Video Super-Resolution Transformer with Masked Inter&Intra-Frame Attention

Xingyu Zhou[1] Leheng Zhang[1] Xiaorui Zhao[1] Keze Wang[2] Leida Li[3] Shuhang Gu[1*]

[1]University of Electronic Science and Technology of China [2]Sun Yat-sen University [3]Xidian University

{xy.chous526, lehengzhang12, zzzhaoxiaorui, kezewang, shuhanggu}@gmail.com, ldli@xidian.edu.cn

Abstract

Recently, Vision Transformer has achieved great success in recovering missing details in low-resolution sequences, i.e., the video super-resolution (VSR) task. Despite its superiority in VSR accuracy, the heavy computational burden as well as the large memory footprint hinder the deployment of Transformer-based VSR models on constrained devices. In this paper, we address the above issue by proposing a novel feature-level masked processing framework: VSR with Masked Intra and inter-frame Attention (MIA-VSR). The core of MIA-VSR is leveraging feature-level temporal continuity between adjacent frames to reduce redundant computations and make more rational use of previously enhanced SR features. Concretely, we propose an intra-frame and inter-frame attention block which takes the respective roles of past features and input features into consideration and only exploits previously enhanced features to provide supplementary information. In addition, an adaptive block-wise mask prediction module is developed to skip unimportant computations according to feature similarity between adjacent frames. We conduct detailed ablation studies to validate our contributions and compare the proposed method with recent state-of-the-art VSR approaches. The experimental results demonstrate that MIA-VSR improves the memory and computation efficiency over state-of-the-art methods, without trading off PSNR accuracy. The code is available at https://github.com/LabShuHangGU/MIA-VSR.

1. Introduction

Image super-resolution (SR) refers to the process of recovering sharp details in high resolution (HR) images from low resolution (LR) observations. Due to its great value in practical usages, e.g., surveillance and high definition display, SR has been a thriving research topic over the past twenty years. Generally, compared with single image super-resolution which only exploits intra-frame information to
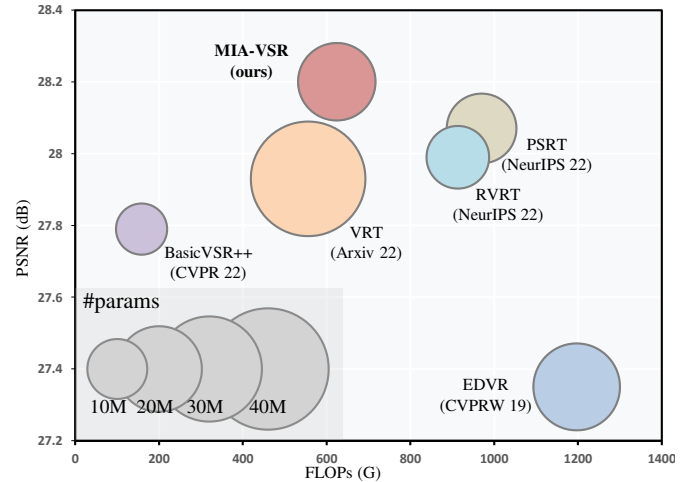

Figure 1. **PSNR(dB) and FLOPs(G) comparison on the Vid4 [24] dataset.** We compare our MIA-VSR model with the state-of-the-art temporal sliding-window and recurrent based VSR models, including EDVR [36], BasicVSR++ [5], VRT [21], RVRT [22] and PSRT [32]. Our MIA-VSR model outperforms these methods and strikes a balance between performance and compute efficiency.

estimate the missing details, video super-resolution (VSR) additionally leverages the temporal information to recover the HR frames and therefore often leads to better SR results.

The key of VSR lies in making rational use of temporal information. Researchers in the early stages utilized convolutional neural networks (CNNs) to extract features and have investigated advanced information propagation strategies [4, 5, 8, 13, 32], sophisticated alignment modules [4, 22, 32, 35, 36, 39], effective training strategies [23, 38] as well as elaborately designed network architectures [15, 19] for the pursuit of highly accurate VSR results. Recently, with the rapid development of Transformers in computer vision, several attempts have been made to exploit Transformers for better recovering missing details in LR sequences [21, 22, 25, 30, 32]. Due to the powerful representation learning capabilities of self-attention structure, these Transformer-based approaches have raised the state-of-the-art in VSR to a new level.

---

*Corresponding author.

