applied sciences

MDPI

Article

LightVSR: A Lightweight Video Super-Resolution Model with Multi-Scale Feature Aggregation

Guanglun Huang [1,2], Nachuan Li [2], Jianming Liu [2], Minghe Zhang [2], Li Zhang [2,*] and Jun Li [2,*]

[1] Post-Doctoral Research Station in Instrumentation Science and Technology, Guilin University of Electronic Technology, Guilin 541004, China; hglun@guet.edu.cn
[2] School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin 541004, China; a1425699467@gmail.com (N.L.); zhangminghe12332@163.com (M.Z.)
* Correspondence: zhangli@guet.edu.cn (L.Z.); lijun5221@126.com (J.L.)

Abstract: Video super-resolution aims to generate high-resolution video sequences with realistic details from existing low-resolution video sequences. However, most existing video super-resolution models require substantial computational power and are not suitable for resource-constrained devices such as smartphones and tablets. In this paper, we propose a lightweight video super-resolution (LightVSR) model that employs a novel feature aggregation module to enhance video quality by efficiently reconstructing high-resolution frames from compressed low-resolution inputs. LightVSR integrates several novel mechanisms, including head-tail convolution, cross-layer shortcut connections, and multi-input attention, to enhance computational efficiency while guaranteeing video super-resolution performance. Extensive experiments show that LightVSR achieves a frame rate of 28.57 FPS and a PSNR of 39.25 dB on the UDM10 dataset and 36.91 dB on the Vimeo-90k dataset, validating its efficiency and effectiveness.

Keywords: video super-resolution; lightweight neural networks; feature aggregation module

check for updates

Received: 12 January 2025
Revised: 28 January 2025
Accepted: 29 January 2025
Published: 1 February 2025

Citation: Huang, G.; Li, N.; Liu, J.; Zhang, M.; Zhang, L.; Li, J. LightVSR: A Lightweight Video Super-Resolution Model with Multi-Scale Feature Aggregation. *Appl. Sci.* **2025**, *15*, 1506. https://doi.org/10.3390/ app15031506

Copyright: © 2025 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

With the rapid development of information technology and the widespread application of digital media, video has become an important medium for information dissemination, entertainment, and social interaction. However, in practical applications, limitations in aspects such as acquisition devices, transmission bandwidth, and storage capacity often result in reduced video resolution, which in turn degrades user viewing experience and the efficiency of video data utilization. To address these challenges, video super-resolution (VSR) technology has become an important research direction in the fields of computer vision and image processing.

Video super-resolution is a technique that converts low-resolution video into high-resolution video through advanced signal processing and image processing methods [1–4]. Since VSR can significantly enhance the visual quality of videos, it has broad application potential in various fields such as video surveillance, medical imaging, remote sensing, and the entertainment industry [5–7]. Although many works have proposed advanced VSR models, these models often require substantial computational resources to reconstruct high-resolution videos, making it difficult to deploy VSR models effectively on hardware platforms with limited computational resources. To solve this problem, there is an urgent need to design lightweight VSR models that can maintain high reconstruction quality while significantly reducing the computational complexity and inference time of VSR models, thereby improving their adaptability on resource-constrained hardware platforms.

In this paper, we propose a lightweight video super-resolution model (LightVSR) to enhance reconstruction quality while optimizing computational efficiency. As illustrated in Figure 1, LightVSR first utilizes the lightweight SPYNET model [8] for optical flow estimation to capture the motion information in the temporal dimension between the current frame and its neighboring frames. Then we design a novel lightweight feature aggregation module to fuse motion information from preceding and subsequent video frames via cross-layer shortcut connections to achieve multi-scale feature fusion, so as to facilitate finer detail restoration while maintaining a small number of model parameters. Subsequently, LightVSR employs a dual-layer pixel shuffle module [9] to perform the upsampling process, followed by a channel reduction module to obtain RGB color information. Finally, LightVSR combines low-resolution input frames up-sampled by bicubic interpolation [10] and high-resolution feature maps generated by the channel reduction module to produce the final high-resolution output frames.

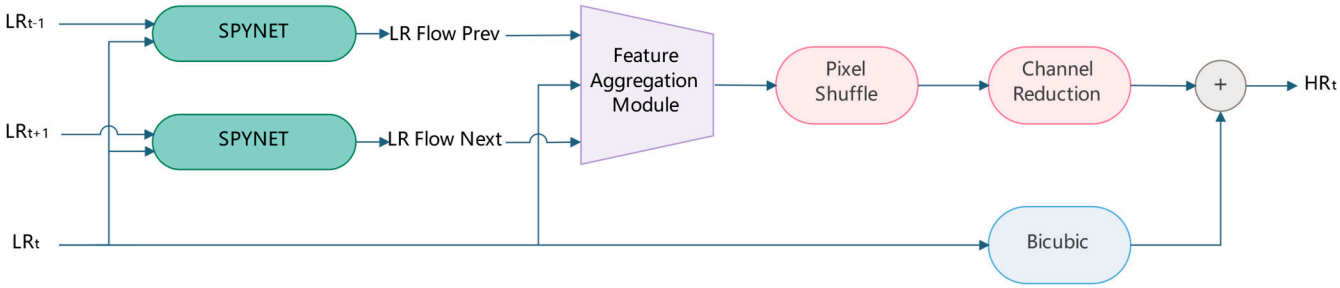

Figure 1. Overview of LightVSR networks.

The main contributions of this paper are as follows:

(1) We propose a lightweight VSR model that utilizes bidirectional optical flow for effective motion compensation, enabling efficient video frame reconstruction even on low-power devices.
(2) We design head-tail convolution, a novel convolutional operation that significantly reduces redundant computations while ensuring the effective representation of critical features across the entire feature tensor.
(3) We design a multi-input attention mechanism that integrates features from multiple sources by utilizing channel-wise attention to achieve comprehensive and efficient feature fusion.
(4) Combined head-tail convolution and multi-input attention mechanisms, we further design a feature aggregation module that leverages cross-layer shortcut connections to enhance computational efficiency while maintaining model performance.

