-adapted models to SOF-VSR [17], TOF [16], TDAN [12], D3Dnet [47], and FRVSR [34]. As shown in Figure 8, the VSR-adapted models reconstruct visually attractive results. The text on the *Calendar* is now easier to read and the details of the *City* are clearer. Additionally, the clothes in the *Walk* image are more recognizable. Moreover, we observed similar trends in the SPMC-11 benchmark, as illustrated in Figure 9. The quality of the reconstructed results of EDSR-VSR is equivalent to that of the compared methods. The RDN-VSR and RCAN-VSR provide results with better quality. The result of SwinIR-VSR has the least artifacts.

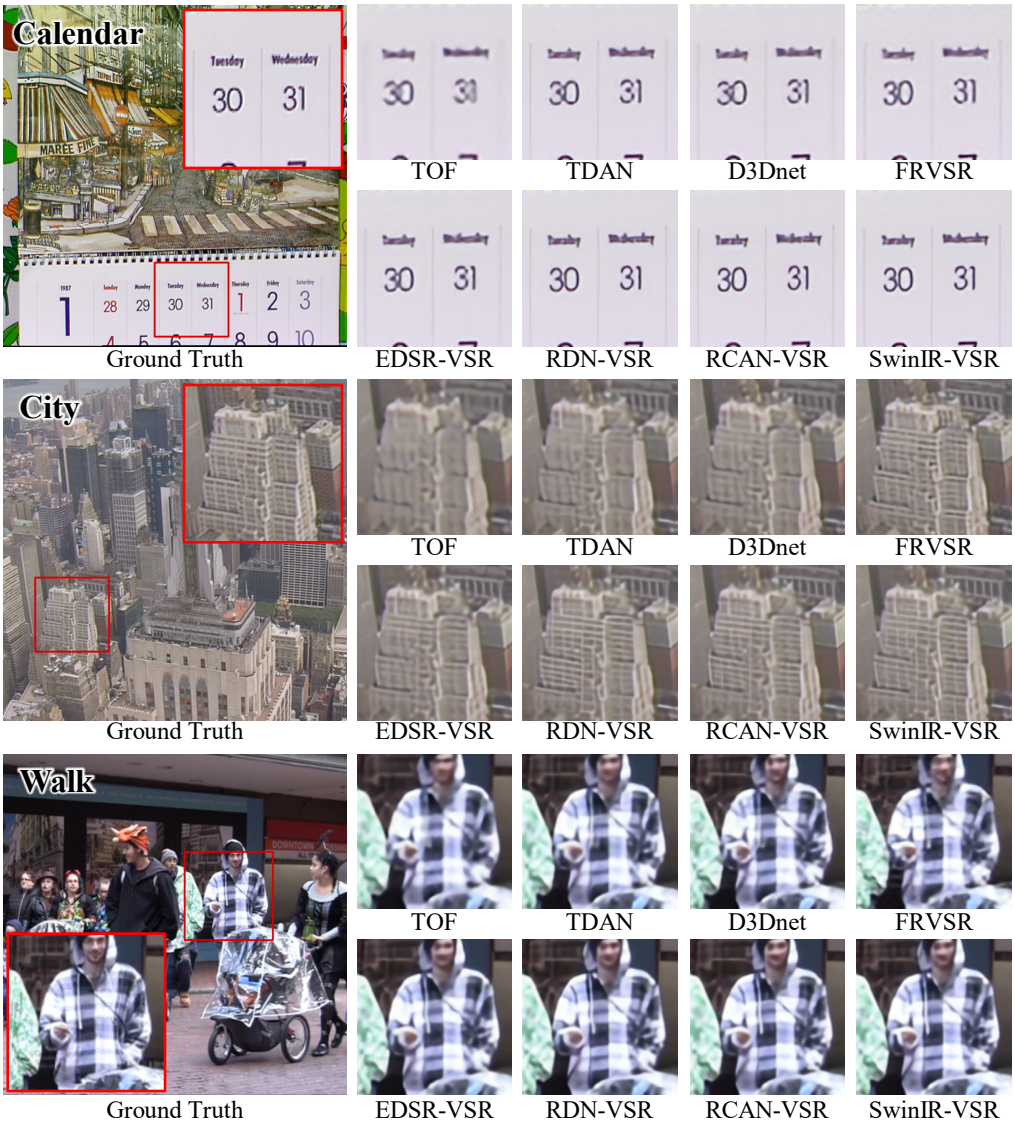

Figure 8. Qualitative Comparison of VSR Models on Vid4 Benchmark.

