

Article

LIF-VSR: A Lightweight Framework for Video Super-Resolution with Implicit Alignment and Attentional Fusion

Songyi Zhang ^{1,2,3}, Hailin Zhang ¹, Xiaolin Wang ², Kailei Song ^{2,3}, Zhizhuo Han ^{2,3} , Zhitao Zhang ² and Wenchi Cheng ^{1,*}

¹ School of Telecommunications Engineering, Xidian University, Xi'an 710071, China; zhangsongyi@stu.xidian.edu.cn (S.Z.); hlzhang@xidian.edu.cn (H.Z.)

² Hebei Far East Communication Engineering Co., Ltd., Shijiazhuang 050200, China; wangxiaolin@hbfec.com.cn (X.W.); songkailei@hbfec.com.cn (K.S.); hanzhizhuo@hbfec.com.cn (Z.H.); zhangzhitao@hbfec.com.cn (Z.Z.)

³ Hebei Private Network Communication Technology Innovation Center, Shijiazhuang 050200, China

* Correspondence: wccheng@xidian.edu.cn

Abstract

Video super-resolution (VSR) has advanced rapidly in enhancing video quality and restoring compressed content, yet leading methods often remain too costly for real-world use. We present LIF-VSR, a lightweight, near-real-time framework built with an efficiency-first philosophy, comprising economical temporal propagation, a new neighboring-frame fusion strategy, and three streamlined core modules. For temporal propagation, a uni-directional recurrent architecture transfers context through a compact inter-frame memory unit, avoiding the heavy compute and memory of multi-frame parallel inputs. For fusion and alignment, we discard 3D convolutions and optical flow, instead using (i) a deformable convolution module for implicit feature-space alignment, and (ii) a sparse attention fusion module that aggregates adjacent-frame information via learned sparse key sampling points, sidestepping dense global computation. For feature enhancement, a cross-attention mechanism selectively calibrates temporal features at far lower cost than global self-attention. Across public benchmarks, LIF-VSR achieves competitive results with only 3.06 M parameters and a very low computational footprint, reaching 27.65 dB on Vid4 and 31.61 dB on SPMCs.

Keywords: video super-resolution; lightweight model; implicit deformable alignment; sparse attention fusion



Academic Editor: Zhe-Ming Lu

Received: 20 October 2025

Revised: 5 December 2025

Accepted: 13 December 2025

Published: 17 January 2026

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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The demand for high-quality, high-resolution video content is steadily increasing in applications such as surveillance, medical imaging, and mobile multimedia. Video super-resolution (VSR), which reconstructs a high-resolution sequence from a low-resolution one, plays a crucial role in these domains. Unlike Single-Image Super-Resolution (SISR) [1–3], which processes frames independently, VSR must simultaneously address cross-frame motion, alignment, and temporal fusion. Naively applying SISR frame-by-frame ignores temporal cues, resulting in temporal inconsistency and artifact accumulation. Consequently, the effective integration of alignment and fusion remains a central challenge in VSR [4–6].

To utilize temporal information effectively, existing VSR methods typically rely on either explicit motion estimation (e.g., optical flow) or implicit alignment (e.g., deformable convolution). However, flow-based approaches often incur substantial computational

costs and are prone to artifacts when estimation fails in occluded regions. On the other hand, while recurrent frameworks have shown promise for efficient temporal propagation, current state-of-the-art designs often suffer from heavy memory pressure due to multi-frame parallel processing or quadratic complexity arising from global attention mechanisms. These computational bottlenecks hinder the practical deployment of VSR models on resource-constrained devices.

To address these challenges, we present LIF-VSR, a lightweight framework tailored for near real-time deployment. Our design centers on *efficient temporal propagation*, *sparse information aggregation*, and *low-cost feature enhancement*. Specifically, we propose a unidirectional recurrent architecture that maintains a minimal state, avoiding the overhead of multi-frame parallel inputs. For alignment and fusion, we introduce Implicit Deformable Alignment (IDA) and Sparse Attention Fusion (SAF). IDA performs cross-frame alignment without optical-flow priors by adaptively regressing sparse sampling offsets. SAF efficiently aggregates temporal cues using a small set of learnable key sampling points, transforming the quadratic complexity of global attention into a near-linear cost. Furthermore, a Temporal Cross Calibration (TCC) mechanism targets and calibrates features via cross-attention, where shallow features of the center frame serve as queries to enhance fused deep features.

Extensive experiments show that LIF-VSR achieves state-of-the-art performance on benchmarks such as Vid4 and SPMCs. Crucially, with only 3.06 M parameters and 17.60 GFLOPs, our model reduces the parameter count by approximately 50% compared to BasicVSR, offering a superior accuracy–efficiency trade-off.

We summarize our key contributions in three main points:

- We propose a minimal-state, unidirectional recurrent propagation scheme coupled with sparse neighboring-frame fusion, which effectively models temporal context without multi-frame parallelism, significantly reducing memory and compute.
- We introduce implicit deformable alignment and sparse attention fusion that regress a small set of sampling offsets from feature correlations to achieve cross-frame alignment and information aggregation without optical-flow priors.
- We adopt an RCAN-based lightweight backbone and propose a cross-attention-driven Temporal Cross Calibration mechanism that uses the center frame as a query to directionally align and enhance temporal features, substantially improving high-frequency detail.

2. Related Work

The goal of video super-resolution is to restore a high-resolution sequence from a low-resolution video by leveraging temporal information across frames—a key advantage over single-image super-resolution (SISR). This approach enables the recovery of more intricate details and ensures temporal smoothness. This section begins with a concise overview of the advancements in SISR, followed by a detailed examination of the core components of VSR, including its alignment strategies, network frameworks, and fusion/enhancement mechanisms.

2.1. Single-Image Super-Resolution

As a classic yet continually evolving research topic, Single-Image Super-Resolution is central to the field of computer vision. Beyond early interpolation and reconstruction methods, deep learning ushered in rapid progress: from the pioneering SRCNN [1] and deeper networks such as VDSR [7] to residual/attention-based designs like EDSR [2] and RCAN [3], SISR performance (e.g., PSNR) has been continuously advanced. More recently, Transformer-based architectures (e.g., SwinIR [8]) further improved long-range dependency modeling and feature representation. However, naively applying SISR frame-by-frame

discards temporal cues, leading to flickering and artifact accumulation in the temporal dimension. Consequently, dedicated VSR models have been developed to explicitly leverage temporal information. Many state-of-the-art VSR systems still draw on successful SISR backbones, in particular, our model adopts an RCAN-style residual channel attention backbone for lightweight feature extraction.