2. Related Work

There have been many research studies on the study of video super-resolution models. SRCNN [11] was the first that applies a convolutional neural network to extract features from low-resolution images to reconstruct high-resolution ones. However, SRCNN only focuses on single-frame processing and does not consider temporal information in video sequences. To address this issue, EDVR [12] and MuCan [13] models were proposed to employ a sliding window for local feature extraction to take temporal information into account. The EDVR model employs the Pyramid, Cascading, and Deformable Convolutions [14] approach to align feature maps across different levels and incorporates temporal and spatial aggregation technology to comprehensively utilize temporal information within video sequences. The MuCan model considers multiple corresponding candidates for each pixel to mitigate reconstruction quality issues stemming from motion estimation errors to improve the accuracy of video super-resolution.

In another branch of VSR research, optical flow is utilized to precisely align and merge temporal information from consecutive frames, capturing fine motion details and enhancing the coherence of reconstructed frames in dynamic scenes. For instance, the FRVSR [15] model leverages temporal correlations between frames by introducing unidirectional optical flow to enhance video super-resolution quality. TecoGAN [16] model generates highly realistic video outputs by incorporating generative adversarial networks [17]. To further improve the accuracy and robustness of feature alignment, the BasicVSR [18] model integrates bidirectional optical flow, considering both forward and backward motion information, to fully extract inter-frame information. The SwinIR [19] model, on the other hand, combines the transformer architecture to enhance the capture of global context through a global attention mechanism. StableVSR [20] was the first model to employ diffusion models within a generative paradigm for video super-resolution (VSR), which utilizes a temporal texture guidance method to incorporate detail-rich and spatially aligned texture information synthesized from adjacent frames to enhance the perceptual quality of images. Xu et al. proposed an alignment method based on implicit resampling, where features and estimated motion are jointly learned through coordinate networks, with alignment implicitly achieved through window-based attention [21]. Lu et al. made the first attempt to tackle the non-trivial problem of learning implicit neural representations from events and RGB frames for VSR at random scales, effectively leveraging the high temporal resolution of events to complement RGB frames through spatial-temporal fusion [22].

However, the aforementioned VSR models generally require substantial computational loads and a large number of parameters, limiting their application on resource-constrained devices such as mobile phones, since these devices typically have limited computational resources and storage space.

There is also some research that focused on lightweight VSR models. Yi et al. proposed the local omniscient VSR and global omniscient VSR models based on the iterative, recurrent, and hybrid omniscient framework by leveraging long-term and short-term memory frames from the past, present, and future, as well as estimated hidden states, to achieve efficient real-time video super-resolution [23]. However, their training process is unstable due to the utilization of mixed-precision training methods. PP-MSVSR [24] employs a lightweight, multi-stage sliding-window framework to effectively utilize temporal information between frames. However, its reliance on multi-stage alignment may degrade the efficiency of video super-resolution in highly dynamic scenes. FDDCC-VSR [25] combines a deformable convolutional network [14], a lightweight spatial attention module, and improved bicubic interpolation to capture complex motion information and enhance high-frequency details. However, the introduction of deformable convolutional networks may compromise training stability, especially when dealing with complex motion patterns. Different from the above lightweight VSR models, we propose a lightweight VSR model that explicitly incorporates bidirectional optical flow estimation and a novel lightweight multi-scale feature aggregation module to achieve accurate motion alignment and effective information fusion across consecutive frames, so as to improve video super-resolution quality while maintaining a relatively small number of parameters.

3. Feature Aggregation Module

In this section, we will introduce the implementation of our designed feature aggregation module in Figure 1 in detail, including head-tail convolution, multi-input attention, and cross-layer shortcut connections.

3.1. Head-Tail Convolution

To design lightweight neural networks, one of the most common and efficient approaches is to reduce the number of convolutional layers, the number of convolutional kernels, or the size of the input data. However, such strategies often lead to a significant drop in model performance. To address this issue, we designed a novel head-tail convolution to perform efficient feature extraction while reducing computational overhead.

The core idea of head-tail convolution is to segment the input feature tensor along the channel dimension into three distinct regions of the head, middle, and tail. Figure 2a presents the conventional convolutional approach. In contrast, as shown in Figure 2b, the head and tail sections of the feature tensor participate in standard convolution operations, whereas the middle section is left unchanged. Although no convolution operations are explicitly applied to the intermediate section, these regions are still effectively captured by the receptive fields of the head and tail convolution, which can significantly reduce redundant computations while ensuring critical features from the entire feature tensor are effectively represented.

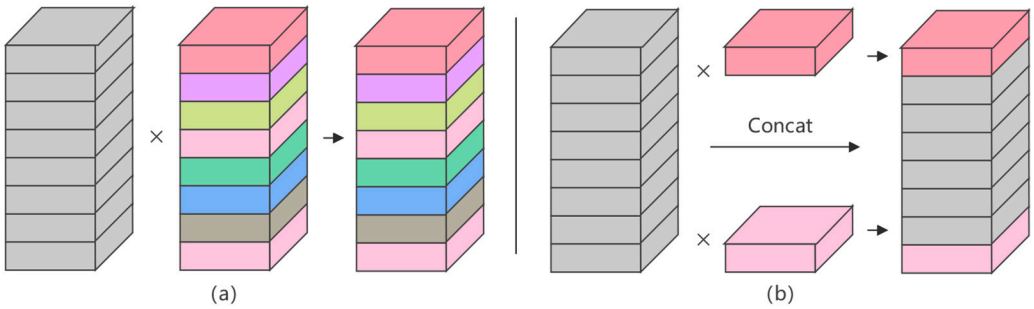

Figure 2. (**a**) Principle of conventional convolution; (**b**) principle of head-tail convolution.

