sensors

Article

Real-World Video Super-Resolution with a Degradation-Adaptive Model

Mingxuan Lu and Peng Zhang *

School of Electronics and Communication Engineering, Shenzhen Campus of Sun Yat-Sen University, Shenzhen 518107, China; lumx7@mail2.sysu.edu.cn

* Correspondence: zhangpeng5@mail.sysu.edu.cn

Abstract: Video super-resolution (VSR) remains challenging for real-world applications due to complex and unknown degradations. Existing methods lack the flexibility to handle video sequences with different degradation levels, thus failing to reflect real-world scenarios. To address this problem, we propose a degradation-adaptive video super-resolution network (DAVSR) based on a bidirectional propagation network. Specifically, we adaptively employ three distinct degradation levels to process input video sequences, aiming to obtain training pairs that reflect a variety of real-world corrupted images. We also equip the network with a pre-cleaning module to reduce noise and artifacts in the low-quality video sequences prior to information propagation. Additionally, compared to previous flow-based methods, we employ an unsupervised optical flow estimator to acquire a more precise optical flow to guide inter-frame alignment. Meanwhile, while maintaining network performance, we streamline the propagation network branches and the structure of the reconstruction module of the baseline network. Experiments are conducted on datasets with diverse degradation types to validate the effectiveness of DAVSR. Our method exhibits an average improvement of 0.18 dB over a recent SOTA approach (DBVSR) in terms of the PSNR metric. Extensive experiments demonstrate the effectiveness of our network in handling real-world video sequences with different degradation levels.

Keywords: real-world video super-resolution; degradation-adaptive; pre-cleaning; lightweight network; deep learning

check for updates

Citation: Lu, M.; Zhang, P. Real-World Video Super-Resolution with a Degradation-Adaptive Model. *Sensors* **2024**, *24*, 2211. https://doi.org/10.3390/s24072211

Academic Editor: Marcin Woźniak

Received: 28 February 2024
Revised: 25 March 2024
Accepted: 27 March 2024
Published: 29 March 2024

1. Introduction

Sensors play a crucial role in numerous aspects of our everyday lives, finding applications in diverse fields such as environmental monitoring, traffic management, medical health, and robotics [1–3]. Although video sensors have been designed to operate efficiently in environments with a low latency and complexity, they continue to impose restrictions on the quality of input videos. In order to tackle this challenge, various video processing techniques have been employed for restoration, including video super-resolution [4–8] and video denoising [9–11]. These methods strive to improve the overall quality of video input, ensuring clearer and more refined outputs across different applications.

Video super-resolution (VSR) is an active area of research within the visual domain, which aims at enhancing the resolution of low-resolution (LR) videos [12]. Due to the introduction and application of convolution neural network (CNN) methodologies, the development of VSR algorithms has been progressing. For instance, in the pioneering work of SOFVSR [13], the motion-compensated low-resolution input undergoes processing within an optical flow reconstruction network, ultimately generating high-resolution video frames. TDAN [7] implicitly integrates motion information into low-quality frames, which can be extracted without computing the optical flow. BasicVSR uses a bidirectional propagation network, which relies on optical flow alignment and is employed to aggregate information across the spatio-temporal dimension [6]. However, most of these ignore complex degradation processes in the real world. They assume an ideal degradation process (bicubic downsampling) from HR to LR [14], which leads to performance drops when processing real-world videos because of mismatched degradation. In contrast, blind video super-resolution aims to restore low-resolution images suffering from unknown and complex degradations [15]. The remarkable work of Real-ESRGAN enhances the image perceptual quality, which introduces complex degradation operations including noise, blur, and down-sampling [16]. However, its recovery ability is limited by the practical usage. In particular, it generates better details in severely degraded LR inputs than in mildly degraded LR inputs.

Considering the aforementioned challenges, we propose a novel video super-resolution framework (DAVSR), which adaptively restores video sequences with varying degrees of degradation. Similar to Real-ESRGAN, our goal is to improve the quality of real-world low-resolution (LR) videos through synthesizing training pairs with a practical degradation process. Real-ESRGAN solely employs second-order degradation processes to synthesize its training data [16], lacking the ability to effectively capture the characteristics of mildly degraded low-resolution (LR) videos. This motivates us to expand the degradation model into an adaptive degradation model to restore LR videos degraded to various degrees. Specifically, we use "first-order" degradation modeling to synthesize training pairs for mildly degraded LR videos and "high-order" degradation modeling for severely degraded videos. The "first-order" degradation results from complex combinations of different degradation processes, including camera blur, sensor noise, and video compression. The "high-order" degradation is represented through the application of multiple repeated degradation operations.

Recently, Huang et al. [17] showed that temporal redundancy brings adverse effects to the information propagation in most existing VSR methods. For example, this information also leads to exaggerated artifacts, arising from the accumulation of errors during propagation. As a result, we used an image pre-cleaning module [18–20] to reduce the adverse effects of noise and artifacts on propagation. Furthermore, to establish accurate connections among multiple frames, we have integrated an unsupervised optical flow estimator to make the most of its potential. Finally, to balance performance and the computational cost, we simplified the propagation network and reconstruction module of BasicVSR, further reducing model complexity.

The primary contributions of this work are outlined below:

- We propose a degradation-adaptive process to model more practical degradations, which notably boosts the model's capability to enhance the resolution of videos affected by various degradation levels. It can attain a superior visual performance compared to prior works, making it more applicable in real-world scenarios.
- Moreover, we present an exquisite image pre-cleaning module for removing degradations in the input images prior to information propagation. It can reduce the adverse effects of temporal redundancy in video super-resolution.
- Finally, the introduced unsupervised optical flow estimator mitigates the inaccuracies in optical flow encountered with previous methods. Furthermore, a new lightweight reconstruction block is proposed to further reduce the complexity. Many experimental results validate the superiority of the proposed method over state-of-the-art methods.

The remainder of this paper is structured as follows: Section 2 introduces some related works on video super-resolution. In Section 3, the proposed DAVSR network is described in detail. Section 4 demonstrates the experimental results of our method. Finally, the conclusions and future work are presented in Section 5.