2.2. Video Super-Resolution

The primary difficulty in VSR lies in effectively managing motion across frames to properly fuse their complementary information. Existing methods developed to address this issue can be broadly categorized as follows.

2.2.1. Alignment-Based VSR

To utilize neighboring frames, one must first compensate for pixel displacements caused by object and camera motion.

Explicit alignment. A prevalent strategy is to first compute the optical flow between adjacent frames and then utilize this motion information to warp features towards the center frame for alignment. Representative methods include VESPCN [9], TOFlow [10], SOF-VSR [11] and EDVR [4], which often rely on pretrained flow estimators such as SPyNet [12] or modern variants (e.g., PWC-Net [13], RAFT [14]). While explicit motion compensation is intuitive and can handle large displacements, its effectiveness hinges on flow accuracy; flow errors lead to warping artifacts and blur. Moreover, accurate flow estimation is computationally heavy and may hinder real-time deployment.

Implicit alignment. To circumvent these issues, implicit, end-to-end alignment in feature space has gained traction. Deformable convolutions (DCN) [15,16] are a mainstream solution: TDAN [17] and BasicVSR++ [6] show that learning sparse sampling offsets enables flexible, content-adaptive aggregation for alignment, offering lower cost and better robustness to complex, non-rigid motion. Deformable 3D convolution extends offset learning to the spatiotemporal domain, enabling content-adaptive sampling across space–time [18], while lightweight kernelizations cut FLOPs and parameters—together boosting efficiency without degrading alignment fidelity [19]. Our Implicit Deformable Alignment (IDA) follows this efficient paradigm, targeting precise alignment under a lightweight budget.

2.2.2. Recurrent VSR Frameworks

Based on their temporal processing methodology, VSR frameworks are typically divided into two main paradigms: sliding-window and recurrent. The sliding-window paradigm operates on fixed-size groups of frames (e.g., 5 or 7) simultaneously. While this approach is adept at capturing local temporal relationships, it suffers from poor scalability in terms of memory and computational cost when applied to longer video sequences. Recurrent frameworks address long-range temporal modeling more efficiently by processing frames sequentially and propagating a hidden state. FRVSR [20], BasicVSR [5], and BasicVSR++ [6] are representative: they aggregate past information via recurrent feature propagation, enabling constant memory/compute regardless of video length—well-suited to efficient and near real-time applications. In addition, RBPN [21] explores multi-stage back-projection with recurrence. Lightweight designs (FDDCC-VSR [19], LightVSR [22], RASVSR [23]) illustrate that careful architectural bottlenecks can retain accuracy without prohibitive compute. LIF-VSR builds upon this efficient unidirectional recurrent scheme, with our contributions focusing on lightweight and innovative alignment, fusion, and enhancement within the recurrent cell.

2.2.3. Fusion and Enhancement in VSR

Aligned multi-frame features must be fused effectively to yield high-quality reconstructions. Early methods either concatenate features before applying convolutions or use 3D convolutions for joint spatiotemporal modeling; the former is often insufficiently adaptive, whereas the latter is computationally expensive. To more selectively exploit temporal cues, attention mechanisms have been introduced into VSR. While Transformer-based designs leverage global self-attention to effectively model long-range spatiotemporal dependencies, their quadratic complexity with respect to the number of tokens makes this approach computationally prohibitive for high-resolution inputs [24]. Despite their strong global context modeling [24,25], the cost remains substantial. Complementary frequency-domain processing has recently been shown to recover periodic textures and subtle high-frequency details that local convolutions may miss [26]. Motivated by these observations, LIF-VSR emphasizes efficient fusion and enhancement via sparse, cross-attentional designs that mitigate computational bottlenecks while preserving—and often improving—reconstruction quality.

3. Proposed Method

We pursue a lightweight design guided by three principles: *minimal state*, *sample-on-demand*, and *directional calibration*. Concretely, LIF-VSR decomposes video super-resolution into (i) a unidirectional recurrent framework for sequence processing, (ii) temporal propagation via implicit deformable alignment and sparse fusion without optical flow, and (iii) feature enhancement and image reconstruction based on a compact RCAN backbone and cross-attentional calibration. This factorization preserves temporal coherence and high-frequency fidelity under tight parameter and compute budgets. Figure 1 illustrates the overall architecture of LIF-VSR.

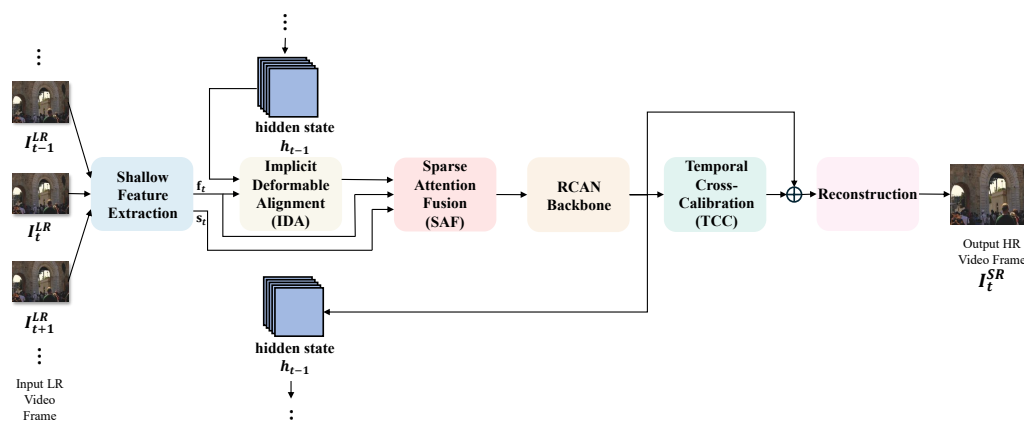


Figure 1. Overall architecture of the LIF-VSR framework.

3.1. Overall Architecture

Given a low-resolution (LR) sequence $\{I_t^{LR}\}_{t=1}^T$, LIF-VSR processes frames sequentially and maintains a compact hidden state $\mathbf{h}_{t-1} \in \mathbb{R}^{C \times H \times W}$ to carry historical context. In contrast to sliding-window parallel schemes, the memory footprint is governed mainly by the spatial and channel dimensions rather than the temporal window length.