4.5. Comparisons of Temporal Consistency

To evaluate the temporal consistency of the proposed method, we generated temporal profiles according to [34] for visualization. As shown in Figure 10, the positions of temporal profiles are highlighted with red lines. The heights of temporal profiles vary due to the video length. As shown in the *Calendar*, the temporal profiles demonstrate that the original SISR models perform poorly because they are unable to capture temporal information. By contrast, the VSR methods and VSR-adapted models produce results with fewer artifacts. However, inappropriate aggregation of temporal information can lead to degraded results. As illustrated in the *City*, the original SISR models and our VSR-adapted models exhibit better temporal consistency than VSR models.

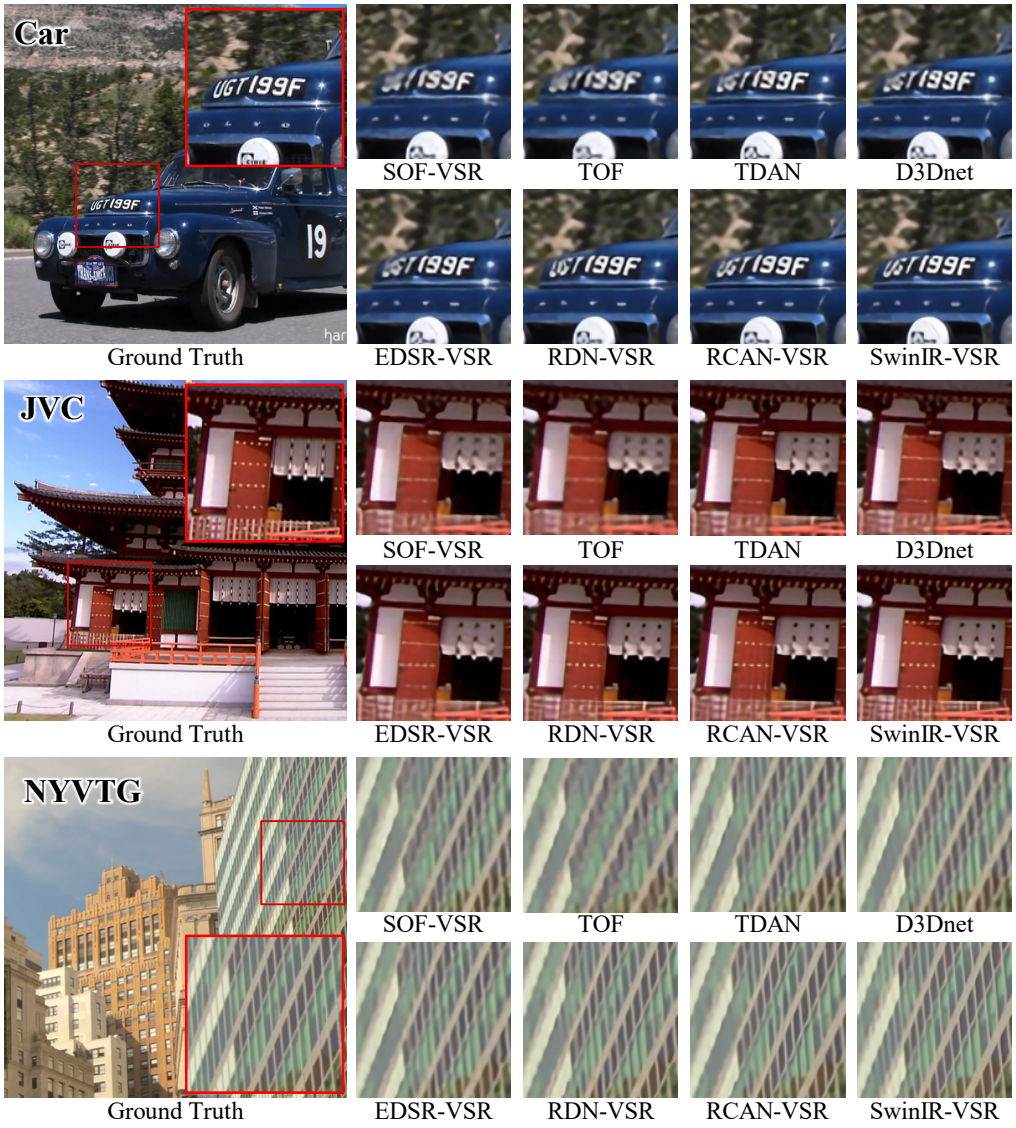

Figure 9. Qualitative Comparison of VSR Models on SPMC-11 Benchmark.

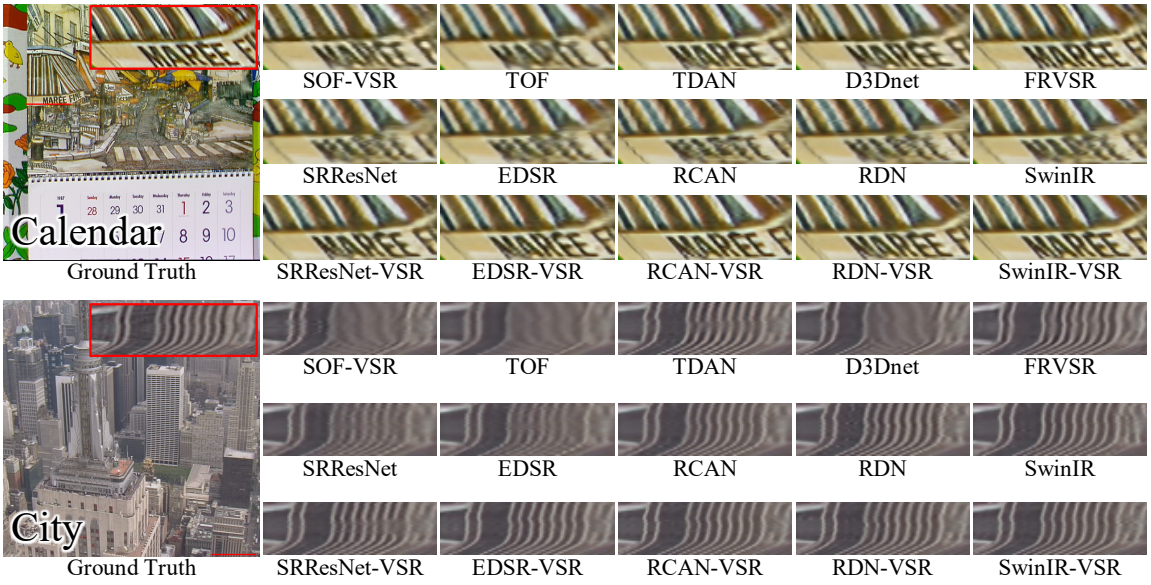

Figure 10. Qualitative Comparison of Temporal Profile on Vid4 Benchmark.

4.6. Ablation Study

We used EDSR [13] as the baseline in the ablation study to evaluate the effectiveness of the proposed temporal feature extraction module, which consists of offset estimation, spatial aggregation, and temporal aggregation submodules. We evaluated three models to determine the effectiveness of each submodule. The first variation is denoted as Model 1. We fed shallow features from neighboring frames to the spatial aggregation submodule without the support of the offset estimation submodule. The neighboring features were then fused with a convolution using a 1×1 kernel. Model 2 is referred to as the second variation. We introduced the offset estimation submodule, which makes use of the center frame and neighboring frames to guide the spatial aggregation. The third variation, denoted as EDSR-VSR, combines all the components, including channel attention and progressive channel shrinking.