2. Related Work

2.1. Real-World Video Super-Resolution

The goal of video super-resolution is to reconstruct high-resolution video frames from the low-resolution ones. Most existing VSR methods are trained with predefined degradations, such as bicubic downsampling. However, they exhibit significant deterioration when confronted with unknown degradations in practical applications. Meanwhile, scaling from

VSR to real-world VSR is not trivial due to the difficulty of modeling complex degradations in the wild. To solve this problem, recent blind super-resolution [14,15,21,22] proposes a solution where the inputs are assumed to be degraded through a recognized process characterized by unknown parameters. Specifically, they train the network through a predefined process with unknown parameters. Despite the fact that this network can restore some degraded videos, it is still insufficient for real-world generalization. Fortunately, Wang et al. proposed a second-order degradation process to simulate complex degradation, including blur, noise, compression, and other factors. These methods exhibit an encouraging performance in real-world super-resolution. However, we have observed that they generate better details in severely degraded low-resolution (LR) inputs compared to those with mild degradation. In this work, we explore the process of degradation modeling and propose a novel video super-resolution framework, which adaptively recovers videos with different levels of degradation.

2.2. Optical Flow Estimation

To explore the temporal dependency between consecutive video frames, numerous video super-resolution methods based on deep learning estimate optical flows from low-resolution (LR) frames for motion compensation. The quality of the optical flow directly influences the efficacy of these methods. With the advancements in deep learning, some optical flow estimation methods trained by CNN networks achieve better results than non-learning methods. The model's generalization ability is constrained by the domain difference between the optical flow of the synthetic dataset and the real-world dataset. Therefore, Yu et al. [23] initially proposed an unsupervised method for optical flow learning, emphasizing brightness constancy and motion smoothness. Wang et al. [24] propose using an unsupervised optical flow estimator to bypass the necessity for labeled data. Further, for the problem of large motion and model occlusion, they proposed a new warping module to improve the performance of the unsupervised optical flow method. Although these methods [23–25] have become sophisticated, the estimated optical flow accuracy is still worse than that of SOTA supervised methods. In addition, they have not solved the difficulty of estimating an accurate optical flow from severely degraded inputs. We enhance the quality of the low-quality (LQ) flows by employing a knowledge distillation approach to learn high-quality (HQ) image characteristics.

2.3. Image Pre-Cleaning

The effectiveness of existing video super-resolution algorithms largely relies on leveraging temporal information from adjacent frames. Inspired by Huang et al. [17], the introduction of temporal redundancy can negatively impact information propagation. Therefore, it is crucial to eliminate degradations before propagation which is beneficial for suppressing artifacts in the outputs. While analogous concepts have been discussed within the realm of single-image super-resolution (SISR) [18,19,26,27], they have not undergone a thorough exploration in the realm of VSR. Furthermore, we devise a dynamic pre-cleaning process for video frames with different levels of degradation. For mildly degraded low-quality video sequences, we can use a cleaning strategy once to remove redundant temporal information. Excessive cleaning can result in the loss of valuable information. In particular, the temporal redundancy introduced by the degradation process in video sequences is mitigated through the utilization of a pre-cleaning module prior to network propagation. A series of ablation experiments were conducted to confirm the effectiveness of our pre-cleaning module.

3. Proposed Method

In this section, we propose an unsupervised flow-aligned degradation-adaptive network for real-world video super-resolution, which we have denoted as DAVSR. As depicted in Figure 1, the proposed method is composed of four modules: the degradation-adaptive module, the image pre-cleaning module, the feature alignment module, and the reconstruction module. Let y_i represent the input video sequence, which go through a degradation module to generate x_i, the low-quality video sequence derived from y_i. Subsequently, x_i passes through an alignment and reconstruction module to yield the super-resolution video sequence SR_i.

In the following sections, the details of the proposed module are introduced.

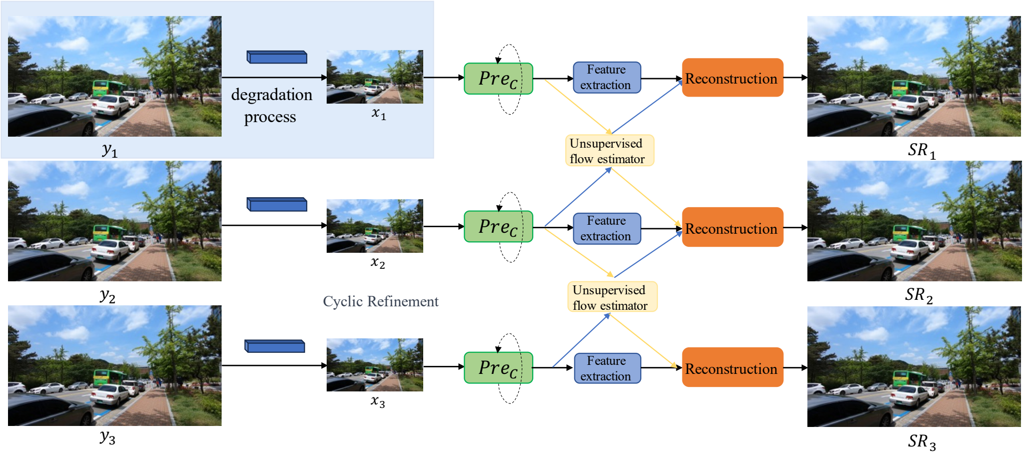

Figure 1. The architecture of the DAVSR network, when the number of input frames N = 3. Degradation process: aims to generate low-resolution frames.

3.1. The Degradation-Adaptive Model

When the classical degradation model is employed to synthesize training pairs [28], it becomes challenging to acquire high-quality real-world LR–HR image pairs. The network model exhibits limitations in learning from image pairs that faithfully represent real-world scenarios, leading to suboptimal performance in the restoration of corrupted video sequences under real-world conditions. The classical degradation model incorporates typical fundamental degradations, such as blur, resize, and compression, and can be considered as first-order modeling. However, degradation processes in real-world scenarios are highly diverse, comprising a multitude of steps such as camera imaging systems, internet transmission, etc. Inspired by Wang et al. [16], it becomes evident that a complex degradation process cannot be effectively modeled with a first-order approach. A higher-order modeling process is proposed to simulate the intricate degradation processes found in the real world, denoted by S. The degradation space S is composed of a series of degradation process $D = [D_1, D_2, ..., D_n]$. Specifically, our method incorporates a range of degradation operations, including noise (both Poisson noise and Gaussian noise) [29], various forms of blurring (isotropic, anisotropic, Gaussian filter), resizing (both bicubic and bilinear), and JPEG and video compression. As in Equation (1), let y stand for the original high-resolution (HR) frame, where D represents a set of various degradation operations and X corresponds to the derived low-resolution (LR) frame after undergoing degradation. Depending on the specific application scenarios, customization of the degradation process is feasible, such as the inclusion of additional noise or adjustments in degradation factor parameters.