At each time step, we first extract shallow features for the current frame $\mathbf{f}_t = \phi_{\text{cur}}(I_t^{LR})$ and build a neighborhood summary $\mathbf{s}_t = \phi_{\text{sup}}(\text{neigh}(I_t^{LR}))$ using mirrored boundary indexing. We then align and fuse historical and neighboring information:

$$\mathbf{h}'_{t-1} = \text{IDA}(\mathbf{h}_{t-1}, \mathbf{f}_t), \quad \mathbf{f}_t^{\text{fus}} = \text{SAF}(\mathbf{f}_t, \mathbf{s}_t, \mathbf{h}'_{t-1}). \quad (1)$$

The fused feature is distilled by a lightweight RCAN backbone and calibrated via cross-attention using the current-frame shallow feature as the query:

$$\mathbf{d}_t = \text{RCAN}(\mathbf{f}_t^{\text{fus}}), \quad \mathbf{z}_t = \text{TCC}(\mathbf{f}_t, \mathbf{d}_t), \quad \mathbf{h}_t = \text{detach}(\mathbf{d}_t). \quad (2)$$

The reconstruction head upsamples by two successive pixel-shuffle stages and adds an LR bypass:

$$I_t^{\text{SR}} = \text{Upsample}(\mathbf{z}_t) + \text{Up}_{\text{bilinear}}(I_t^{\text{LR}}). \quad (3)$$

From a complexity viewpoint, the recurrent paradigm confines the dominant memory and latency terms to single-frame tensors (C, H, W) and avoids the bandwidth pressure of multi-frame stacks and 3D convolutions.

To formalize this workflow, we summarize the entire recurrent inference process in Algorithm 1. The subsequent sections will elaborate on the core components: IDA, SAF, and TCC.

Algorithm 1 LIF-VSR Recurrent Inference Algorithm

Require: Low-resolution video sequence $LQs = \{I_t^{\text{LR}}\}_{t=1}^T$

Ensure: High-resolution video sequence $SRs = \{I_t^{\text{SR}}\}_{t=1}^T$

```

1: Initialize hidden state  $\mathbf{h}_0 \leftarrow \mathbf{0}$ 
2: Initialize an empty output list  $SRs \leftarrow []$ 
3: for  $t \leftarrow 1$  to  $T$  do
4:    $\mathbf{f}_t \leftarrow \phi_{\text{cur}}(I_t^{\text{LR}})$  ▷ Extract shallow features for the current frame
5:   if  $t > 1$  then
6:      $\mathbf{h}'_{t-1} \leftarrow \text{IDA}(\mathbf{h}_{t-1}, \mathbf{f}_t)$  ▷ Align the hidden state from the previous step
7:   else
8:      $\mathbf{h}'_{t-1} \leftarrow \mathbf{h}_0$  ▷ No history for the first frame
9:   end if
10:   $\mathbf{s}_t \leftarrow \phi_{\text{sup}}(\text{neigh}(I_t^{\text{LR}}))$  ▷ Extract features from neighboring frames
11:   $\mathbf{f}_t^{\text{fus}} \leftarrow \text{SAF}(\mathbf{f}_t, \mathbf{s}_t, \mathbf{h}'_{t-1})$  ▷ Fuse current, neighbor, and historical info
12:   $\mathbf{d}_t \leftarrow \text{RCAN}(\mathbf{f}_t^{\text{fus}})$  ▷ Extract deep features via the backbone
13:   $\mathbf{z}_t \leftarrow \text{TCC}(\mathbf{d}_t, \text{query} = \mathbf{f}_t)$  ▷ Calibrate with cross-attention
14:   $I_t^{\text{SR}} \leftarrow \text{Upsample}(\mathbf{z}_t) + \text{Up}_{\text{bilinear}}(I_t^{\text{LR}})$  ▷ Reconstruct the HR image
15:  Append  $I_t^{\text{SR}}$  to the  $SRs$  list
16:   $\mathbf{h}_t \leftarrow \text{detach}(\mathbf{d}_t)$  ▷ Update the hidden state for the next frame
17: end for
18: return  $SRs$ 

```

3.2. Temporal Propagation, Alignment and Fusion

This component aligns the historical state to the current content and fuses temporal cues efficiently, all in the feature domain without optical flow or warping.

3.2.1. Implicit Deformable Alignment (IDA)

To align the historical state $\mathbf{h}_{t-1}$ with the current frame $\mathbf{f}_t$, we propose the Implicit Deformable Alignment (IDA) module (see Figure 2). While previous methods such as EDVR [4] and BasicVSR++ [6] also employ deformable convolutions, they typically rely on *cascaded pyramidal processing* or *explicit optical flow guidance* (e.g., using SPyNet [12]) to predict offsets. These approaches incur significant computational overhead and are prone to instability when optical flow estimation fails in occluded or textureless regions.

In contrast, our IDA is designed for efficiency and stability. It discards the heavy pre-trained flow network and instead learns to regress offsets directly from local feature correlations. Specifically, conditioned on the concatenation $[\mathbf{h}_{t-1}, \mathbf{f}_t]$, a lightweight con-

volutional head predicts the residual sampling offsets $\Delta \mathbf{p}_k$ and modulation masks m_k . The alignment is performed via:

$$\mathbf{y}(\mathbf{p}_0) = \sum_{k=1}^K w_k \left[\mathbf{x}(\mathbf{p}_0 + \mathbf{p}_k + \Delta \mathbf{p}_k) \right] m_k, \quad (4)$$

where $\mathbf{x} = \mathbf{h}_{t-1}$.

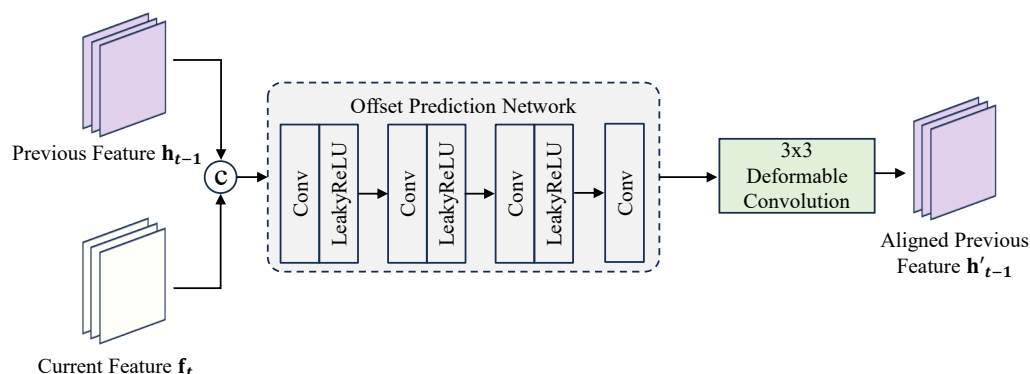


Figure 2. The internal structure of the Implicit Deformable Alignment (IDA) module. Unlike flow-guided approaches, IDA directly regresses offsets from feature correlations, ensuring lightweight and stable alignment.