In spite of its superior SR results, the heavy computational burden and large memory footprint [21, 22, 32] of Transformer-based VSR approaches limit their application to constrained devices. With the common availability of video cameras, efficient video processing has become an increasingly important research topic in the literature on computer vision. Paralleling to network pruning and quantization, which is applicable to all kinds of application, exploiting temporal redundancy for avoiding unnecessary computation is a specific strategy for accelerating video processing. In their inspiring work, Habibian *et al*. [10] proposed a skip-convolution method that restricts the computation of video processing only to regions with significant changes while skipping the others. Although the skip-covolution strategy could save computations without significant performance drop in high-level tasks, e.g. object detection and human pose estimation, as low-level vision tasks such as video super-resolution are highly sensitive to minor changes in image content, whether such a skip-processing mechanism is applicable to VSR is still an open question.

In this paper, we provide an affirmative answer to the above question with a novel masked VSR framework, i.e., **M**asked **I**nter&Intra frame **A**ttention (MIA) model. Our MIA-VSR method could make use of temporal continuity to recover missing HR details, while, at the same time, leveraging temporal redundancy for reducing unnecessary computations. Concretely, our MIA-VSR model advances the existing VSR approaches in the following two aspects. First, we develop a tailored inter-frame and intra-frame attention block (IIAB) for making more rational use of previously enhanced features. Instead of directly inheriting the Swin-Transformer block [27] to process concatenated hidden states and image feature, our proposed IIAB block takes into account the respective roles of past features and input features and only utilizes the image feature of the current frame to generate the query token. As a result, the proposed IIAB block not only reduces the computational consumption of the original joint self-attention strategy by a large margin, but also aligns the intermediate feature with spatial coordinates to enable efficient masked processing. Secondly, we propose a feature-level adaptive masked processing mechanism to reduce redundant computations according to the continuity between adjacent frames. Differently from previous efficient processing models, which simply determine skipable regions of the whole network according to pixel intensities, we adopt a feature-level selective skipping strategy and pass by computations of a specific stage adaptively. The proposed feature-level adaptive masking strategy enables our MIA-VSR to save computations and at the same time achieve good VSR results. Our contributions can be summarized as follows.

- We present a novel feature-level masked processing framework for efficient VSR, which is able to take ad-

vantage of the temporal continuity to reduce redundant computations in the VSR task.
- We propose an intra-frame and inter-frame attention block, which could effectively extract spatial and temporal supplementary information to enhance SR features.
- We propose an adaptive mask prediction module, which masks out unimportant regions according to the feature similarity between adjacent frames for different stages of processing in the VSR model.
- We compare our proposed MIA-VSR model with state-of-the-art VSR models against which our approach could generate superior results with less computations and memory footprints.

2. Related Work

2.1. Video Super-Resolution

According to how the temporal information is utilized, existing deep learning based VSR methods can be grouped into two categories: temporal sliding-window based methods and recurrent based methods.

Temporal sliding-window based VSR. Given an LR video sequence, a category of approaches processes the LR frames in a temporal sliding-window manner which aligns adjacent LR frames with the reference frame to estimate a single HR output. The alignment module plays an essential role in temporal sliding-window based method. Earlier work [1, 26, 34] explicitly estimated the optical flow to align adjacent frames. Recently, implicit alignment modules [39] were proposed to perform alignment in the high-dimensional feature space. Dynamic filters [18], deformable convolutions [7, 35, 36] and attention modules [14, 19] have been developed to conduct motion compensation implicitly in the feature space. In addition to the alignment module, another important research direction is investigating sophisticated network architectures to process aligned images. Many works were aimed at estimating HR images from multiple input images in a temporal sliding-window manner [2, 19, 21, 36]. Although the alignment module enables sliding window based VSR networks to better leverage temporal information from adjacent frames, accessible information to the VSR models is limited by the size of temporal sliding window and these methods could only make use of temporal information from limited number of input video frames.