$$X = D^n(y) = (D_n ... D_2 D_1)(y) \tag{1}$$

In contrast to Real-ESRGAN, we have introduced video compression into the degradation process. This is because video compression implicitly takes into account the spatio-temporal information of video frames, bringing benefits to the degradation. While the second-order degradation process utilized in Real-ESRGAN can generate an extensive degradation space, effectively adapting to videos with diverse degradation levels remains challenging. Hence, we divide the degradation space S into three levels, represented as $[S_1, S_2, S_3]$, by configuring the parameters D accordingly. Among these, S_3 is generated by second-order degradation, while S_1 and S_2 are, respectively, produced by first-order degradation processes with different parameters. As depicted in Figure 2, the S_1 degradation process is visible. The S_3 degradation space corresponds to videos with severe degradation, the S_2 degradation space corresponds to videos with moderate degradation, and the S_1 degradation space corresponds to mild degradation. Therefore, our degradation space S better captures real-world degradation processes.

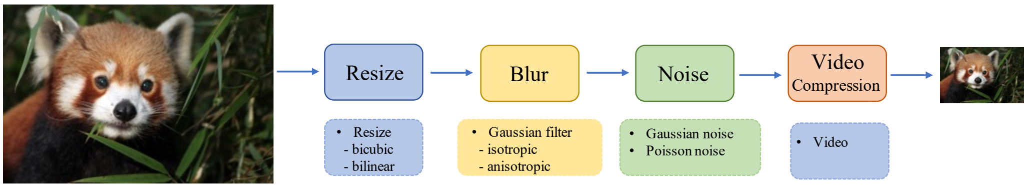

Figure 2. The S_1 degradation process. The S_3 degradation process involves two repeated degradation processes.

3.2. The Image Pre-Cleaning Module

Inspired by Huang et al. [17], the introduction of temporal redundancy can have adverse effects on information propagation. Therefore, removing degradations prior to propagation is crucial, as it aids in suppressing artifacts in the outputs. A simple pre-cleaning module is proposed to suppress degradation; its effectiveness is shown by the experimental results. Our pre-cleaning module consists of 15 cascaded residual blocks. The residual blocks are illustrated in Figure 3.

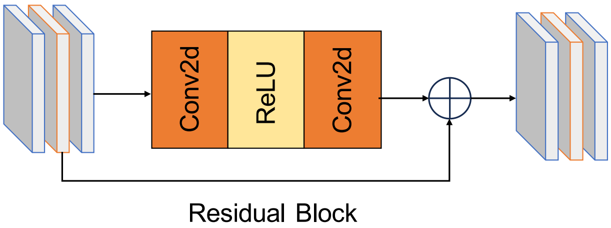

Figure 3. The pre-cleaning module consists of 15 cascaded residual blocks.

The degraded images are passed to the pre-cleaning module to remove redundant temporal information. x_i represents the i-th image of the input sequence, $\hat{x}_i$ is the clean image, and Pre_C is pre-cleaning module, as in Equation (2).

$$\hat{x}_i = Pre_C(x_i) \tag{2}$$

Given BasicVSR's remarkable achievements in video super-resolution (VSR) facilitated by feature propagation, we employ an identical residual block as a pre-cleaning module. Furthermore, to achieve a more lightweight module, we use a streamlined reconstruction module while maintaining performance. Subsequently, the clean images are forwarded to the VSR network for super-resolution. To guide the training of the image pre-cleaning module, we employ the Charbonnier loss [30] to constrain the output of the pre-cleaning module.

$$y_i = S(\{\hat{x}_i\}) \tag{3}$$

3.3. Unsupervised Optical Flow Estimator

The feature alignment module in VSR primarily adopts optical flow-based alignment methods. As analyzed in Section 2, the quality of optical flow significantly impacts the performance of these methods. However, previous supervised flow-based methods are limited by the distinction in the domain between synthetic optical flow and real-world optical flow datasets. Consequently, we introduce an unsupervised optical flow estimator [31]. Our purpose is to construct an optical flow network F_l capable of accurately estimating motion information F_{12}^x from low-quality videos $\{x_1, x_2\}$.

$$F_{12}^x = F_l(x_1, x_2) \tag{4}$$

Based on the fact that the motion information displayed is more accurate between high-quality (HQ) frames $\{y_1, y_2\}$, we employ the optical flow between HQ video frames to guide the training of the optical flow estimator F_l. Firstly, we train a teacher optical flow estimator F_t on the high-quality frames. Once this optical flow estimator F_t converges, it is then frozen and employed as the teacher network in the subsequent phase. F^y_{12} represents the optical flow field for two consecutive frames in the HQ videos. As previously mentioned,

$$F^y_{12} = F_t(y_1, y_2) \tag{5}$$

and the optical flow information between successive frames of HQ frames is more precise. We employ the more precise F^y_{12} as a pseudo-label for the flows F^x_{12} and introduce the concept of the distillation loss: the upsampling operation Up ensures that F^x_{12} matches the dimensions of F^y_{12} in the VSR task.