Formally, denote the input feature tensor as $X = [X_1, X_2, \ldots, X_N]$, where N represents the tensor length. We partition X into head features, middle features and tail features as follows:

$$X_{head} = [X_1, X_2, \ldots, X_\alpha] \tag{1}$$

$$X_{mid} = [X_{\alpha+1}, \ldots, X_{N-\beta}] \tag{2}$$

$$X_{tail} = [X_{N-\beta+1}, \ldots, X_{N-1}, X_N] \tag{3}$$

where α and β denote the lengths of the head and tail segments, respectively. We then apply convolutional operations to the head and tail features:

$$X'_{head} = f_{conv}(X_{head}) \tag{4}$$

$$X'_{tail} = f_{conv}(X_{tail}) \tag{5}$$

The middle features remain unchanged:

$$X'_{mid} = X_{mid} \tag{6}$$

Finally, we concatenate above features to form the output feature vector:

$$X\prime = [X'_{head}, X'_{mid}, X'_{tail}] \tag{7}$$

To fully exploit head-tail convolution, we integrate it into a basic building block, called the HT-Block, as shown in Figure 3, in which the input tensor first applies head-tail convolution to extract features, then a GELU (Gaussian error linear unit) activation function layer is applied to accelerate training and improve generalization. Finally, a full 3×3 convolution layer is utilized to further fuse information of the head, middle, and

tail from head-tail convolution. Note that the HT-Block is designed as a general-purpose module that can be seamlessly integrated into various network architectures.

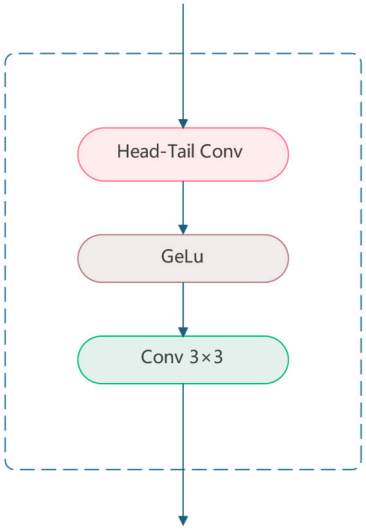

Figure 3. Architecture of HT-Block.

3.2. Multi-Input Attention

In video super-resolution, one of the feature fusion issues is the uneven distribution of information across different channels of feature maps. Some channels with critical information (e.g., complex details, dynamic objects) contribute more, while others with non-critical information (e.g., background, static objects) contribute less to the overall visual quality. This imbalance in information density may affect the quality of feature fusion and the subsequent reconstruction of high-resolution frames. To tackle this issue, we designed a multi-input attention mechanism to assign relative priority to each channel based on its contribution. By dynamically assigning higher weights to channels with complex textures or significant motion, the module can perform efficient feature fusion and thus benefit detailed reconstruction in key areas.

As shown in Figure 4, the multi-input attention mechanism dynamically integrates multiple feature tensors by assigning adaptive weights to each channel. For input tensor $T_i (i = 1, \ldots, \mathrm{n})$, where each tensor T_i has the shape B × C × H × W (batch size, channels, height, width), we first fuse all tensor via an element-wise summation:

$$T = \sum_{k=1}^{n} T_k \tag{8}$$

and squeeze the tensor along the channel dimension:

$$X_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} T_c(i, j) \tag{9}$$

where T_c represent tensor on channel c. Then we apply the SoftMax function for all X to obtain the weight S_i of X_i.

$$S_i = \frac{exp(X_i)}{\Sigma_{j=1}^{c} exp(X_j)} \quad for\ i = 1, 2, \ldots, C \tag{10}$$

where C is the number of channels. The resulting weight vector $S = [S_1, S_2, \ldots, S_C]$ is used to re-weigh each input channel such that the model can more effectively focus on key information and reduce the weight of channels with unimportant information. The final weighted output tensor F is computed as:

$$F = \sum_{k=1}^{n} \sum_{j=1}^{c} T_j^k \cdot S_j^k \tag{11}$$

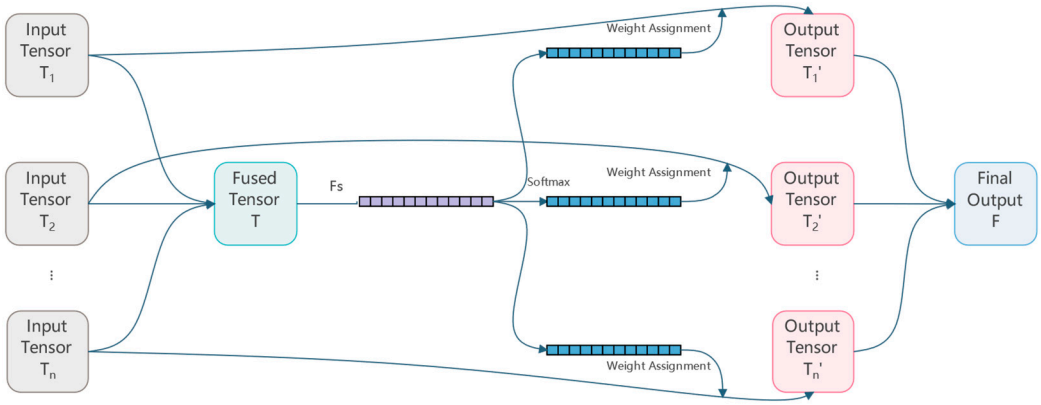

Figure 4. Architecture of multi-input attention.

3.3. Feature Aggregation

By utilizing the proposed lightweight HT-Block and multi-input attention mechanisms, we design an efficient feature aggregation module (FAM) with cross-layer shortcut connections to efficiently capture video frame features through multi-scale feature fusion. As shown in Figure 5, the FAM module adopts encoder–decoder architecture to achieve multi-scale feature extraction through dense cross-layer shortcut connections for video frame reconstruction and utilizes a series of HT-Blocks to extract features of low-resolution video frames. Then the features from the low-resolution and reconstructed video frames, along with the raw video frames, are all fed into multi-input attention for further multi-scale feature aggregation. Note that different from traditional paradigms of deepening the network layers (like ResNet [26]) or broadening the architecture (like Inception [27]) to boost performance, FAM leverages dense cross-layer short-cut connections and feature reuse to learn optical flow features from video frames while significantly reducing the number of parameters, so as to simplify the training process and enable more efficient feature learning.