Table 3 indicates that relying solely on the spatial aggregation submodule does not lead to performance improvement. However, with the support of the offset estimation submodule, there is a significant performance improvement. Furthermore, the temporal aggregation submodule further improved the performance. Three submodules play an irreplaceable role in our presented temporal feature extraction module.

Table 3. The Effectiveness of Each Component in Temporal Feature Extraction Module.

Dataset	Model	Spatial Aggregation	Offset Estimation	Temporal Aggregation	PSNR (dB)	SSIM
Vid4	EDSR [13]	✗	✗	✗	25.27	0.726
	Model 1	✓	✗	✗	25.31	0.725
	Model 2	✓	✓	✗	26.49	0.793
	EDSR-VSR	✓	✓	✓	26.58	0.798
SPMC-11	EDSR [13]	✗	✗	✗	27.85	0.813
	Model 1	✓	✗	✗	27.88	0.813
	Model 2	✓	✓	✗	28.97	0.849
	EDSR-VSR	✓	✓	✓	29.14	0.853

To evaluate the efficiency of the proposed method, we conducted a comparison on the Vid4 benchmark. We evaluated three models, i.e., EDSR [13], EDSR-VSR, and EDSR-VSR 2. The EDSR-VSR 2 employs the temporal feature extraction module after the deep feature extraction module. Table 4 shows the performance and average latency of inference. As we can see, the EDSR-VSR is about $1.6\times$ faster than the EDSR-VSR 2. Although the EDSR-VSR is slower than EDSR [13], it reaches 24 frames per second. Specifically, we analyzed the latency of each part of EDSR-VSR. Overall, 0.89% of the latency is consumed by the shallow feature extraction module from the SISR model. The subsequent offset estimation submodule, spatial aggregation submodule, and temporal aggregation submodule occupied 21.25%, 39.99%, and 15.21% of the latency, respectively. Additionally, 22.66% of the time is spent on the deep feature extraction and reconstruction module from the SISR model. Note that the temporal feature extraction module has to process all input frames, so each submodule takes a longer time to complete the computation. Thus, the proposed method balances the accuracy and latency.

Table 4. The Efficiency of Proposed Method on Vid4 Benchmark.

	EDSR [13]	EDSR-VSR	EDSR-VSR 2
PSNR (dB)	25.27	26.58	26.61
SSIM	0.726	0.798	0.798
Latency (ms)	9.872	41.543	65.003

5. Discussion and Limitation

The proposed method builds a bridge between the SISR model and the VSR model. We revisited many SISR models and summarized a common architecture of SISR models. The proposed method leverages the inherent similarities and differences between the two tasks, and the plug-and-play temporal feature extraction module is presented to allow the VSR-adapted model to utilize information from neighboring frames. We applied it to five representative SISR models to evaluate our method, including a generator of GAN [19], three representative SISR models [13,20,21], and a Transformer-based model [15]. Compared with state-of-the-art VSR models, our VSR-adapted models achieve competitive performance.

There are several strong points of the proposed method. Firstly, the proposed architecture of VSR-adapted models provides a novel scheme to develop VSR models. As long as a SISR model follows the common architecture, it can be easily adapted to a VSR model. It reduces the delay of applications of new SISR technologies. Secondly, with the development of VSR, better temporal feature extraction techniques will be proposed, leading to better VSR performance. It divides the development of the VSR model into two independent tasks. Thirdly, the plug-and-play characteristic enables a single model to perform both SISR and VSR tasks.

Although the VSR-adapted models show promising results, we observed some failure cases in experiments. As illustrated in Figure 11, these models fail to recover tiny details. In these cases, the contrast is low in the ground truth, and the contrast is further reduced in LR frames, making SR reconstruction very challenging. Furthermore, all VSR-adapted models fail to provide clear results.