$$L_{dis}(F^x_{12}, F^y_{12}) = \sum \mid F^y_{12} - Up(F^x_{12}) \mid \tag{6}$$

3.4. Lightweight Reconstruction Module

In the realm of non-blind VSR, BasicVSR exhibits remarkable performance in the feature alignment module. Inspired by this, to strike a balance between computational complexity and performance in DAVSR, we aim to make its feature alignment module lightweight while maintaining an unchanged performance. Neural networks are composed of blocks, and the feature alignment module within BasicVSR conforms to this architectural convention. Each block is composed of common components, i.e., 3×3 convolution, ReLU, and shortcut. To maintain a simple structure, our guiding principle is not to introduce entities unless necessary. ReLU, the activation function employed in the BasicVSR block, is widely adopted in the domain of computer vision. However, recent state-of-the-art (SOTA) methods mostly adopt gated linear units (GLUs) as the activation function. Given their advantages in the low vision field, we contemplated replacing ReLU with a GLU. The formulation for gated linear units is expressed as follows, where X represents the input feature map, σ is a non-linear activation function, f and g are linear transformers, and $\odot$ indicates element-wise multiplication:

$$Gated(X, \sigma, f, g) = f(X) \odot \sigma(g(X)) \tag{7}$$

Although incorporating GLUs into the baseline may enhance the performance, it concurrently raises the block complexity, contradicting our original purpose. To solve this, we optimize the function structure, specifically adopting the GELU:

$$GELU(x) = x\phi(x) \tag{8}$$

From Equations (7) and (8), it is evident that the GELU is a specific instance of a GLU. f, g can be regarded as the same function as input and output and σ is regarded as ϕ. Drawing a parallel, we hypothesize from a different standpoint that GLUs could be seen as a generalization of activation functions, potentially serving as a replacement for nonlinear activation functions. Even if σ is removed, $Gated(X) = f(X) \odot g(X)$ inherently contains nonlinearity. Therefore, we believe that directly partitioning the feature maps along the channel dimension and multiplying them is equivalent to a GLU. This formula is expressed in Equation (9), offering a simplicity that exceeds the implementation of the GELU.

$$SimpleGate(X, Y) = X \odot Y \tag{9}$$

Considering the recent popularity of transformer architectures within the realm of computer vision, the attention mechanism has been widely employed in recent SOTA methods. We incorporate channel attention into the reconstruction block as it enables capturing global information and exhibits efficient computation. A simplified channel attention mechanism can be expressed as follows, where X is an input feature map, *pool* represents the global average pooling operation, and *mlp* is a fully connected layer. The global pooling layer conducts dimension compression operations, while *mlp* is used for calculating the channel attention. In comparison to the block structure employed in BasicVSR, our proposed block structure is shown in Figure 4. The forthcoming ablation experiments will confirm that, even with a more lightweight design, our model achieves a comparable performance.

$$\mathrm{CA}(X) = \sigma(mlp(pool(X))) \cdot X \tag{10}$$

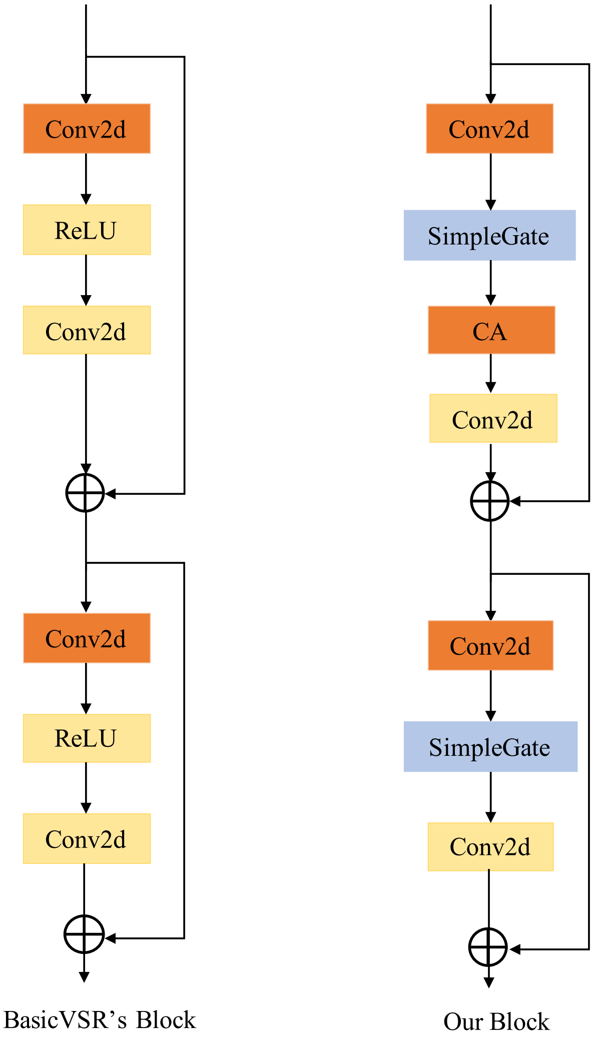

Figure 4. BasicVSR's block contains the most common elements. Our reconstruction module replaces ReLU with SimpleGate, eliminating the presence of a non-linear activation function, and introduces a channel attention mechanism.

4. Experimental Results

In this section, we compare the proposed method with the state of the art both quantitatively and qualitatively, including image models ESRGAN [32], TDAN [7], IKC [33], and Real-ESGRAN [16] and video models RealVSR [34] and DBVSR [15]. The experimental details are outlined below.

4.1. Testing Datasets

Existing non-blind VSR methods are assessed on specific synthetic datasets, which cannot reflect the performance of the real-world video resolution enhancement. The testing datasets evaluated by some blind super-resolution methods are only specifically designed synthetic data. For instance, IKC is assessed by synthetic LR images that are blurred and downsampled. As far as we know, a comprehensive low-resolution video dataset, encompassing noise degradations and diverse blurs, is currently unavailable. For quantitative evaluation, we synthesized LR–HR pairs by subjecting the validation datasets to three levels of degradation in vid4 [35] and vimeo-90k [36]. An illustration depicting video frames with different types of degradations is shown in Figure 5. Specifically, the three degradation types for vid4 and vimeo-90k include (1) Type I: a first-order degradation model equipped with blur, noise with Gaussian σ [1, 10], and video compression. It corresponds to slightly corrupted frames. (2) Type II: a first-order degradation model equipped with blur, noise with Gaussian σ [1, 20], resize, and video compression. It corresponds to moderately corrupted frames. (3) Type III: a two-order degradation model, a repeated first-order degradation model, which corresponds to real-world badly corrupted frames. Although the proposed three-level degradation model is not flawless, it strives to cover the entire degradation space of the real world. We contend that the performance of testing datasets can reflect the advantages of our method.