This design offers three key advantages over prior DCN-based methods:

1. **Direct Feature Regression:** By bypassing explicit flow warping, IDA avoids the error propagation associated with flow estimation artifacts.
2. **Computational Efficiency:** IDA eliminates the need for bulky flow estimators. Specifically, it reduces the parameter count of the alignment module by approximately 70% compared to the standard pre-trained SPyNet [12] used in flow-guided counterparts.
3. **Stability Mechanism:** The offsets are learned end-to-end within the reconstruction objective, allowing the network to suppress sampling in unreliable regions via the modulation masks m_k , thereby providing superior robustness against large motions and occlusions.

3.2.2. Sparse Attention Fusion (SAF)

Dense 3D convolutions incur high memory bandwidth, and global self-attention scales quadratically with spatial positions, which is prohibitive at mid/high resolutions. We reformulate global correspondence as sparse keypoint sampling and weighted aggregation (see Figure 3).

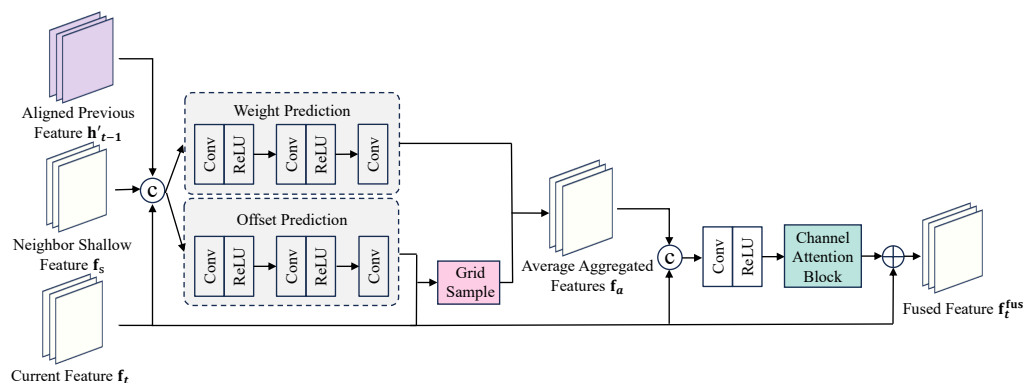


Figure 3. Illustration of the Sparse Attention Fusion (SAF) module. It learns sparse offsets and weights to aggregate multi-source temporal information.

For the current feature and a candidate temporal branch (neighboring or aligned history), an offset head predicts K offsets $\{\Delta \mathbf{p}_k\}_{k=1}^K$ and a weight head yields $\{\alpha_k\}_{k=1}^K$ with Softmax normalization. First, for each temporal branch (e.g., neighbor features or the aligned hidden state), we compute an aggregated feature, denoted as $\bar{\mathbf{a}}$, using a sparse sampling mechanism. At a location $\mathbf{p}_0$, this is calculated as:

$$\bar{\mathbf{a}}(\mathbf{p}_0) = \sum_{k=1}^K \alpha_k \left[\mathbf{g}(\mathbf{p}_0 + \Delta \mathbf{p}_k) \right], \quad (5)$$

where $\mathbf{g}$ is the feature map of the temporal branch being sampled. After computing the aggregated feature $\bar{\mathbf{a}}_i$ for each of the B temporal branches, we concatenate them with the central feature $\mathbf{f}_t$. This combined tensor is subsequently fed into a fusion block $\Gamma(\cdot)$ (a 1×1 convolution followed by channel attention) to produce the final fused feature for the current timestep, $\mathbf{f}_t^{\text{fus}}$:

$$\mathbf{f}_t^{\text{fus}} = \Gamma(\text{Concat}[\mathbf{f}_t, \bar{\mathbf{a}}_1, \dots, \bar{\mathbf{a}}_B]). \quad (6)$$

3.3. Feature Enhancement and Reconstruction

3.3.1. Lightweight RCAN Backbone

We adopt a compact RCAN as our feature extraction backbone. It serves as a standardized baseline, ensuring that performance gains are primarily derived from our proposed temporal alignment and fusion schemes rather than a heavier spatial backbone. It is composed of several residual groups; each group stacks multiple RCABs (see Figure 4). For an input $\mathbf{u}$, the channel rescaling in an RCAB reads

$$\mathbf{s} = \sigma(\mathbf{W}_2 \delta(\mathbf{W}_1 \text{GAP}(\phi(\mathbf{u})))), \quad \text{RCAB}(\mathbf{u}) = \mathbf{u} + \mathbf{s} \odot \phi(\mathbf{u}), \quad (7)$$

where GAP denotes global average pooling, $\delta(\cdot)$ is ReLU, $\sigma(\cdot)$ is Sigmoid, $\phi(\cdot)$ denotes a convolutional transform, and $\odot$ is channel-wise multiplication.

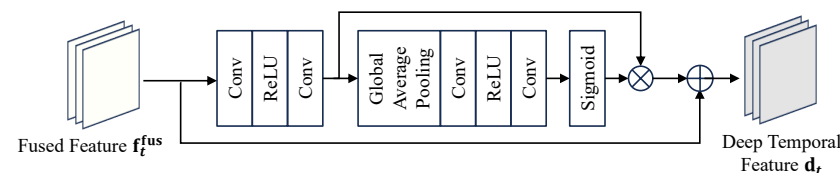


Figure 4. The architecture of the Residual Channel Attention Block (RCAB).

Group-level residual connections stabilize deep optimization. With moderate channels (e.g., $C = 64$), the $O(C^2HW)$ cost remains tractable while effectively boosting texture/structure responses.

3.3.2. Temporal Cross Calibration (TCC)

Deep temporal fusion may dilute high-frequency cues of the current frame, whereas self-attention on $\mathbf{d}_t$ alone is costly. We instead perform cross-attention with the shallow current feature as the query and the deep temporal feature as key/value (see Figure 5).