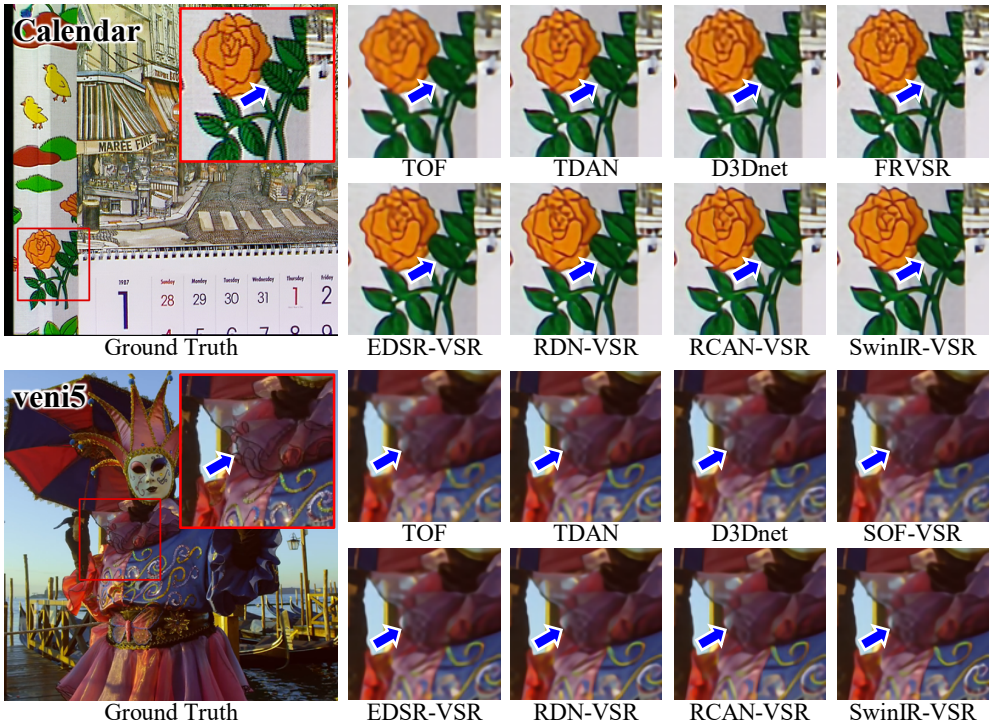

Figure 11. The Qualitative Comparison of Details in Low-Contrast Areas.

6. Conclusions

In this paper, we propose a method for adapting SISR models to the VSR task. For effectiveness on various SISR models, we summarize the common architecture of SISR models. The VSR-adapted models leverage the capability of SISR models to learn the mapping between LR and HR images. Then, the proposed plug-and-play temporal feature extraction module allows VSR-adapted models to access spatial–temporal information.

Thus, the performance in the VSR task is improved by the incorporation of the SISR model and the temporal feature extraction module. The experiments on several SISR models and benchmarks show that VSR-adapted models surpass the original SISR models. The achieved performance is positively related to the capacity of SISR models, indicating the effectiveness of the proposed method. Further, the VSR-adapted models achieved better results than the SOTA VSR models. In the future, we plan to solve the problem of poor performance in low-contrast areas.

Author Contributions: Conceptualization, W.W.; methodology, W.W.; software, W.W. and Y.H.; validation, W.W. and H.L.; formal analysis, Z.L. and R.L.; investigation, W.W., H.L. and Y.H.; resources, Z.L. and R.L.; data curation, W.W., H.L. and Y.H.; writing—original draft preparation, W.W. and H.L.; writing—review and editing, Z.L., R.L., H.L., Y.H. and W.W.; visualization, H.L., Y.H. and W.W.; supervision, Z.L.; project administration, W.W.; funding acquisition, Z.L. and R.L. 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 (61866009, 62172120, 82272075), Guangxi Science Fund for Distinguished Young Scholars (2019GXNSFFA245014), Guangxi Key Research and Development Program (AB21220037), and Innovation Project of Guangxi Graduate Education (YCBZ2022112).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The public data used in this work are listed here: Vimeo90k http://toflow.csail.mit.edu/ (accessed on 12 December 2022), Vid4 https://drive.google.com/file/d/1ZuvNNLgR85TV_whJoH-M7uVb-XW1y70DW/view?usp=sharing (accessed on 12 December 2022), and SPMC-11 https://pan.baidu.com/s/1PK-ZeTo8HVklHU5Pe26qUtw (accessed on 12 December 2022) (Code: 4l5r).