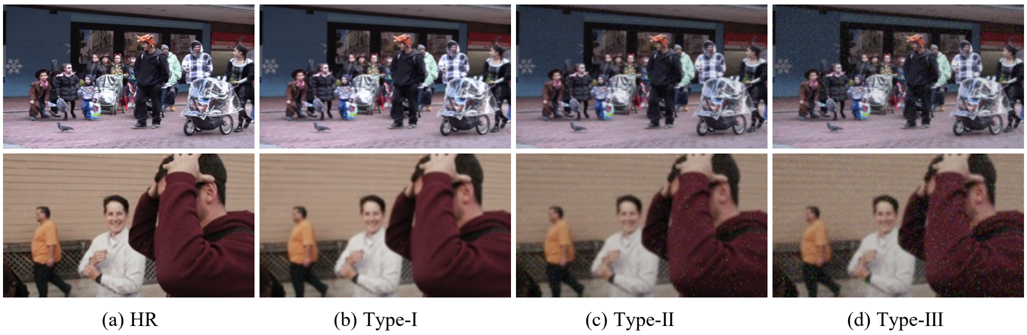
(a) HR (b) Type-I (c) Type-II (d) Type-III

Figure 5. Images with different levels of degradation in vid4 and vimeo-90k.

4.2. Training Details and Quantitative Metric

All experiments were performed within the PyTorch framework, with training facilitated by NVIDIA 3090 GPUs. Following previous works [15], we employed the REDS dataset [37] for training. For efficiency, we employed the lightweight SR network as our backbone and adaptive degradation processes for simplicity and effectiveness. We adopted the Adam [38] optimizer to train the network. The training process is divided into two stages: Initially, we trained the network using the distillation loss of optical flow and output loss for 300K iterations with a learning rate of 10^{-4}. After training the network, we fine-tuned the network with the perceptual loss $L_{perceptual}$ and the adversarial loss $L_{adversarial}$ on different levels of degradation for 200K iterations. The generator was assigned a learning rate of 5×10^{-5}, while the discriminator was set a learning rate of 10^{-4}. Following common practice, we utilized the widely adopted the PSNR and SSIM as the evaluation metrics. The PSNR (peak signal-to-noise ratio, where higher values indicate a better quality) is commonly used to measure the quality of reconstructed images after compression, typically represented in dB [39]. The SSIM (structural similarity index, where higher values indicate a better quality) is commonly employed to measure the similarity between two images, with its values ranging between 0 and 1 [40].

4.3. Comparison to the State of the Art

4.3.1. Quantitative Comparison

Due to the fact that VSR methods demonstrate superior performance exclusively on degraded video frames through bicubic interpolation, and despite the increased complexity in the degradation space of recent blind VSR methods, their performance remains constrained when applied to video frames with varying degrees of degradation in real-world scenarios. To demonstrate the superiority of our degradation-adaptive model, we conduct experimental comparisons on datasets with different levels of degradations. Type III degradation corresponds to videos with severe degradation, type II degradation corresponds to videos with moderate degradation, and type I degradation corresponds to videos with mild degradation. Moreover, these three types of degraded video frames can effectively represent a diverse range of damaged video frames encountered in the real world. Specifically, Table 1 illustrates the restoration performance of various methods under different degradation types, assessed through the evaluative metrics of the PSNR and SSIM. As shown in Table 1, current methods demonstrated superior performance solely under specific degradation conditions, while exhibiting shortcomings in alternative scenarios. In comparison to other methods, our method exhibits q stable and superior performance under three types of degraded images. Meanwhile, as shown in Table 2, a comparison of the inference speed for different networks was conducted. It can be observed that the size of our model (DAVSR) is 36.8% of DBVSR. However, our network's inference speed is slightly faster compared to this. Moreover, the inference speed is five times faster than RealVSR. Therefore, our network processing speed has an advantage compared to the other two blind video super-resolution algorithms.

Specifically, ESRGAN and RealVSR exhibit a favorable performance on bicubic-downsampled datasets, while experiencing a notable decline in performance on other degraded video frames. For instance, ESRGAN exhibits a performance enhancement of 12% when applied to bicubic-degraded data in comparison to its performance on type I and type II datasets. Even as a blind VSR method, RealVSR exhibits a 9% performance decline on datasets of type III. Recent blind super-resolution methods have shown remarkable restoration effects on severely damaged images. However, a decline in performance is observed when confronted with 'type I' and 'type II' degradation, involving mild noise and blurring. Our proposed method, DAVSR, maintains a consistent performance on datasets with different degradation types, exhibiting only a performance decline for type III datasets compared to Real-ESRGAN.

Table 1. Comparison of different methods' quantitative performance on vid4 datasets with different degradations. The optimal results are indicated in bold. 'Types I', 'II', and 'III' are employed to signify datasets characterized by mild, medium, and severe degradations, respectively. 'Bicubic' denotes the vid4 dataset with bicubic interpolation.

D-Type	ESRGAN	TDAN	IKC	RealVSR	DBVSR	Real-ESRGAN	DAVSR
Bicubic	25.03/0.721	26.16/0.782	24.89/0.712	26.65/0.795	26.87/0.805	26.93/0.809	**27.04/0.816**
Type-I	20.36/0.522	26.20/0.783	24.72/0.703	26.51/0.784	26.74/0.798	26.84/0.803	**26.95/0.810**
Type-II	21.98/0.587	25.87/0.770	24.55/0.696	26.44/0.779	26.66/0.796	26.78/0.801	**26.83/0.803**
Type-III	23.21/0.604	23.14/0.602	23.08/0.597	24.12/0.652	24.27/0.658	**25.45/0.751**	25.32/0.744

Table 2. Comparison of different methods' inference speeds with an input size of 180 × 320.

Method	TDAN	RealVSR	DBVSR	Real-ESRGAN	DAVSR
Params (M)	1.97	2.7	25.5	16.7	9.4
Runtime (ms)	138	1082	239	149	216

4.3.2. Qualitative Comparison

Figure 6 provides a comparative analysis of the qualitative aspects of different methods applied to video frames exhibiting various degradations. Specifically, both RealVSR and Real-ESRGAN fail to generate a satisfactory texture and realistic details on video frames degraded by bicubic downsampling. From the perspective of the restoration results, there is no difference between blind video super-resolution and non-blind video super-resolution networks. This is attributed to the fact that the training data for non-blind super-resolution underwent degradation through bicubic downsampling, resulting in an excellent performance on low-quality video sequences subjected to the same degradation. Conversely, the Real-ESRGAN method was trained on severely degraded datasets. Consequently, when applied to restore mildly damaged video frames, the network tends to amplify noise, resulting in a simultaneous reduction in the accuracy of detail generation. In contrast, our DAVSR is trained on datasets with different degradation types. Therefore, it exhibits a stable performance in restoring video frames with different degrees of degradation. As illustrated in Figure 6(1), our DAVSR achieves restoration closer to the high-resolution (HR) frames and exhibits richer details. (For instance, the outline of the eaves appears more distinct and natural, while the font of the store is sharper and more refined.) Similarly, our network demonstrates a superior restoration performance in video frames with moderate and severe degradation, e.g., in Figure 6(4), the sharp edge of the building window lines.