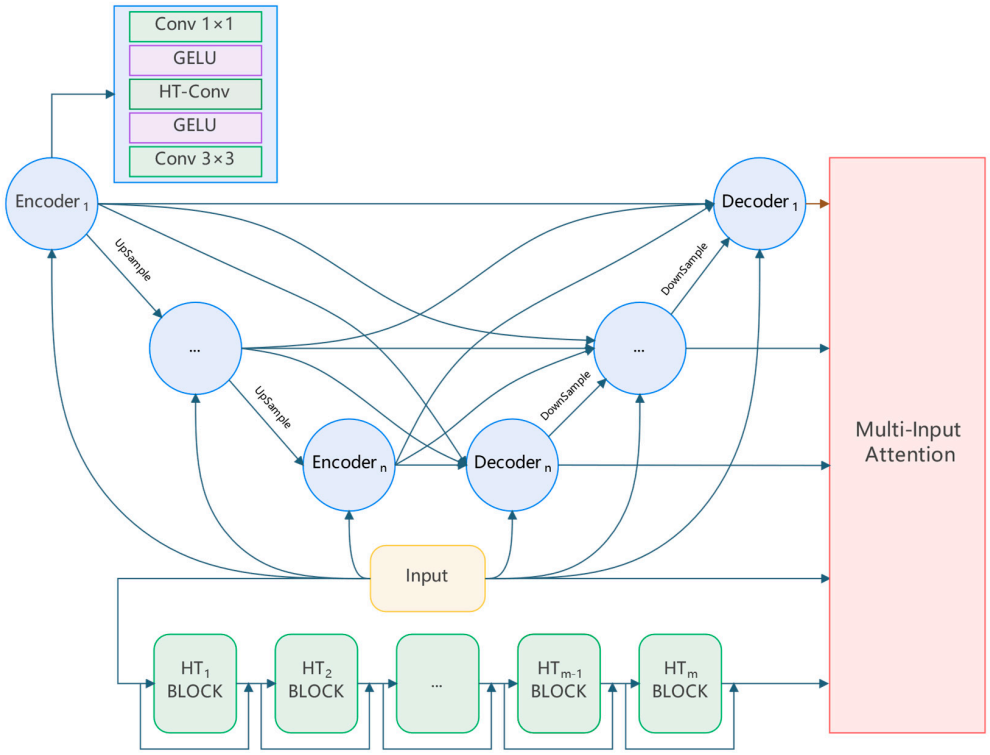

Figure 5. Architecture of feature aggregation network.

Specifically, denote the input to the FAM as:

$$input = concat(LR,\ LR\ flow\ next,\ LR\ flow\ prev) \tag{12}$$

Then the output of the i layer encoder can be represented as:

$$encoder_i = \varphi(concat(encoder_{i-1},\ input)) \tag{13}$$

where φ represents a series of convolutional operation, *concat* is the tensor concatenation operation, $encoder_{i-1}$ is the output of the $i-1$ layer encoder, and *input* are the collection of bi-directional optical flow information and low-resolution video frames. The output of the i layer decoder can be represented as:

$$decoder_i = \varphi(concat(encoder_{1:n}, decoder_{i-1},\ input)) \tag{14}$$

where $encoder_{1:n}$ are the output of all encoders, $decoder_{i-1}$ is the output of the $i-1$ layer decoder. For the j layer HT-Block, its output HT_j is represented as follows:

$$HT_j = Conv(GELU(\phi(HT_{j-1}))) + HT_{j-1} \tag{15}$$

where ϕ represents the head-tail convolution operation. Finally, we integrate the outputs from the input layer, decoders, and HT-Blocks into the multi-input attention mechanism as follows:

$$Output = Att(decoder_{1:n},\ input,\ HT_m) \tag{16}$$

where m represents the total number of HT layers, $decoder_{1:n}$ are the output of all decoders. This integration leverages both shallow features that capture fine-grained texture details and deep features that encode high-level structural and semantic information to allow the multi-input attention mechanism to dynamically weigh the contributions from each component, achieving efficient fusion of shallow and deep features.

4. Results

4.1. Experiment

In the experiments, we used the REDS [28] and Vimeo-90K [29] datasets for training and the Vid4, UDM10, and Vimeo-90K-T datasets for testing. To evaluate the model's performance under different downsampling conditions, we conducted 4× downsampling tests using both bicubic downsampling (BI) and blur downsampling (BD).

Our proposed LightVSR model employs a pre-trained SPYNET as the optical flow prediction module and is trained on four NVIDIA GeForce GTX Titan Xp GPUs (Nvidia, Santa Clara, CA, USA) in Cuda 11.3 with Python 3.8.13 and Pytorch 1.12.1. During training, we set the batch size to 2, and each batch contains 15 consecutive frames. The number of training iterations is 300,000, with AdamW as the optimizer and an initial learning rate of 2×10^{-4}. To ensure stable training, we set the moving average of gradients to 0.9 and the moving average of gradient squares to 0.99. Additionally, to fine-tune the model, the minimum learning rate was set to 1×10^{-7}. We used the Charbonnier loss function as in [30] to improve reconstruction quality.

To comprehensively evaluate the performance of the LightVSR model, we compared LightVSR with a range of state-of-the-art video super-resolution models, including bicubic interpolation [10], VESPCN [31], SPMC [32], TOFlow [29], FRVSR [15], DUF [33], RBPN [34], EDVR-M [12], EDVR [5], PFNL [35], MuCAN [13], TGA [36], RLSP [37], RSDN [38], RRN [39], D3Dnet [40], RSTT [41], STDAN [42], and FDDCC-VSR [25], and adopted peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) to evaluate the quality of reconstruction video.