The TCC module is built upon a scaled dot-product attention mechanism, whose output acts as a calibration signal that is subsequently added back to the original deep feature $\mathbf{d}_t$ via a residual connection to produce the final calibrated feature, denoted as $\mathbf{z}_t$. This process is formulated as:

$$\mathbf{z}_t = \mathbf{d}_t + \Psi(\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V})), \quad (8)$$

where $\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}$. The Query, Key, and Value are derived as $\mathbf{Q} = \mathbf{W}_q \mathbf{f}_t$, $\mathbf{K} = \mathbf{W}_k \mathbf{d}_t$, and $\mathbf{V} = \mathbf{W}_v \mathbf{d}_t$, with $\Psi(\cdot)$ representing a final linear projection. This design emphasizes content most relevant to the current frame while avoiding the quadratic complexity and memory of global self-attention on $\mathbf{d}_t$.

It is worth noting that TCC is explicitly designed as a residual module: $\mathbf{z}_t = \mathbf{d}_t + \Psi(\dots)$. Since the main feature stream $\mathbf{d}_t$ is derived from the recurrent hidden state which aggregates long-term temporal context, it inherently maintains strong temporal stability. The attention map acts solely as a high-frequency calibration signal. Consequently, this residual structure ensures that the single-frame guidance enhances spatial details without disrupting the underlying temporal coherence of the video sequence.

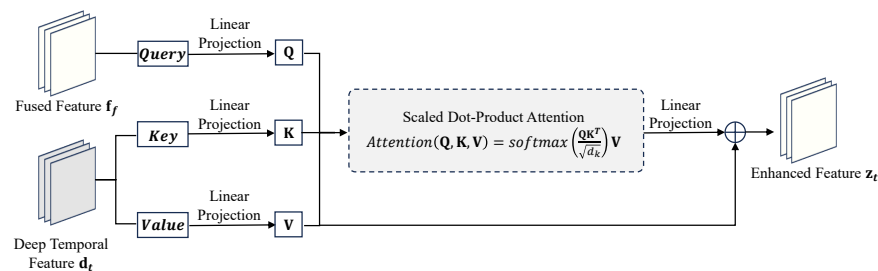


Figure 5. The cross-attention mechanism in the Temporal Cross Calibration (TCC) module. Shallow features serve as the Query to calibrate the deep features.

3.3.3. Upsampling and Image Reconstruction

The reconstruction head employs two pixel-shuffle stages, each scaling by $\times 2$. Let $\tilde{\mathbf{z}}_t^{(1)}$ and $\tilde{\mathbf{z}}_t^{(2)}$ be the channel-expanded tensors prior to shuffling, and let $\Pi_{\times 2}$ denote channel-to-space rearrangement:

$$\mathbf{z}_t^{(1)} = \Pi_{\times 2}(\tilde{\mathbf{z}}_t^{(1)}), \quad \mathbf{z}_t^{(2)} = \Pi_{\times 2}(\tilde{\mathbf{z}}_t^{(2)}). \quad (9)$$

A high-resolution convolutional head maps $\mathbf{z}_t^{(2)}$ to RGB, and an LR bilinear bypass is added as residual:

$$I_t^{\text{SR}} = \Phi(\mathbf{z}_t^{(2)}) + \text{Up}_{\text{bilinear}}(I_t^{\text{LR}}). \quad (10)$$

where $\Phi(\cdot)$ denotes the final HR convolution. Pixel-shuffle converts spatial enlargement into channel rearrangement, avoiding checkerboard artifacts and reducing the amount of computation performed at high resolution; the residual design enables the network to better concentrate on learning high-frequency details.

3.4. Dataset Description

We use REDS [27] for training and evaluate on Vid4 [9], SPMCs [28], and REDS4 [27]. All datasets are downsampled using bicubic interpolation (BI) to create LR-HR training pairs. REDS covers diverse scenes and strong/non-rigid motion, with a resolution around 720 p and typically 100 frames per sequence. We adopt the official training split (about 240 sequences), generate $\times 4$ low-resolution (LR) inputs from high-resolution (HR) frames via bicubic downsampling. For data augmentation, we employ a set of geometric transformations, including vertical flips and random horizontal, as well as rotations by 90° .

For evaluation, Vid4 consists of four classic sequences, each with roughly 30 frames at standard-definition resolution. It contains moderate camera/object motion and many fine repetitive textures, and is commonly used to assess temporal consistency and detail fidelity. SPMCs includes 11 sequences (about 31 frames per sequence, resolution $\approx 960 \times 540$) spanning translation, scaling, panning/tilting, and noticeable parallax, thus emphasizing

ing alignment robustness and temporal aggregation. REDS4 is a four-sequence subset selected from the REDS validation set, featuring fast motion, illumination changes, and motion blur—challenges that effectively gauge the generalization of lightweight methods in complex real-world scenarios.

3.5. Evaluation Metrics

The proposed system was implemented using the PyTorch 2.7 framework and trained on 8 NVIDIA GeForce RTX 4090 GPUs. For the training data, we extracted video clips of 30 frames in length. Input frames were randomly cropped to a resolution of 256×256 and augmented through random flips and rotations. Optimization uses Adam [29] with an initial learning rate of 1×10^{-4} and a Cosine Annealing with Warm Restarts schedule [30]. We used a batch size of 32 and trained the model for a total of 10^5 iterations, which corresponds to approximately ten epochs over the dataset.

We report reconstruction quality using PSNR and SSIM. PSNR is computed from the mean squared error (MSE) between the reconstruction $\hat{I}$ and ground truth I :

$$\text{PSNR} = 10 \log_{10} \left(\frac{\text{MAX}_I^2}{\text{MSE}} \right), \quad \text{MSE} = \frac{1}{N} \sum_{p=1}^N (\hat{I}_p - I_p)^2. \quad (11)$$

Higher PSNR indicates lower distortion. SSIM measures perceptual similarity by comparing luminance, contrast, and structure, which is calculated as:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \quad (12)$$

where μ_x and μ_y represent the mean values of x and y , σ_x^2 and σ_y^2 denote their variances, σ_{xy} is the covariance, and C_1, C_2 are constants to stabilize the division. SSIM is bounded in $[0, 1]$, and higher values are better.

4. Results

4.1. Quantitative Comparison

We compare the performance of LIF-VSR with state-of-the-art (SOTA) VSR methods on Vid4, SPMCs, and REDS4 benchmarks. Table 1 summarizes the PSNR and SSIM scores.