Recurrent framework based VSR. Another category of approaches applies recurrent neural networks to exploit temporal information from more frames. FRVSR [31] first proposed a recurrent framework that utilizes optical flow to align the previous HR estimation and the current LR input for VSR. RLSP [8] propagates high-dimensional hidden states instead of the previous HR estimation to better ex-
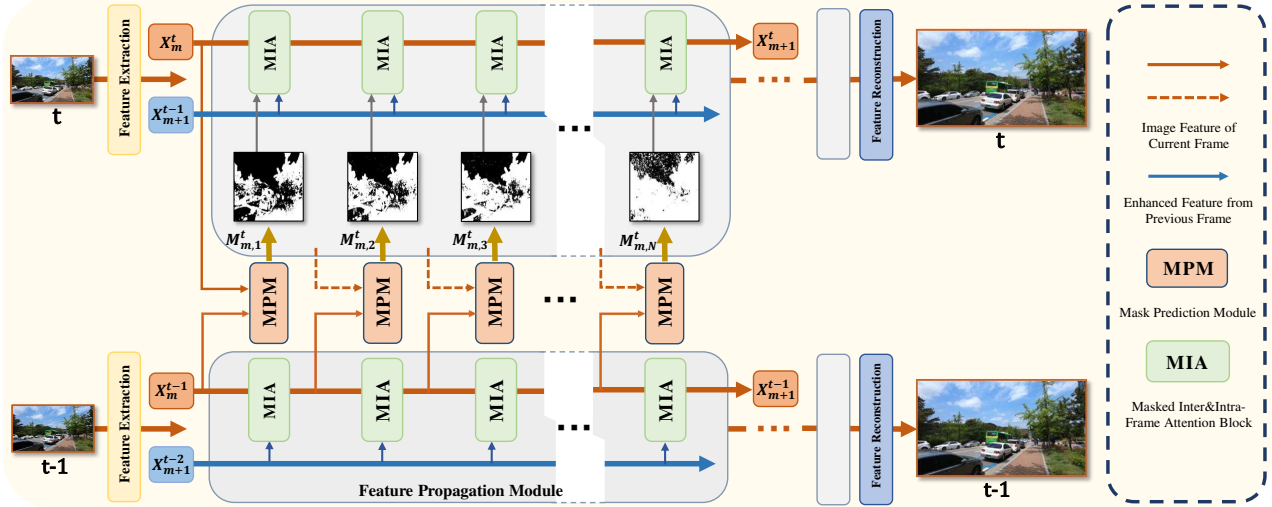
Figure 2. **The overall architecture of MIA-VSR.** We develop a feature-level masked processing framework which uses the mask prediction module (MPM) to reduce redundant computations by leveraging temporal continuity, and propose a masked intra-frame and inter-frame (MIA) block to make more rational use of previous enhanced features to support the feature enhancement of the current frame. Our MIA-VSR model can be easily extended to the bi-directional second-order grid propagation framework as [5]. More details of our proposed MIA-VSR can be found in Section 3.

ploit long-term information. RSDN [13] further extended RLSP [8] by decomposing the LR frames into structure and detail layers and introduced an adaptation module to selectively use the information from hidden state. BasicVSR [3] utilized bi-directional hidden states, and BasicVSR++ [5] further improved BasicVSR with second-order grid propagation and flow-guided deformable alignment. PSRT [32] adopted the bi-directional second-order grid propagation framework of BasicVSR++ and utilized multi-frame self-attention block to jointly process the outputs of the feature propagation of the previous frames and the input features. Generally, by passing the high dimensional hidden states or output feature, the recurrent based VSR could incorporate temporal information from more frames for estimating the missing details and therefore achieve better VSR results. Our proposed MIA-VSR follows the general framework of bi-directional second-order hidden feature propagation while introducing masked processing strategy and intra&inter-frame attention block for the pursuit of better trade-off between VSR accuracy, computational burden and memory footprint.

2.2. Efficient Video Processing

Various strategies of reducing temporal redundancy have been explored for efficient video processing. A category of approaches adopts the general network optimization strategy and utilizes pruning [37] or distillation [11] methods to train light-weight networks for efficient video processing. In order to take extra benefit from the temporal continuity of video frames, another category of methods only computes expensive backbone features on key-frames and applies concatenation methods [16], optical-flow based alignment methods [12, 16, 20, 29, 41, 42], dynamic convolution methods [29] and self-attention methods [12, 25] to enhance features of other frames with keyframe features. Most recently, [10] proposed a skip-convolution approach which only conducts computation in regions with significant changes between frames to achieve the goal of efficient video processing. However, most of the advanced efficient video processing schemes described above were designed for high-level vision tasks such as object detection and pose estimation. To the best of our knowledge, our study is the first work that leverages the temporal continuity across different areas for each block to reduce redundant computation for low-level VSR tasks.

3. Methodology

3.1. Overall Architecture

Given T frames of low-resolution video sequence $\boldsymbol{I}_{LR} \in \mathbb{R}^{T \times H \times W \times 3}$, our goal is to reconstruct the corresponding HR video sequence $\boldsymbol{I}_{HR} \in \mathbb{R}^{T \times sH \times sW \times 3}$, where s is the scaling factor and H, W, 3 are the height, width and channel number of input frames.