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

References

1. Yang, C.; Huang, Z.; Wang, N. QueryDet: Cascaded Sparse Query for Accelerating High-Resolution Small Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, 18–24 June 2022; pp. 13658–13667. [CrossRef]
2. Shermeyer, J.; Etten, A.V. The Effects of Super-Resolution on Object Detection Performance in Satellite Imagery. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Computer Vision Foundation/IEEE, Long Beach, CA, USA, 16–20 June 2019; pp. 1432–1441. [CrossRef]
3. Dong, H.; Xie, K.; Xie, A.; Wen, C.; He, J.; Zhang, W.; Yi, D.; Yang, S. Detection of Occluded Small Commodities Based on Feature Enhancement under Super-Resolution. *Sensors* **2023**, *23*, 2439. [CrossRef] [PubMed]
4. Yuan, X.; Fu, D.; Han, S. LRF-SRNet: Large-Scale Super-Resolution Network for Estimating Aircraft Pose on the Airport Surface. *Sensors* **2023**, *23*, 1248. [CrossRef] [PubMed]
5. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. [CrossRef]
6. Cheng, H.K.; Schwing, A.G. XMem: Long-Term Video Object Segmentation with an Atkinson-Shiffrin Memory Model. In Proceedings of the Computer Vision-ECCV 2022—17th European Conference, Tel Aviv, Israel, 23–27 October 2022; Proceedings, Part XXVIII; Lecture Notes in Computer Science; Avidan, S., Brostow, G.J., Cissé, M., Farinella, G.M., Hassner, T., Eds.; Springer: Berlin/Heidelberg, Germany, 2022; Volume 13688, pp. 640–658. [CrossRef]
7. Chen, Y.; Xia, R.; Zou, K.; Yang, K. FFTI: Image inpainting algorithm via features fusion and two-steps inpainting. *J. Vis. Commun. Image Represent.* **2023**, *91*, 103776. [CrossRef]
8. Imran, A.; Sulaman, M.; Yang, S.; Bukhtiar, A.; Qasim, M.; Elshahat, S.; Khan, M.S.A.; Dastgeer, G.; Zou, B.; Yousaf, M. Molecular beam epitaxy growth of high mobility InN film for high-performance broadband heterointerface photodetectors. *Surf. Interfaces* **2022**, *29*, 101772. [CrossRef]
9. Liu, H.; Ruan, Z.; Zhao, P.; Dong, C.; Shang, F.; Liu, Y.; Yang, L.; Timofte, R. Video super-resolution based on deep learning: A comprehensive survey. *Artif. Intell. Rev.* **2022**, *55*, 5981–6035. [CrossRef]
10. Haris, M.; Shakhnarovich, G.; Ukita, N. Recurrent Back-Projection Network for Video Super-Resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Computer Vision Foundation/IEEE, Long Beach, CA, USA, 16–20 June 2019; pp. 3897–3906. [CrossRef]