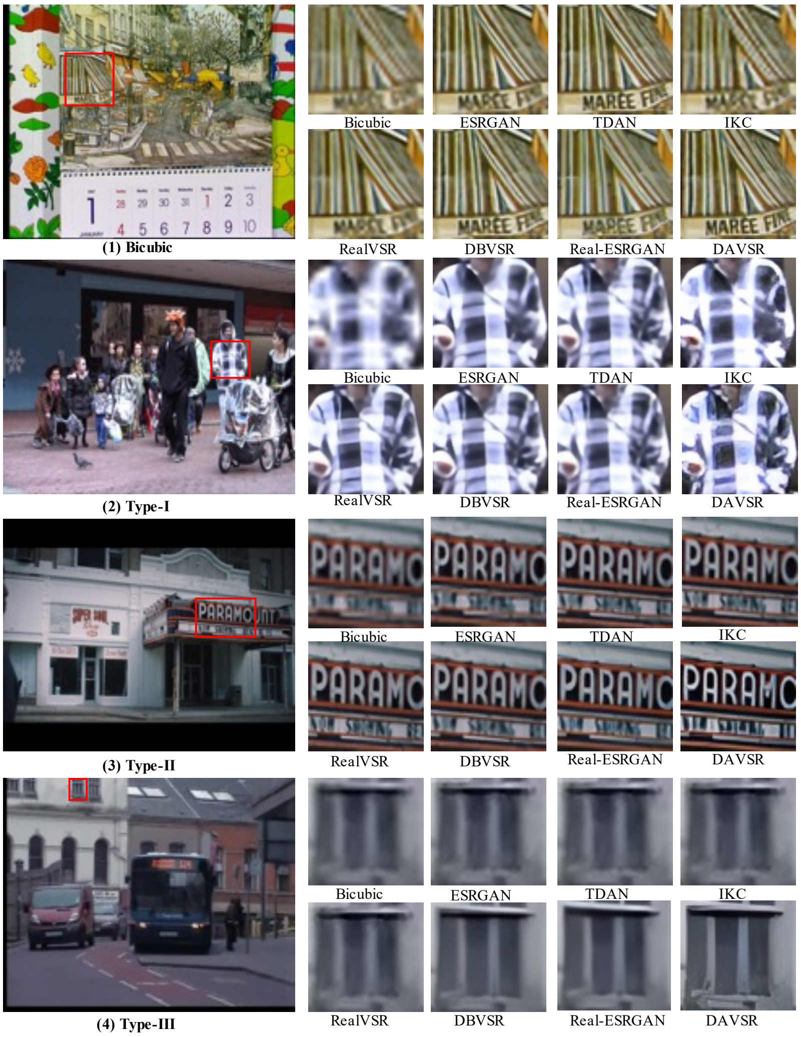

Figure 6. A qualitative comparison of different methods on video frames with diverse degradations. The proposed DAVSR demonstrates superior performance in recovering intricate image structures and achieving enhanced visual quality.

4.4. Effectiveness of Pre-Cleaning Strategy

We conducted experiments to validate the effectiveness of our pre-processing strategy. Initially, the network was trained with the inclusion of the pre-cleaning model. The degraded video frame sequences underwent optical flow alignment and reconstruction directly without passing through the pre-cleaning module, resulting in temporal redundancy being introduced by the degradation process. Examples are depicted in Figure 7. The network without a cleaning strategy lacks sufficient suppression of artifacts and noise, resulting in perceptually impactful noise in the recovered images. This is attributed to the network without a cleaning strategy, which fails to eliminate redundant temporal information, consequently amplifying noise and artifacts. Consequently, it yields a comparatively diminished efficacy in restoration compared to the cleaning network.

Furthermore, we kept the cleaning module and explored the influence of the number of cleaning iterations on the removal of temporal information. The restoration results for images with degradation type I are shown in Figure 8. On the one hand, when the cleaning module is applied only once, the noise is not removed and blurriness persists. On the other hand, employing the cleaning module four times results in excessive noise removal, leading to an unrealistic over-smoothing of images. We observe that for mildly degraded images, applying the cleaning module twice is optimal. As anticipated, for images representing degradation type III, characterized by severe degradation, the application of the pre-cleaning module twice is insufficient. Our observations indicate that, in most scenarios, a maximum of three iterations suffices. We ascertain that for images exhibiting degradation types I and II, a two-fold application of the pre-processing module will suffice. Conversely, for images representing degradation type III, a three-fold application of the module is deemed adequate.

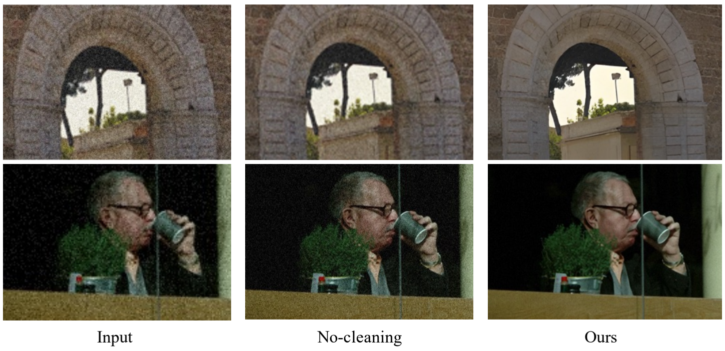

Figure 7. Exploring the effectiveness of the pre-cleaning strategy. The cleaning module plays a crucial role in suppressing artifacts and noise, which is essential for subsequent information propagation in degraded low-resolution images.