PSNR is a standard metric for assessing the quality of reconstructed video frames. It is defined as the logarithmic ratio between the peak signal and noise levels, where a

higher PSNR value indicates that the reconstructed video frame is much more similar to the reference video frame. The PSNR calculation formula is as follows:

$$PSNR = 10 * log_{10}\left(\frac{MAX^2}{MSE}\right) \tag{17}$$

where MAX denotes the maximum possible pixel value in the video frame; MSE represents the mean squared error between the reconstructed and ground truth frame.

SSIM can be used to measure the similarity between a reconstructed frame and a reference frame in terms of luminance, contrast, and structural attributes, and its value ranges from 0 to 1. The larger SSIM value indicates the higher similarity. The calculation formula for SSIM is:

$$\mathrm{SSIM}(x, y) = \frac{\left(2\mu_x\mu_y + c_1\right)\left(2\sigma_{xy} + c_2\right)}{\left(\mu_x^2 + \mu_y^2 + c_1\right)\left(\sigma_x^2 + \sigma_y^2 + c_2\right)} \tag{18}$$

where μ_x and σ_x represent the mean value and variance of pixels in frame x, respectively, σ_{xy} is the pixel covariance between frame x and y, and c_1 and c_2 are stabilization constants.

As illustrated in Table 1, LightVSR demonstrates competitive performance compared with other VSR models across all test datasets. Specifically, under the BI degradation model, LightVSR achieved PSNR/SSIM values of 30.71/0.8780 on the REDS4 dataset, 36.69/0.9406 on Vimeo-90K-T, 26.95/0.8188 on Vid4, and 39.25/0.9656 on UDM10. Under the BD degradation model, LightVSR achieved PSNR/SSIM values of 36.91/0.9444 on Vimeo-90K-T and 27.37/0.8360 on Vid4. Although EDVR performs better than LightVSR in terms of PSNR/SSIM, it has a parameter size of 20.6 M and requires 378 ms for inference per frame. In contrast, LightVSR contains only 3.5 M parameters and achieves an inference time of just 35 ms per frame. The key advantage of our approach lies in its lightweight design, which significantly reduces computational requirements. In addition, LightVSR also strikes a good balance between model size and runtime efficiency. It can be seen that compared with RSTT and STDAN, our proposed model achieves better performance in terms of PSNR and SSIM across all test datasets under the BI degradation model while using fewer parameters and faster runtime. Compared with FDDCC-VSR, although LightVSR has slightly more parameters, it performs better in most cases.

From a subjective comparison perspective, it is seen that LightVSR offers significant advantages over other VSR models, such as FRVSR, VESRCN, and SOFVSR. As shown in Figure 6a, LightVSR can preserve sharp details, especially in facial features, where it surpasses its competitors in rendering finer details like eyes and facial contours with greater clarity while also effectively reducing blurring in areas with motion or complex textures, such as clothing. As shown in Figure 6b–d, the texture of trees, buildings, and fonts in the results produced by LightVSR are clearer, indicating better structural preservation compared to the competing models that tend to smooth out or blur these areas. Consequently, the LightVSR model provides a superior balance of detail retention, sharpness, and natural visual quality, making it highly effective for video super-resolution tasks.

4.2. Ablation Study on HT-Convolution

To validate the effectiveness and contributions of the HT-Conv module, we conducted an ablation study to compare the performance of the LightVSR model with and without the HT-Conv module. The ablation study was performed on three widely used benchmark datasets: REDS4, Vimeo-90k-T, and Vid4. The results are summarized in Table 2 below.

From the ablation study, it is seen that the inclusion of the HT-Conv module significantly improves the performance of the VSR model across all evaluated metrics and datasets. Specifically, the application of the HT-Conv module increases the parameter num-

ber by 0.14 million. Despite this modest increase in parameters, the model demonstrates a significant improvement in both PSNR and SSIM scores. For example, in the REDS4 dataset, the HT-Conv module achieves a 3.56 dB gain in PSNR and a 0.11 increase in SSIM. These results highlight the effectiveness of the HT-Conv module in enhancing the model's generalization ability and overall performance. The experimental results demonstrate that the HT-Conv module significantly enhances the quality of video super-resolution, particularly in preserving fine details and structural integrity.