Table 1. Performance comparison of our LIF-VSR against SOTA methods on the task of $4\times$ video super-resolution. The evaluation is conducted on the Vid4, SPMCs, and REDS4 benchmark datasets, with PSNR and SSIM as the primary metrics. The top three performing methods are highlighted: **first** (bold), second (underlined), and *third* (italic).

Method	Year	Vid4		SPMCs		REDS4	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Bicubic	-	21.80	0.5246	25.67	0.7260	26.14	0.7292
VESPCN [9]	2017	25.09	0.7323	27.83	0.8230	29.29	0.8451
DUF [31]	2018	26.80	0.8134	<u>30.10</u>	0.8556	29.50	0.8283
DBPN [32]	2018	25.32	0.7362	27.92	0.8225	29.44	0.8475
RCAN [3]	2018	25.46	0.7404	28.36	0.8287	29.51	0.8489
PFNL [33]	2019	26.40	0.8284	28.66	0.8478	29.63	0.8502
SOF-VSR [11]	2020	26.02	0.7713	28.21	0.8324	29.73	0.8519
ZS-Mo [34]	2020	26.14	0.7974	28.80	0.8635	29.93	0.8634
TDAN [17]	2020	26.16	0.7821	28.51	0.8412	29.87	0.8533
D3Dnet [18]	2020	26.52	0.7993	28.78	0.8523	30.51	0.8657
FastDVDnet [35]	2021	26.14	0.7719	28.53	0.8465	29.57	0.8474
TMNet [36]	2021	26.23	0.8041	28.78	<i>0.8640</i>	29.91	0.8633
RSTT [25]	2022	26.20	0.7991	28.86	0.8634	30.11	0.8613
STDAN [37]	2023	26.28	0.8041	28.94	<u>0.8687</u>	29.98	0.8613
LRGAN [38]	2023	26.80	0.8149	28.95	0.8608	29.82	0.8383
L-VSR [22]	2025	26.95	0.8188	29.02	0.8583	<u>30.71</u>	0.8780
FDI-VSR [39]	2025	<u>27.29</u>	0.8230	29.84	0.8597	31.11	0.8674
LIF-VSR (ours)	2025	27.65	<u>0.8283</u>	31.61	0.8874	30.69	0.8566

As shown in Table 1:

- On **Vid4**, LIF-VSR achieves the highest PSNR of 27.65 dB, surpassing the second-best FDI-VSR by 0.36 dB. It also obtains a competitive SSIM of 0.8283.
- On **SPMCs**, our method secures the top spot in both metrics with a PSNR of 31.61 dB and an SSIM of 0.8874, outperforming L-VSR by a margin of 2.59 dB.
- On **REDS4**, LIF-VSR achieves a PSNR of 30.69 dB, ranking third behind FDI-VSR (31.11 dB) and L-VSR (30.71 dB).

4.2. Efficiency Analysis

Table 2 presents the model complexity in terms of parameter count and FLOPs (calculated on a 64×64 input with $\times 4$ upscaling).

Table 2. Comparison of model parameters and computational cost. All metrics are for a single LR frame input of size 64×64 and a scale factor of $\times 4$.

Method	Input Size	Scale	Parameter (M)	Computational Cost Per Frame (GFLOPs)
BasicVSR++ [6]	64×64	$\times 4$	7.03	28.26
BasicVSR [5]	64×64	$\times 4$	6.29	26.26
DUF [31]	64×64	$\times 4$	5.82	16.90
EDVR-M [4]	64×64	$\times 4$	3.15	32.91
LIF-VSR (ours)	64×64	$\times 4$	3.06	17.60

LIF-VSR contains 3.06 million parameters, which is approximately 49% of BasicVSR (6.29 M) and 44% of BasicVSR++ (7.03 M). In terms of computational cost, our model requires 17.60 GFLOPs per frame, which is lower than BasicVSR (26.26 GFLOPs) and EDVR-M (32.91 GFLOPs), and comparable to the lightweight DUF (16.90 GFLOPs).

4.3. Ablation Experiments

In this section, we conduct a comprehensive ablation study to verify the effectiveness of our proposed framework. We first analyze the contribution of each core component, followed by a detailed discussion on the sampling strategy within the Sparse Attention Fusion (SAF) module.

4.3.1. Effectiveness of Core Components

We investigate the contribution of Implicit Deformable Alignment (IDA), Sparse Attention Fusion (SAF), and Temporal Cross Calibration (TCC) by incrementally adding them to the baseline. The quantitative results are reported in Table 3.

Table 3. Ablation experiments of the core components of LIF-VSR on the Vid4, SPMCs, and REDS4 datasets. ‘✓’ indicates the module is active, while ‘×’ indicates it is deactivated.

Dataset	IDA	SAF	TCC	PSNR	SSIM
Vid4	×	×	×	25.81	0.7502
	✓	×	×	27.13	0.8115
	✓	✓	×	27.52	0.8248
	✓	✓	✓	27.65	0.8283
SPMCs	×	×	×	28.95	0.8133
	✓	×	×	30.79	0.8641
	✓	✓	×	31.45	0.8832
	✓	✓	✓	31.61	0.8874
REDS4	×	×	×	28.62	0.7989
	✓	×	×	30.08	0.8381
	✓	✓	×	30.56	0.8525
	✓	✓	✓	30.69	0.8566

As observed in Table 3:

- **Baseline + IDA:** Introducing the IDA module increases the PSNR by 1.32 dB on Vid4 and 1.84 dB on SPMCs compared to the baseline, confirming that feature-space alignment significantly outperforms simple concatenation.
- **Baseline + IDA + SAF:** Adding the SAF module further improves PSNR by 0.39 dB on Vid4 and 0.66 dB on SPMCs, validating the efficacy of sparse temporal aggregation.
- **Full Model (+ TCC):** The inclusion of TCC yields the final performance of 27.65 dB on Vid4 and 31.61 dB on SPMCs, demonstrating the benefit of high-frequency calibration.

4.3.2. Analysis of Sampling Granularity in SAF

The number of sampling points K in the SAF module determines the granularity of temporal aggregation. We interpret $K = 9$ as the sparse equivalent of a standard 3×3 convolution kernel, which is a proven “sweet spot” in computer vision for capturing local geometric transformations. To verify this, we conducted an ablation study with $K \in \{4, 9, 25\}$, corresponding to 2×2 , 3×3 , and 5×5 receptive fields, respectively.