11. Haris, M.; Shakhnarovich, G.; Ukita, N. Deep Back-Projection Networks for Super-Resolution. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Computer Vision Foundation/IEEE Computer Society, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1664–1673. [CrossRef]
12. Tian, Y.; Zhang, Y.; Fu, Y.; 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 2020, Computer Vision Foundation/IEEE, Seattle, WA, USA, 13–19 June 2020; pp. 3357–3366. [CrossRef]
13. 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, CVPR Workshops 2017, IEEE Computer Society, Honolulu, HI, USA, 21–26 July 2017; pp. 1132–1140. [CrossRef]
14. Liang, J.; Fan, Y.; Xiang, X.; Ranjan, R.; Ilg, E.; Green, S.; Cao, J.; Zhang, K.; Timofte, R.; Gool, L.V. Recurrent Video Restoration Transformer with Guided Deformable Attention. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 378–393.
15. Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Gool, L.V.; Timofte, R. SwinIR: Image Restoration Using Swin Transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021, Montreal , BC, Canada, 11–17 October 2021; pp. 1833–1844. [CrossRef]
16. 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]
17. 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]
18. Lian, W.; Lian, W. Sliding Window Recurrent Network for Efficient Video Super-Resolution. In Proceedings of the Computer Vision-ECCV 2022 Workshops, Tel Aviv, Israel , 23–27 October 2022; Proceedings, Part II; Lecture Notes in Computer Science; Karlinsky, L., Michaeli, T., Nishino, K., Eds.; Springer: Berlin/Heidelberg, Germany, 2022; Volume 13802, pp. 591–601. [CrossRef]
19. Ledig, C.; Theis, L.; Huszar, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.P.; Tejani, A.; Totz, J.; Wang, Z.; et al. 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 2017, IEEE Computer Society, Honolulu, HI, USA, 21–26 July 2017; pp. 105–114. [CrossRef]
20. Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; Fu, Y. Image Super-Resolution Using Very Deep Residual Channel Attention Networks. In Proceedings of the Computer Vision-ECCV 2018—15th European Conference, Munich, Germany, 8–14 September 2018; Proceedings, Part VII; Lecture Notes in Computer Science; Ferrari, V.; Hebert, M.; Sminchisescu, C.; Weiss, Y., Eds. Springer: Berlin/Heidelberg, Germany, 2018; Volume 11211, pp. 294–310. [CrossRef]
21. Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; Fu, Y. Residual Dense Network for Image Super-Resolution. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Computer Vision Foundation/IEEE Computer Society, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2472–2481. [CrossRef]
22. Horé, A.; Ziou, D. Image Quality Metrics: PSNR vs. SSIM. In Proceedings of the 2010 20th International Conference on Pattern Recognition, Istanbul, Turkey, 23–26 August 2010; pp. 2366–2369. [CrossRef]
23. Wang, Z.; Bovik, A.; Sheikh, H.; Simoncelli, E. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef]
24. Liu, Y.; Chu, Z.; Li, B. A Local and Non-Local Features Based Feedback Network on Super-Resolution. *Sensors* **2022**, *22*, 9604. [CrossRef]
25. Chen, Y.; Xia, R.; Yang, K.; Zou, K. MFFN: Image super-resolution via multi-level features fusion network. *Vis. Comput.* **2023**, 1–16. [CrossRef]
26. Shi, W.; Caballero, J.; Huszar, 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 2016, IEEE Computer Society, Las Vegas, NV, USA, 27–30 June 2016; pp. 1874–1883. [CrossRef]
27. Lan, R.; Sun, L.; Liu, Z.; Lu, H.; Pang, C.; Luo, X. MADNet: A Fast and Lightweight Network for Single-Image Super Resolution. *IEEE Trans. Cybern.* **2021**, *51*, 1443–1453. [CrossRef] [PubMed]
28. Lan, R.; Sun, L.; Liu, Z.; Lu, H.; Su, Z.; Pang, C.; Luo, X. Cascading and Enhanced Residual Networks for Accurate Single-Image Super-Resolution. *IEEE Trans. Cybern.* **2021**, *51*, 115–125. [CrossRef] [PubMed]
29. Sun, L.; Liu, Z.; Sun, X.; Liu, L.; Lan, R.; Luo, X. Lightweight Image Super-Resolution via Weighted Multi-Scale Residual Network. *IEEE/CAA J. Autom. Sin.* **2021**, *8*, 1271–1280. [CrossRef]
30. 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]
31. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, IEEE Computer Society, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
32. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.

33. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, IEEE, Montreal, QC, Canada, 10–17 October 2021; pp. 9992–10002. [CrossRef]
34. Sajjadi, M.S.M.; Vemulapalli, R.; Brown, M. Frame-Recurrent Video Super-Resolution. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Computer Vision Foundation/IEEE Computer Society, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6626–6634. [CrossRef]
35. Sajjadi, M.S.M.; Schölkopf, B.; Hirsch, M. EnhanceNet: Single Image Super-Resolution Through Automated Texture Synthesis. In Proceedings of the IEEE International Conference on Computer Vision, ICCV 2017, IEEE Computer Society, Venice, Italy, 22–29 October 2017; pp. 4501–4510. [CrossRef]
36. 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 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Computer Vision Foundation/IEEE, Long Beach, CA, USA, 16–20 June 2019; pp. 1954–1963. [CrossRef]
37. Choi, Y.J.; Lee, Y.; Kim, B. Wavelet Attention Embedding Networks for Video Super-Resolution. In Proceedings of the 25th International Conference on Pattern Recognition, ICPR 2020, Milan, Italy, 10–15 January 2021; pp. 7314–7320. [CrossRef]
38. Xu, W.; Song, H.; Jin, Y.; Yan, F. Video Super-Resolution with Frame-Wise Dynamic Fusion and Self-Calibrated Deformable Alignment. *Neural Process. Lett.* **2022**, *54*, 2803–2815. [CrossRef]
39. Cao, Y.; Wang, C.; Song, C.; Tang, Y.; Li, H. Real-Time Super-Resolution System of 4K-Video Based on Deep Learning. In Proceedings of the 32nd IEEE International Conference on Application-specific Systems, Architectures and Processors, ASAP 2021, Virtual, 7–9 July 2021; pp. 69–76. [CrossRef]
40. 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 Conference on Computer Vision and Pattern Recognition, CVPR 2018, Computer Vision Foundation/IEEE Computer Society, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3224–3232. [CrossRef]
41. Kim, S.Y.; Lim, J.; Na, T.; Kim, M. Video Super-Resolution Based on 3D-CNNS with Consideration of Scene Change. In Proceedings of the 2019 IEEE International Conference on Image Processing, ICIP 2019, Taipei, Taiwan, 22–25 September 2019; pp. 2831–2835. [CrossRef]
42. 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 2020, Computer Vision Foundation/IEEE, Seattle, WA, USA, 13–19 June 2020; pp. 8005–8014. [CrossRef]
43. 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 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Computer Vision Foundation/IEEE, Virtual, 19–25 June 2021; pp. 4947–4956. [CrossRef]
44. Liu, Z.; Siu, W.; Chan, Y. Efficient Video Super-Resolution via Hierarchical Temporal Residual Networks. *IEEE Access* **2021**, *9*, 106049–106064. [CrossRef]
45. Lee, Y.; Cho, S.; Jun, D. Video Super-Resolution Method Using Deformable Convolution-Based Alignment Network. *Sensors* **2022**, *22*, 8476. [CrossRef]
46. Anwar, S.; Khan, S.H.; Barnes, N. A Deep Journey into Super-resolution: A Survey. *ACM Comput. Surv.* **2021**, *53*, 60:1–60:34. [CrossRef]
47. 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]
48. Liu, C.; Sun, D. On Bayesian Adaptive Video Super Resolution. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 346–360. [CrossRef]
49. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015*; Conference Track Proceedings; Bengio, Y., LeCun, Y., Eds.; Springer: Berlin/Heidelberg, Germany, 2015.
50. Wen, W.; Ren, W.; Shi, Y.; Nie, Y.; Zhang, J.; Cao, X. Video Super-Resolution via a Spatio-Temporal Alignment Network. *IEEE Trans. Image Process.* **2022**, *31*, 1761–1773. [CrossRef] [PubMed]
51. Zhu, X.; Li, Z.; Lou, J.; Shen, Q. Video super-resolution based on a spatio-temporal matching network. *Pattern Recognit.* **2021**, *110*, 107619. [CrossRef]