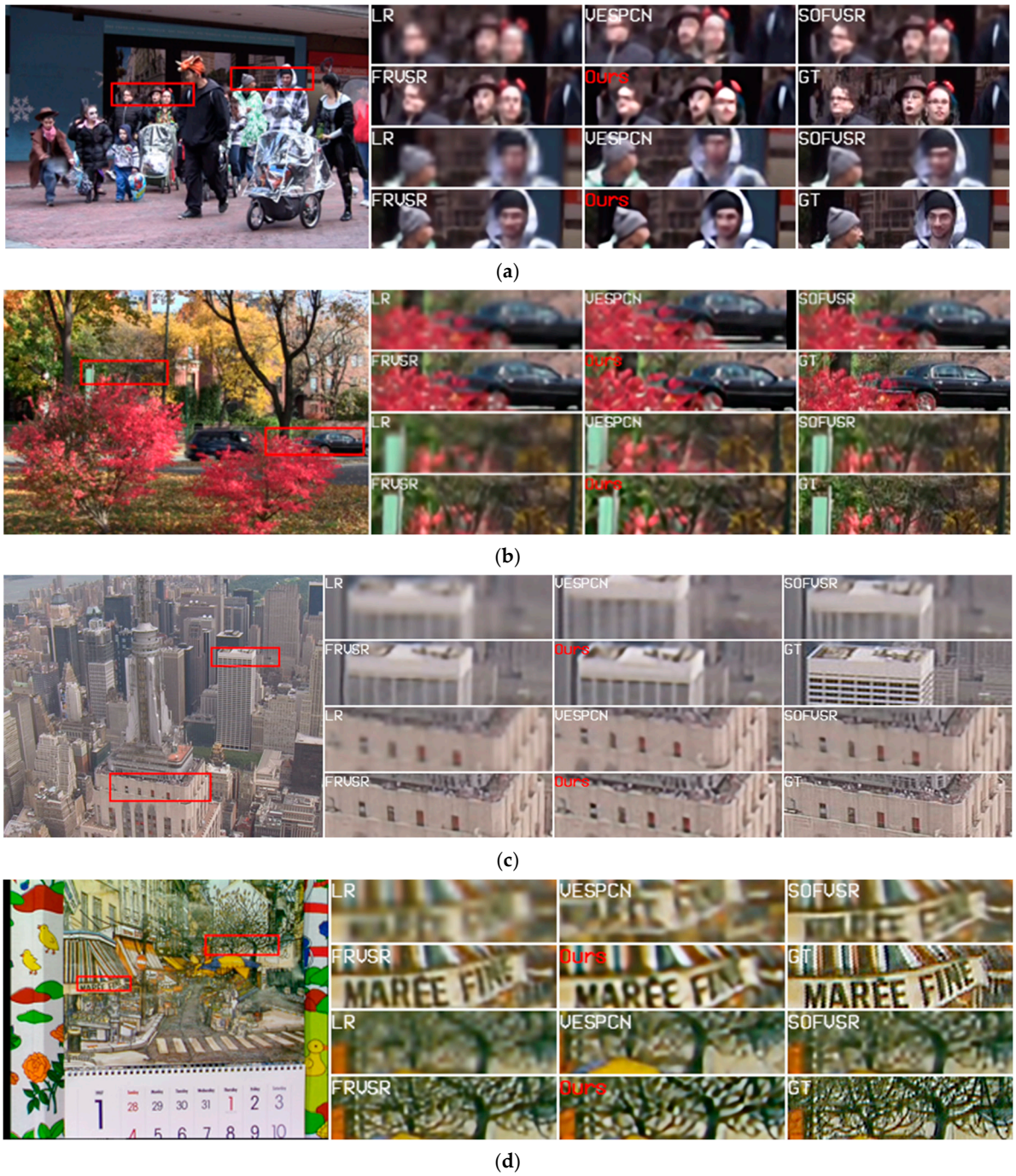

Figure 6. Qualitative comparison of the Vid4 datasets: (**a**) walk, (**b**) foliage, (**c**) city, (**d**) calendar.

Table 1. Performance comparison (PSNR/SSIM). All results are calculated on Y-channel except REDS4 (RGB-channel). The runtime is computed on an LR size of 180 × 320. Notably, Params is expressed in millions (M), and the runtime is reported in milliseconds (ms).

				BI Degradation			BD Degradation		
	Params	**Runtime**	**Fps**	**REDS4**	**Vimeo-90K**	**Vid4**	**UDM10**	**Vimeo-90K**	**Vid4**
(1981) Bicubic [10]	-	-	-	26.14/0.7292	31.32/0.8684	23.78/0.6347	28.47/0.8253	31.30/0.8687	21.80/0.5246
(2017) VESPCN [31]	-	-	-	-	-	25.35/0.7557	-	-	-
(2017) SPMC [32]	-	-	-	-	-	25.88/0.7752	-	-	-
(2019) TOFlow [29]	-	-	-	27.98/0.7990	33.08/0.9054	25.89/0.7651	36.26/0.9438	34.62/0.9212	-
(2018) FRVSR [15]	5.1	137	7.30	-	-	-	37.09/0.9522	35.64/0.9319	26.69/0.8103
(2018) DUF [33]	5.8	974	1.03	28.63/0.8251	-	-	38.48/0.9605	36.87/0.9447	27.38/0.8329
(2019) RBPN [34]	12.2	1507	0.66	30.09/0.8590	37.07/0.9435	27.12/0.8180	38.66/0.9596	37.20/0.9458	
(2019) EDVR-M [12]	3.3	118	8.47	30.53/0.8699	37.09/0.9446	27.10/0.8186	39.40/0.9663	37.33/0.9484	27.45/0.8406
(2019) EDVR [12]	20.6	378	2.65	31.09/0.8800	37.61/0.9489	27.35/0.8264	39.89/0.9686	37.81/0.9523	27.85/0.8503
(2019) PFNL [35]	3.0	295	3.39	29.63/0.8502	36.14/0.9363	26.73/0.8029	38.74/0.9627	-	27.16/0.8355
(2020) MuCAN [13]	-	-	-	30.88/0.8750	37.32/0.9465	-	-	-	-
(2020) TGA [36]	5.8	-	-	-	-	-	-	37.59/0.9516	27.63/0.8423
(2019) RLSP [37]	4.2	49	20.41	-	-	-	38.48/0.9606	36.49/0.9403	27.48/0.8388
(2020) RSDN [38]	6.2	94	10.64	-	-	-	39.35/0.9653	37.23/0.9471	27.92/0.8505
(2020) RRN [39]	3.4	45	22.22	-	-	-	38.96/0.9644	-	27.69/0.8488
(2020) D3Dnet [40]	2.6	-	-	30.51/0.8657	35.65/0.9331	26.52/0.7993	-	-	-
(2022) RSTT [41]	4.5	38	26.32	30.11/0.8613	36.58/0.9381	26.29/0.7941	-	-	-
(2023) STDAN [42]	8.3	72	13.89	29.98/0.8613	35.70/0.9387	26.28/0.8041	-	-	-
(2024) FDDCC-VSR [25]	1.8	-	-	30.55/0.8663	35.73/0.9357	26.79/0.8334	-	-	-
LightVSR (ours)	3.5	35	28.57	30.71/0.8780	36.69/0.9406	26.95/0.8188	39.25/0.9656	36.91/0.9444	27.37/0.8360

Table 2. Performance comparison of HT-Conv application.

	Params	**REDS4**				**Vimeo-90k-T**				**Vid4**			
		PSNR ↑	**SSIM ↑**	**MAE ↓**	**MSE ↓**	**PSNR ↑**	**SSIM ↑**	**MAE ↓**	**MSE ↓**	**PSNR ↑**	**SSIM ↑**	**MAE ↓**	**MSE ↓**
HT-Conv Applied	3.55 M	30.72	0.8784	0.0176	0.0009	35.99	0.9366	0.0126	0.0007	26.84	0.8121	0.0355	0.0035
HT-Conv Not Applied	3.41 M	30.13	0.8640	0.0188	0.0011	35.52	0.9309	0.0133	0.0008	26.29	0.7902	0.0377	0.0040

4.3. Comparison of Feature Fusion Methods

In this section, we compare the proposed multi-input attention mechanism with the concatenation mechanism, where tensors are concatenated, as well as the direct tensor addition mechanism, which involves simply adding the tensors together. The structural comparisons of these three mechanisms are depicted in Figure 7.