The overall architecture of our proposed MIA-VSR model is shown in Fig.2. We built our MIA-VSR framework upon the bi-directional second-order grid propagation framework of BasicVSR++ [5], which has also been adopted in the recent state-of-the-art method PSRT [32]. The whole model consists of three parts, i.e., the shallow
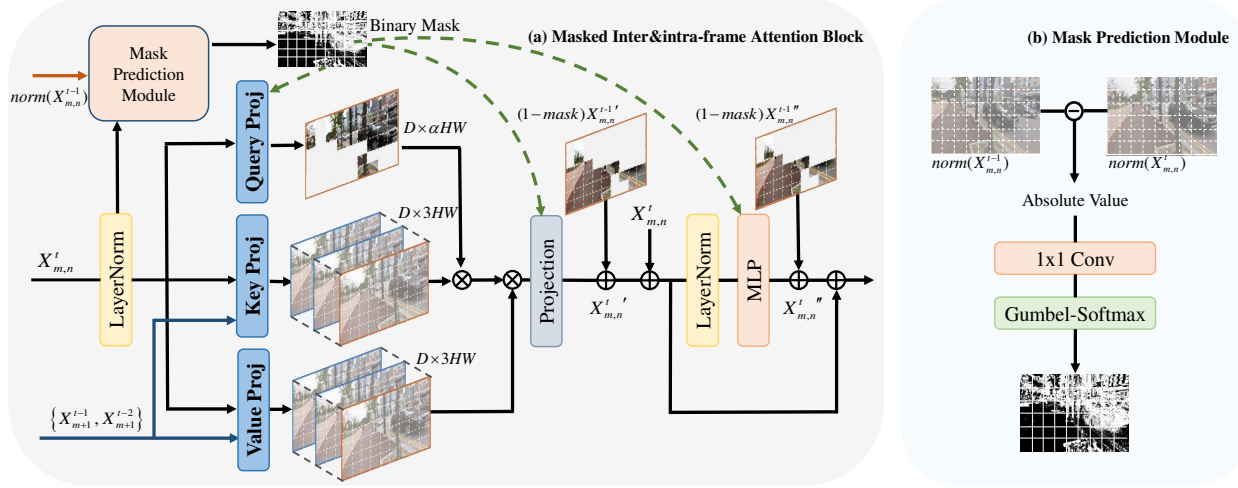
Figure 3. **Illustration of the inter&intra-frame attention block with adaptive masked processing module.** The adaptive mask prediction module **(b)** in the IIAB block acts in the `Attention` module's linear layer which produce the Query, the projection layer and the linear layer in the `FFN` module during inference to reduce temporal and sptical redundancy calculations **(a)**. ${X_{m,n}^{t}}'$ and ${X_{m,n}^{t}}''$ refer to the processed hidden feature in the `Attention` and `FFN` module.

feature extraction part, the recurrent feature refinement part and the feature reconstruction part. We follow previous works BasicVSR++ [5] and PSRT [32] which use a plain convolution operation to extract shallow features and adopt a pixel-shuffle layer [9, 33] to reconstruct HR output with refined features. The patch alignment method used in PSRT [33] is used to align the adjacent frames. We improve the efficiency of existing works by reducing redundant computations in the recurrent feature refinement part.

Generally, the recurrent feature refinement part comprises **M** feature propagation modules and each feature propagation module consists of **N** cascaded processing blocks. The feature propagation module takes the enhanced outputs from previous frames as well as image feature of the current frame as input. For the t-th frame, let us denote the input feature for the m-th feature propagation module as X_m^t. The feature propagation module takes X_m^t, X_{m+1}^{t-1} and X_{m+1}^{t-2} as inputs to calculate the enhanced feature:

$$X_{m+1}^t = \text{FPM}(X_m^t, X_{m+1}^{t-1}, X_{m+1}^{t-2}), \quad (1)$$

where X_{m+1}^t is the enhanced output of the current frame, X_{m+1}^{t-1} and X_{m+1}^{t-2} are enhanced feature from the past two frames. To be more specific, in each feature propagation module, cascaded processing blocks utilize outputs from the previous frames to enhance input features. In comparison to the previous methods that directly inherit the Swin transformer block to process concatenated features, we propose a tailored intra&inter frame attention block (IIAB) to more efficiently enhance X_m^t with the help of X_{m+1}^{t-1} and X_{m+1}^{t-2}:

$$X_{m,n+1}^t = \text{IIAB}(X_{m,n}^t, X_{m+1}^{t-1}, X_{m+1}^{t-2}), \quad (2)$$

where $X_{m,n}^t$ is the input to the n-th IIAB in the m-th feature propagation module, $X_{m,0}^t = X_m^t$ and $X_{m,N}^t = X_{m+1}^t$ are the input and output of the whole FPM, respectively. More details of the proposed IIAB will be introduced in the following subsection 3.2.