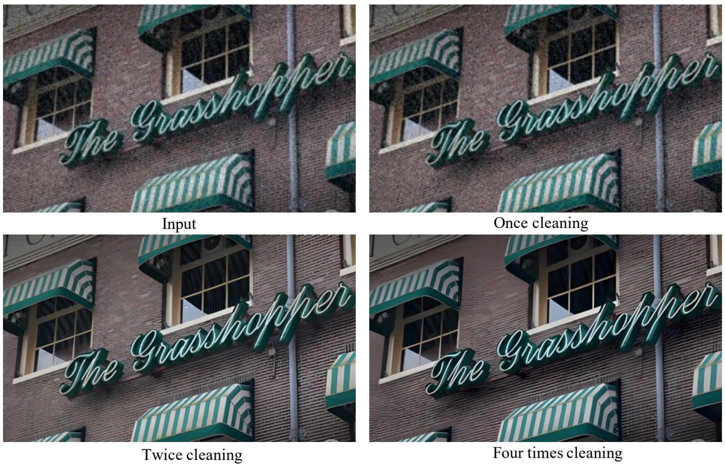

Figure 8. Exploring the impact of the number of pre-cleaning module iterations. For mildly degraded images, our network achieves optimal results when applying the cleaning strategy twice.

4.5. Ablation Studies

4.5.1. Unsupervised Optical Flow Estimator

To validate the effectiveness of our unsupervised optical flow estimator, we retrained a supervised optical flow estimator, specifically SpyNet. This is a widely used optical flow network in the state of the art (SOTA), such as BasicVSR. As shown in Table 3, SpyNet trained with an unsupervised approach can outperform the supervised counterpart by 0.13 dB on the vid4 dataset. Furthermore, on the vimeo-90k dataset, the performance surpasses

0.15 dB. The experiments conducted on various datasets suggest that the unsupervised optical flow alignment method achieves better frame alignment results. Additionally, this improvement is attributed to the utilization of optical flow between high-quality video frames to guide the training of our optical flow estimator.

Table 3. A comparative analysis of unsupervised and supervised training approaches for SpyNet networks in video super-resolution. "Sup" represents supervised training; "Unsup" represents unsupervised training.

Dataset	**Vid4**		**Vimeo-90k**	
Method	Sup	Unsup	Sup	Unsup
PSNR(db)	26.91	27.04	34.78	34.93
SSIM	0.804	0.816	0.929	0.937

4.5.2. Lightweight Reconstruction Module

From BasivVSR's Block to our Block: Our reconstruction module is composed of concatenating multiple blocks. A comparison of the reconstruction block between BasicVSR and our DAVSR is illustrated in Figure 4. Notably, we have introduced a channel attention mechanism and replaced the ReLU activation with a simplified SimpleGate in our block structure. Ablation studies were conducted on the vid4 and vimeo-90k datasets. We employed the training approach mentioned in Section 4.2 for training. The effectiveness of channel attention and SimpleGate is demonstrated in Table 4. A performance improvement is observed when either replacing ReLU in the reconstruction module with SimpleGate (SG) or adding channel attention (CA) to the reconstruction module for both vid4 and vimeo-90k datasets.

Table 4. The effectiveness of channel attention (CA) and SimpleGate (SG) has been validated. In the table, Original Block refers to the reconstruction module of BasicVSR.

	ReLU → SG	**CA**	**Vid4** **PSNR SSIM**	**Vimeo-90k** **PSNR SSIM**
Original Block			26.82/0.803	33.74/0.925
↓	✓		26.98/0.812	34.56/0.931
		✓	26.92/0.808	34.37/0.929
Our Block	✓	✓	27.04/0.816	34.93/0.937

Number of Blocks: Due to the fact that our reconstruction module is formed by concatenating multiple blocks, we investigated the effect of the number of blocks in the reconstruction module on its restoration effectiveness in Table 5. To ensure the lower bounds of the restoration performance within the reconstruction module, we designed an experimental setup with the number of blocks ranging from 10 to 30. It is evident from our experiments that augmenting the number of blocks to 30 does not yield a substantial enhancement in network performance. On the contrary, the model params increase by 50%. Clearly, the benefits of increasing the number of blocks are limited. Therefore, we set the number of blocks in the DAVSR network to 20, reaching the optimal performance for the reconstruction module.

Table 5. The impact of residual block quantity on the performance of the reconstruction module.

	Number of Blocks	**Params (M)**	**Vid4** **PSNR SSIM**	**Vimeo-90k** **PSNR SSIM**
	10	2.82	26.73/0.798	33.25/0.916
DAVSR	20	5.64	27.04/0.816	34.93/0.937
	30	8.46	27.05/0.816	35.05/0.938

5. Conclusions

In this paper, we propose degradation-adaptive blind video super-resolution, namely DVASR, to handle various types of degraded video frames in the real world. In order to capture various degradation processes in the real world, we propose a degradation-adaptive model for modeling. The proposed DAVSR exhibits favorable restoration when applied to video frames with different levels of degradation. Moreover, we utilize a pre-cleaning strategy to mitigate the adverse effects stemming from temporal redundancy between video frames. Meanwhile, we employ a lightweight reconstruction module to balance the computational complexity and the performance of DAVSR. Extensive experiments on several benchmark datasets achieve a state-of-the-art SR performance.

Due to the significantly higher computational complexity of super-resolution networks compared to image classification and object detection networks, in the future, we will explore more lightweight architectures and real-time super-resolution networks.

Author Contributions: Conceptualization, P.Z. and M.L.; methodology, M.L.; software, M.L.; validation, P.Z. and M.L.; formal analysis, P.Z. and M.L.; writing—original draft preparation, M.L.; writing—review and editing, P.Z. and M.L.; funding acquisition, P.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Shenzhen Science and Technology Program (Grant No. KQTD20190929172704911) and the Science and Technology Planning Project of Guangdong Science and Technology Department under Grant Guangdong Key Laboratory of Advanced IntelliSense Technology (2019B121203006).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available from the authors upon reasonable request.