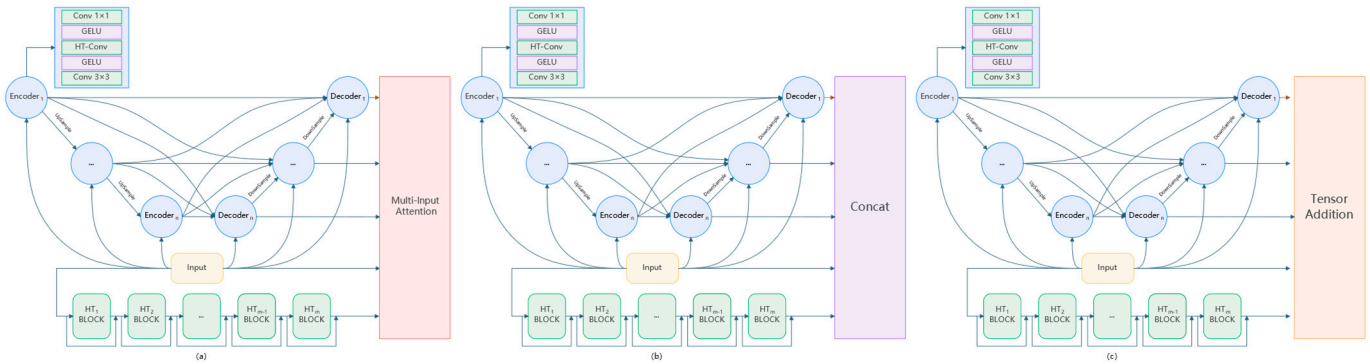

Figure 7. Comparison of different feature fusion methods in FAM: (**a**) multi-input attention; (**b**) concatenate; (**c**) tensor addition.

As illustrated in Table 3, our proposed multi-input attention mechanism achieves the best performance compared to the concatenation and direct tensor addition mechanisms. This is because the proposed attention mechanism can adaptively assign weights to different channels for achieving more refined and efficient feature fusion compared to traditional

methods that employ fixed or uniform feature combination strategies. In contrast, the concatenation and direct tensor addition methods lack such adaptability, often leading to suboptimal feature combination, which in turn affects performance. By employing the attention mechanism, our model achieves more precise feature integration, ultimately enhancing the quality of the video super-resolution results.

Table 3. Comparison of three feature fusion methods.

	REDS4				Vimeo-90k-T				Vid4			
	PSNR ↑	**SSIM ↑**	**MAE ↓**	**MSE ↓**	**PSNR ↑**	**SSIM ↑**	**MAE ↓**	**MSE ↓**	**PSNR ↑**	**SSIM ↑**	**MAE ↓**	**MSE ↓**
Attention mechanism	30.71	0.8780	0.0176	0.0009	36.00	0.9368	0.0125	0.0007	26.95	0.8188	0.0349	0.0034
Concatenate	30.65	0.8769	0.0177	0.0010	36.03	0.9369	0.0147	0.0007	26.93	0.8188	0.0351	0.0034
Tensor addition	30.61	0.8764	0.0178	0.0010	35.93	0.9359	0.0126	0.0007	26.80	0.8143	0.0354	0.0035

5. Conclusions

In this paper, we proposed a lightweight video super-resolution model that utilizes an innovative feature aggregation module to improve video quality by efficiently reconstructing high-definition frames from compressed, low-resolution inputs. LightVSR incorporates several novel mechanisms, such as head-tail convolution, cross-layer shortcut connections, and multi-input attention, to boost computational speed while maintaining superior video super-resolution capabilities. Comprehensive tests demonstrate that LightVSR attains a frame rate of 28.57 FPS and a PSNR of 39.25 dB on the UDM10 dataset and 36.91 dB on the Vimeo-90k dataset, thereby confirming its efficiency and efficacy.

Author Contributions: Conceptualization, G.H. and N.L.; methodology, G.H. and N.L.; data curation, M.Z.; formal analysis, L.Z.; funding acquisition, G.H., J.L. (Jianming Liu) and J.L. (Jun Li); investigation, N.L.; resources, G.H.; software, N.L.; project administration, L.Z. and J.L. (Jun Li); visualization, N.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the Key Research and Development Program of Guangxi under Grant GuiKe AD22035118 and AB23075178, the Guangxi Innovation Driven Development Special Fund Project under Grant GuiKe AA19046004, the Guangxi Key Laboratory of Trusted Software (No. KX202324).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/strix214/LightVSR/ (accessed on 24 January 2025).

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

References

1. Jiang, Y.; Nawała, J.; Feng, C.; Zhang, F.; Zhu, X.; Sole, J.; Bull, D. RTSR: A Real-Time Super-Resolution Model for AV1 Compressed Content. *arXiv* **2024**, arXiv:2411.13362.
2. Sun, J.; Yuan, Q.; Shen, H.; Li, J.; Zhang, L. A Single-Frame and Multi-Frame Cascaded Image Super-Resolution Method. *Sensors* **2024**, *24*, 5566. [CrossRef] [PubMed]
3. Ko, H.-k.; Park, D.; Park, Y.; Lee, B.; Han, J.; Park, E. Sequence Matters: Harnessing Video Models in Super-Resolution. *arXiv* **2024**, arXiv:2412.11525.
4. Li, Y.; Yang, X.; Liu, W.; Jin, X.; Jia, X.; Lai, Y.; Liu, H.; Rosin, P.L.; Zhou, W. TINQ: Temporal Inconsistency Guided Blind Video Quality Assessment. *arXiv* **2024**, arXiv:2412.18933.
5. Wen, Y.; Zhao, Y.; Liu, Y.; Jia, F.; Wang, Y.; Luo, C.; Zhang, C.; Wang, T.; Sun, X.; Zhang, X. Panacea: Panoramic and controllable video generation for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 6902–6912.