As shown in Table 4, reducing K to 4 results in a noticeable performance drop (0.18 dB on Vid4), indicating insufficient context aggregation. Conversely, increasing K to 25 yields only marginal gains (+0.03 dB) but increases computational complexity. Therefore, $K = 9$ offers the optimal trade-off.

It is also worth noting that we employ *multi-group deformable attention* with $G = 16$ groups. This means that while each head samples only 9 points, the model effectively aggregates information from $16 \times 9 = 144$ distinct spatial locations per position, providing a rich context summary without the prohibitive memory cost of dense global attention.

Table 4. Ablation study on the number of sampling points (K) in the SAF module. Evaluated on Vid4 and SPMCs.

Sampling Points (K)	Params (M)	GFLOPs	Vid4 PSNR (dB)	SPMCs PSNR (dB)
$K = 4 (2 \times 2)$	3.04	17.42	27.47	31.42
$K = 9 (3 \times 3)$ (Default)	3.06	17.60	27.65	31.61
$K = 25 (5 \times 5)$	3.12	18.15	27.68	31.70

4.4. Visualization Analysis

Visual comparisons on sequences from the Vid4, SPMCs, and REDS4 datasets are shown in Figure 6, Figure 7, and Figure 8, respectively.

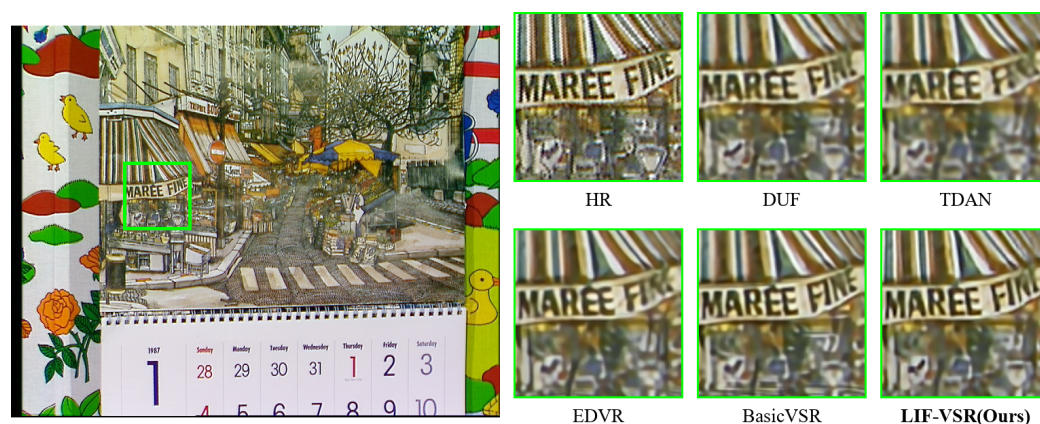


Figure 6. Qualitative comparison on the “Calendar” sequence from Vid4. Our LIF-VSR excels at restoring sharp text characters (e.g., “MAREE FINE”) and fine awning patterns.



Figure 7. Qualitative comparison on a sequence from the SPMCs dataset. LIF-VSR demonstrates superior ability in preserving the geometry of the neon signage and the surrounding brick textures.



Figure 8. Qualitative comparison on a sequence from REDS4. LIF-VSR effectively handles the distant building details, resulting in significantly clearer window grids compared to other methods.

In the “Calendar” sequence from Vid4 (Figure 6), methods such as DUF and TDAN produce blurred text where the characters are barely recognizable. In contrast, LIF-VSR effectively reconstructs the sharp edges of the text “MAREE FINE” and the distinct striped patterns on the awning, demonstrating superior recovery of high-frequency details.

For the sequence from SPMCs (Figure 7), which features a complex brick facade and neon signage, competing methods like EDVR and BasicVSR exhibit aliasing artifacts and blurring around the cursive letters. LIF-VSR maintains the structural integrity of the “Grasshopper” signage and separates the neon tubes from the background brick texture more clearly.

In the REDS4 example (Figure 8), the scene contains distant building structures that challenge temporal consistency. While other methods struggle to resolve the grid patterns of the windows, resulting in severe motion blur, LIF-VSR successfully restores the rectangular window shapes and sharp building edges, proving its robustness in handling complex outdoor scenes.

To address the concern that the single-frame guidance in the TCC module might introduce temporal instability, we visualized the temporal profiles (vertical slices of the video volume over time) across the Vid4, SPMCS, and REDS4 datasets. As shown in Figure 9, the temporal trajectories generated by LIF-VSR are visually indistinguishable from those of the Ground Truth (HR).

Specifically, in the Vid4 and REDS4 samples, which contain complex motion (manifested as curved patterns in the temporal profile), LIF-VSR accurately reconstructs the continuous flow of pixels without exhibiting jagged artifacts or discontinuities. Similarly, in the static regions of the SPMCS sample, the straight vertical lines remain stable and sharp. This strong alignment with the HR reference confirms that the residual design of the

TCC module enhances spatial details while strictly preserving the underlying temporal coherence of the video.



Figure 9. Temporal profile comparison across three datasets. The red lines in the reference frames indicate the sampling positions. The temporal profiles of LIF-VSR exhibit smooth, continuous structures that closely match the Ground Truth (HR), verifying that the proposed method maintains excellent temporal consistency.

5. Discussion

The experimental results presented in Section 4 demonstrate that LIF-VSR achieves a superior balance between reconstruction quality and model complexity. As illustrated in Table 1, our method outperforms existing lightweight and prominent recurrent networks on the Vid4 and SPMCS benchmarks. A critical insight from our efficiency analysis is that high performance in VSR does not necessarily require heavy optical flow estimation or dense global attention. By reducing the parameter count to 3.06 M—approximately 50% of BasicVSR—LIF-VSR proves that a unidirectional recurrent scheme with minimalist state propagation is sufficient for capturing long-range temporal dependencies, provided that the alignment and fusion mechanisms are carefully designed.

The effectiveness of our design philosophy is further supported by the ablation study. The dominance of the Implicit Deformable Alignment (IDA) module in performance gains confirms that feature-space alignment is more robust than explicit motion compensation, particularly for complex, non-rigid motions. Furthermore, the success of the Sparse Attention Fusion (SAF) module validates our hypothesis that global attention is redundant for VSR. By selecting only a small set of key sampling points, SAF effectively aggregates relevant temporal cues without the quadratic complexity of full self-attention. The Temporal Cross Calibration (TCC) further refines these features by using the current frame’s high-frequency information to prevent the “smoothing out” effect.