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

References

1. Abad, A.C.; Ranasinghe, A. Visuotactile sensors with emphasis on gelsight sensor: A review. *IEEE Sensors J.* **2020**, *20*, 7628–7638. [CrossRef]
2. Kollolu, R. A Review on wide variety and heterogeneity of iot platforms. *Int. J. Anal. Exp. Modal Anal. Anal.* **2020**, *12*, 3753–3760. [CrossRef]
3. Ashraf, S.; Saleem, S.; Ahmed, T.; Aslam, Z.; Shuaeeb, M. Iris and foot based sustainable biometric identification approach. In Proceedings of the 2020 International Conference on Software, Telecommunications and Computer Networks (SoftCOM), Split, Croatia, 17–19 September 2020; IEEE: Toulouse, France, 2020; pp. 1–6.
4. 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.
5. 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; Springer: Berlin/Heidelberg, Germany, 2020; pp. 645–660.
6. Chan, K.C.; Wang, X.; Yu, K.; Dong, C.; Loy, C.C. Basicvsr: The search for essential components in video super-resolution and beyond. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 4947–4956.
7. Tian, Y.; Zhang, Y.; Fu, Y.; Xu, C. Tdan: Temporally-deformable alignment network for video super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3360–3369.
8. Chan, K.C.; Zhou, S.; Xu, X.; Loy, C.C. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5972–5981.
9. Tassano, M.; Delon, J.; Veit, T. Fastdvdnet: Towards real-time deep video denoising without flow estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 1354–1363.
10. Maggioni, M.; Huang, Y.; Li, C.; Xiao, S.; Fu, Z.; Song, F. Efficient multi-stage video denoising with recurrent spatio-temporal fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 3466–3475.

11. Sheth, D.Y.; Mohan, S.; Vincent, J.L.; Manzorro, R.; Crozier, P.A.; Khapra, M.M.; Simoncelli, E.P.; Fernandez-Granda, C. Unsupervised deep video denoising. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 1759–1768.
12. 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]
13. Wang, L.; Guo, Y.; Lin, Z.; Deng, X.; An, W. Learning for video super-resolution through HR optical flow estimation. In Proceedings of the Computer Vision—ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, 2–6 December 2018; Revised Selected Papers, Part I 14; Springer: Berlin/Heidelberg, Germany, 2019; pp. 514–529.
14. Liu, A.; Liu, Y.; Gu, J.; Qiao, Y.; Dong, C. Blind image super-resolution: A survey and beyond. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 5461–5480. [CrossRef] [PubMed]
15. Pan, J.; Bai, H.; Dong, J.; Zhang, J.; Tang, J. Deep blind video super-resolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 4811–4820.
16. Wang, X.; Xie, L.; Dong, C.; Shan, Y. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 1905–1914.
17. Huang, Y.; Dong, H.; Pan, J.; Zhu, C.; Liang, B.; Guo, Y.; Liu, D.; Fu, L.; Wang, F. Boosting Video Super Resolution with Patch-Based Temporal Redundancy Optimization. In Proceedings of the International Conference on Artificial Neural Networks, Heraklion, Greece, 26–29 September 2023; Springer: Berlin/Heidelberg, Germany, 2023; pp. 362–375.
18. Kim, G.; Park, J.; Lee, K.; Lee, J.; Min, J.; Lee, B.; Han, D.K.; Ko, H. Unsupervised real-world super resolution with cycle generative adversarial network and domain discriminator. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 456–457.
19. Maeda, S. Unpaired image super-resolution using pseudo-supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 291–300.
20. Rad, M.S.; Yu, T.; Musat, C.; Ekenel, H.K.; Bozorgtabar, B.; Thiran, J.P. Benefiting from bicubically down-sampled images for learning real-world image super-resolution. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2021; pp. 1590–1599.
21. Lee, S.; Choi, M.; Lee, K.M. Dynavsr: Dynamic adaptive blind video super-resolution. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2021; pp. 2093–2102.
22. Xiao, Y.; Yuan, Q.; Jiang, K.; He, J.; Wang, Y.; Zhang, L. From degrade to upgrade: Learning a self-supervised degradation guided adaptive network for blind remote sensing image super-resolution. *Inf. Fusion* **2023**, *96*, 297–311. [CrossRef]
23. Yu, J.J.; Harley, A.W.; Derpanis, K.G. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In Proceedings of the Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, 8–10. 15–16 October 2016; Proceedings, Part III 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 3–10.
24. Wang, Y.; Yang, Y.; Yang, Z.; Zhao, L.; Wang, P.; Xu, W. Occlusion aware unsupervised learning of optical flow. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4884–4893.
25. Liu, L.; Zhang, J.; He, R.; Liu, Y.; Wang, Y.; Tai, Y.; Luo, D.; Wang, C.; Li, J.; Huang, F. Learning by analogy: Reliable supervision from transformations for unsupervised optical flow estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6489–6498.
26. Lugmayr, A.; Danelljan, M.; Timofte, R. Unsupervised learning for real-world super-resolution. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Republic of Korea, 27–28 October 2019; IEEE: Toulouse, France, 2019; pp. 3408–3416.
27. Yuan, Y.; Liu, S.; Zhang, J.; Zhang, Y.; Dong, C.; Lin, L. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 701–710.
28. 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]
29. Huang, Y.; Li, S.; Wang, L.; Tan, T. Unfolding the alternating optimization for blind super resolution. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 5632–5643.
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; IEEE: Toulouse, France, 1994; Volume 2, pp. 168–172.
31. Lin, J.; Hu, X.; Cai, Y.; Wang, H.; Yan, Y.; Zou, X.; Zhang, Y.; Van Gool, L. Unsupervised flow-aligned sequence-to-sequence learning for video restoration. In Proceedings of the International Conference on Machine Learning, PMLR, Grenoble, France, 19–23 September 2022; pp. 13394–13404.
32. Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; Change Loy, C. Esrgan: Enhanced super-resolution generative adversarial networks. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018.
33. Gu, J.; Lu, H.; Zuo, W.; Dong, C. Blind super-resolution with iterative kernel correction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 5–20 June 2019; pp. 1604–1613.

34. Yang, X.; Xiang, W.; Zeng, H.; Zhang, L. Real-world video super-resolution: A benchmark dataset and a decomposition based learning scheme. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 4781–4790.
35. Liu, C.; Sun, D. On Bayesian adaptive video super resolution. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *36*, 346–360. [CrossRef] [PubMed]
36. 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]
37. 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.
38. Kinga, D.; Adam, J.B. A method for stochastic optimization. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 22 December 2015; Volume 5, p. 6.
39. Hore, 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; IEEE: Toulouse, France, 2010; pp. 2366–2369.
40. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef]

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.