6. Feng, R.; Li, C.; Loy, C.C. Kalman-inspired feature propagation for video face super-resolution. In Proceedings of the European Conference on Computer Vision, Paris, France, 26–27 March 2025; pp. 202–218.
7. Wang, Q.; Yin, Q.; Huang, Z.; Jiang, W.; Su, Y.; Ma, S.; Zhang, J. Compressed Domain Prior-Guided Video Super-Resolution for Cloud Gaming Content. *arXiv* **2025**, arXiv:2501.01773.
8. Ranjan, A.; Black, M.J. Optical Flow Estimation using a Spatial Pyramid Network. In Proceedings of the 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2720–2729.
9. Shi, W.Z.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A.P.; Bishop, R.; Rueckert, D.; Wang, Z.H. 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), Seattle, WA, USA, 27–30 June 2016; pp. 1874–1883.
10. Keys, R. Cubic convolution interpolation for digital image processing. *IEEE Trans.Acoust. Speech Signal Process.* **1981**, *29*, 1153–1160. [CrossRef]
11. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 295–307. [CrossRef] [PubMed]
12. Wang, X.; Chan, K.C.; Yu, K.; Dong, C.; Change Loy, C. Edvr: Video restoration with enhanced deformable convolutional networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019.
13. 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 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part X 16, pp. 335–351.
14. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 764–773.
15. Sajjadi, M.S.; Vemulapalli, R.; Brown, M. Frame-recurrent video super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6626–6634.
16. Chu, M.; Xie, Y.; Leal-Taixé, L.; Thuerey, N. Temporally coherent gans for video super-resolution (tecogan). *arXiv* **2018**, arXiv:1811.09393.
17. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems 27 (NIPS 2014), Montreal, QC, Canada, 8–13 December 2014; Volume 27.
18. Chan, K.C.; Wang, X.; Yu, K.; Dong, C.; Loy, C.C. Basicvsr: The search for essential components in video super-resolution and beyond. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 4947–4956.
19. Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; Timofte, R. Swinir: Image restoration using swin transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 10–17 October 2021; pp. 1833–1844.
20. Rota, C.; Buzzelli, M.; van de Weijer, J. Enhancing Perceptual Quality in Video Super-Resolution through Temporally-Consistent Detail Synthesis using Diffusion Models. *arXiv* **2023**, arXiv:2311.15908.
21. Xu, K.; Yu, Z.; Wang, X.; Mi, M.B.; Yao, A. Enhancing Video Super-Resolution via Implicit Resampling-based Alignment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 2546–2555.
22. Lu, Y.; Wang, Z.; Liu, M.; Wang, H.; Wang, L. Learning spatial-temporal implicit neural representations for event-guided video super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 1557–1567.
23. Yi, P.; Wang, Z.; Jiang, K.; Jiang, J.; Lu, T.; Tian, X.; Ma, J. Omniscient video super-resolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 10–17 October 2021; pp. 4429–4438.
24. Jiang, L.; Wang, N.; Dang, Q.; Liu, R.; Lai, B. PP-MSVSR: Multi-stage video super-resolution. *arXiv* **2021**, arXiv:2112.02828.
25. Wang, X.; Yang, X.; Li, H.; Li, T. FDDCC-VSR: A lightweight video super-resolution network based on deformable 3D convolution and cheap convolution. *Vis. Comput.* **2024**, 1–13. [CrossRef]
26. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
27. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
28. Nah, S.; Baik, S.; Hong, S.; Moon, G.; Son, S.; Timofte, R.; Mu Lee, K. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019.

29. 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]
30. Charbonnier, P.; Blanc-Feraud, L.; Aubert, G.; Barlaud, M. Two deterministic half-quadratic regularization algorithms for computed imaging. In Proceedings of the 1st International Conference on Image Processing, Austin, TX, USA, 13–16 November 1994; pp. 168–172.
31. Caballero, J.; Ledig, C.; Aitken, A.; Acosta, A.; Totz, J.; Wang, Z.; Shi, W. Real-time video super-resolution with spatio-temporal networks and motion compensation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4778–4787.
32. Tao, X.; Gao, H.; Liao, R.; Wang, J.; Jia, J. Detail-revealing deep video super-resolution. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4472–4480.
33. Jo, Y.; Oh, S.W.; Kang, J.; Kim, S.J. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3224–3232.
34. Haris, M.; Shakhnarovich, G.; Ukita, N. Recurrent back-projection network for video super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 3897–3906.
35. Yi, P.; Wang, Z.; Jiang, K.; Jiang, J.; Ma, J. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3106–3115.
36. Isobe, T.; Li, S.; Jia, X.; Yuan, S.; Slabaugh, G.; Xu, C.; Li, Y.-L.; Wang, S.; Tian, Q. Video super-resolution with temporal group attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8008–8017.
37. Fuoli, D.; Gu, S.; Timofte, R. Efficient video super-resolution through recurrent latent space propagation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Republic of Korea, 27–28 October 2019; pp. 3476–3485.
38. Isobe, T.; Jia, X.; Gu, S.; Li, S.; Wang, S.; Tian, Q. Video super-resolution with recurrent structure-detail network. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XII 16, pp. 645–660.
39. Isobe, T.; Zhu, F.; Jia, X.; Wang, S. Revisiting temporal modeling for video super-resolution. *arXiv* **2020**, arXiv:2008.05765.
40. 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]
41. Geng, Z.; Liang, L.; Ding, T.; Zharkov, I. Rstt: Real-time spatial temporal transformer for space-time video super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 17441–17451.
42. Wang, H.; Xiang, X.; Tian, Y.; Yang, W.; Liao, Q. Stdan: Deformable attention network for space-time video super-resolution. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *35*, 10616. [CrossRef] [PubMed]

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.