Despite these promising results, LIF-VSR has limitations. The random memory access patterns inherent in deformable convolutions and sparse sampling may exhibit lower inference speedup on certain hardware accelerators compared to standard convolutions. Additionally, while our unidirectional scheme is efficient, extremely large or abrupt scene changes might challenge the propagation stability compared to bidirectional approaches.

6. Conclusions

In this paper, we addressed the efficiency bottleneck in video super-resolution by proposing LIF-VSR, a lightweight framework tailored for practical deployment. By synergizing Implicit Deformable Alignment, Sparse Attention Fusion, and Temporal Cross Calibration, we successfully eliminated the need for heavy optical flow estimation and

dense attention mechanisms. Our method achieves state-of-the-art performance on standard benchmarks while maintaining a low computational footprint.

Future work will focus on two directions: exploring lightweight bidirectional propagation strategies to further handle complex scene transitions, and optimizing the inference latency of deformable operators on edge hardware devices.

Author Contributions: Conceptualization, S.Z. and H.Z.; methodology, S.Z. and W.C.; software, S.Z. and X.W.; validation, S.Z., K.S. and Z.Z.; investigation, S.Z., H.Z., Z.Z. and Z.H.; data curation, Z.H. and K.S.; writing—original draft, S.Z. and X.W.; writing—review and editing, S.Z., W.C. and K.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the S&T Program of Hebei, grant number 252Y0401D.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All datasets utilized in this research are publicly available. For training and evaluation, we used the Vid4 (<https://people.csail.mit.edu/celiu/CVPR2011/videoSR.zip> (accessed on 12 December 2025)), SPMCs (https://github.com/jiangsutx/SPMC_VideoSR (accessed on 12 December 2025)), and REDS4 (<https://seungjunna.github.io/Datasets/reds.html> (accessed on 12 December 2025)) datasets, all of which are accessible at their respective URLs.

Conflicts of Interest: Authors Songyi Zhang, Xiaolin Wang, Kailei Song, Zhizhuo Han and Zhitao Zhang are employed by Hebei Far East Communication Engineering Co., Ltd. company. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Dong, C.; Loy, C.C.; He, K.; Tang, X. Learning a deep convolutional network for image super-resolution. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 184–199.
2. Lim, B.; Son, S.; Kim, H.; Nah, S.; Mu Lee, K. Enhanced deep residual networks for single image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 136–144.
3. Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; Fu, Y. Image super-resolution using very deep residual channel attention networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 286–301.
4. 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, 16–20 June 2019; pp. 1954–1963.
5. 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, Virtual, 19–25 June 2021; pp. 4947–4956.
6. 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.
7. Kim, J.; Lee, J.K.; Lee, K.M. Accurate image super-resolution using very deep convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1646–1654.
8. 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, Virtual, 11–17 October 2021; pp. 1833–1844.
9. 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.
10. 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](#)]
11. Wang, L.; Guo, Y.; Liu, L.; Lin, Z.; Deng, X.; An, W. Deep video super-resolution using HR optical flow estimation. *IEEE Trans. Image Process.* **2020**, *29*, 4323–4336. [[CrossRef](#)] [[PubMed](#)]

12. Ranjan, A.; Black, M.J. Optical flow estimation using a spatial pyramid network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4161–4170.
13. Sun, D.; Yang, X.; Liu, M.Y.; Kautz, J. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8934–8943.
14. Teed, Z.; Deng, J. Raft: Recurrent all-pairs field transforms for optical flow. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 402–419.
15. 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.
16. Zhu, X.; Hu, H.; Lin, S.; Dai, J. Deformable convnets v2: More deformable, better results. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 9308–9316.
17. 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, 14–19 June 2020; pp. 3360–3369.
18. 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](#)]
19. 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.* **2025**, *41*, 3581–3593. [[CrossRef](#)]
20. 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, 18–22 June 2018; pp. 6626–6634.
21. 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, 15–20 June 2019; pp. 3897–3906.
22. 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. [[CrossRef](#)]
23. Wang, H.; Li, S.; Zhao, M. A lightweight recurrent aggregation network for satellite video super-resolution. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *17*, 685–695. [[CrossRef](#)]
24. Liang, J.; Cao, J.; Fan, Y.; Zhang, K.; Ranjan, R.; Li, Y.; Timofte, R.; Van Gool, L. Vrt: A video restoration transformer. *IEEE Trans. Image Process.* **2024**, *33*, 2171–2182. [[CrossRef](#)] [[PubMed](#)]
25. 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.
26. Zhou, M.; Huang, J.; Guo, C.L.; Li, C. Fourmer: An efficient global modeling paradigm for image restoration. In Proceedings of the International Conference on Machine Learning, PMLR, Honolulu, HI, USA, 23–29 July 2023; pp. 42589–42601.
27. 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 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 12–16 June 2019.
28. 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.
29. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
30. Loshchilov, I.; Hutter, F. SGDR: Stochastic Gradient Descent with Warm Restarts. In Proceedings of the International Conference on Learning Representations, San Juan, PR, USA, 2–4 May 2016.
31. 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, 18–22 June 2018; pp. 3224–3232.
32. Haris, M.; Shakhnarovich, G.; Ukita, N. Deep back-projection networks for super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1664–1673.
33. Yi, P.; Wang, Z.; Jiang, K.; Jiang, J.; Ma, J. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019; pp. 3106–3115.
34. Xiang, X.; Tian, Y.; Zhang, Y.; Fu, Y.; Allebach, J.P.; Xu, C. Zooming slow-mo: Fast and accurate one-stage space-time video super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 3370–3379.
35. 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, Virtual, 11–17 October 2021; pp. 4429–4438.
36. Xu, G.; Xu, J.; Li, Z.; Wang, L.; Sun, X.; Cheng, M.M. Temporal modulation network for controllable space-time video super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 6388–6397.

37. 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*, 10606–10616. [[CrossRef](#)] [[PubMed](#)]
38. Zhu, Y.; Li, G. A lightweight recurrent grouping attention network for video super-resolution. *Sensors* **2023**, *23*, 8574. [[CrossRef](#)] [[PubMed](#)]
39. Lim, D.; Choi, J. FDI-VSR: Video Super-Resolution Through Frequency-Domain Integration and Dynamic Offset Estimation. *Sensors* **2025**, *25*, 2402. [[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.