In order to reduce redundant computations according to temporal continuity, we further develop an adaptive mask prediction module to generate block-wise masks $M_{m,n}^t$, with which we could directly utilize the outputs from past frame and selectively skip unimportant computation:

$$\hat{X}_{m,n}^t = M_{m,n}^t \odot X_{m,n}^t + (1 - M_{m,n}^t) \odot \hat{X}_{m,n}^{t-1}, \quad (3)$$

where $\hat{X}_{m,n}^t$ is the processed hidden feature in the IIAB block (shown as ${X_{m,n}^t}'$ and ${X_{m,n}^t}''$ in Fig.3) and $\odot$ is the point-wise multiplication operation. More details of our mask predicting module will be presented in Section 3.3.

3.2. Inter&Intra-Frame Attention Block

As introduced in the previous section, cascade processing blocks play a key role in extracting supplementary information from previous frames to enhance SR features. To achieve this goal, the previous method [22, 32] simply adopts the multi-head self-attention block with shifted local windows in the Swin Transformer to process concatenated hidden states (enhanced features of previous frames) and the current input feature. In this paper, we take the respective role of the previously enhanced feature and the current feature into consideration and propose a intra&inter frame attention block (IIAB) for efficient VSR. To be more specific, we think the enhanced features of the previous frames X_{m+1}^{t-1} and X_{m+1}^{t-2} should only used for providing

supplementary information and do not need to be further enhanced. Therefore, we only utilize feature of the current frame to generate Query Tokens, and adopt enhanced features from the previous frames as well as feature of the current frame to generate Key and Value Tokens:

$$\boldsymbol{Q}_{m,n}^t = \boldsymbol{X}_{m,n}^t \boldsymbol{W}_{m,n}^Q,$$
$$\begin{cases} \boldsymbol{K}_{m,n}^{t,intra} = \boldsymbol{X}_{m,n}^t \boldsymbol{W}_{m,n}^K, \\ \boldsymbol{K}_{m,n}^{t,inter} = \left[\boldsymbol{X}_{m+1}^{t-1}; \boldsymbol{X}_{m+1}^{t-2} \right] \boldsymbol{W}_{m,n}^K, \end{cases} \quad (4)$$
$$\begin{cases} \boldsymbol{V}_{m,n}^{t,intra} = \boldsymbol{X}_{m,n}^t \boldsymbol{W}_{m,n}^V, \\ \boldsymbol{V}_{m,n}^{t,inter} = \left[\boldsymbol{X}_{m+1}^{t-1}; \boldsymbol{X}_{m+1}^{t-2} \right] \boldsymbol{W}_{m,n}^V, \end{cases}$$

where $\boldsymbol{W}_{m,n}^Q$, $\boldsymbol{W}_{m,n}^K$ and $\boldsymbol{W}_{m,n}^V$ are the respective projection matrices; $\boldsymbol{Q}_{m,n}^t$ represents the Query Tokens generated from the current input feature; $\boldsymbol{K}_{m,n}^{t,intra}$ and $\boldsymbol{K}_{m,n}^{t,inter}$ are the intra-frame and inter-frame Key Tokens and $\boldsymbol{V}_{m,n}^{t,intra}$ and $\boldsymbol{V}_{m,n}^{t,inter}$ are the intra-frame and inter-frame Value Tokens. IIAB jointly calculates the attention map between query token and intra&inter-frame keys to generate the updated feature:

$$\text{IIAB}_{\text{Attention}} = \text{SoftMax}(\boldsymbol{Q}_{m,n}^t \boldsymbol{K}_{m,n}^t{}^T / \sqrt{d} + B) \boldsymbol{V}_{m,n}^t, \quad (5)$$

where $\boldsymbol{K}_{m,n}^t = [\boldsymbol{K}_{m,n}^{t,inter}; \boldsymbol{K}_{m,n}^{t,intra}]$ and $\boldsymbol{V}_{m,n}^t = [\boldsymbol{V}_{m,n}^{t,inter}; \boldsymbol{V}_{m,n}^{t,intra}]$ are the concatenated intra&inter-frame tokens; d is the channel dimension of the Token and B is the learnable relative positional encoding. It should be noted that in Eq.4, all the intermediate inter-frame tokes $\{\boldsymbol{V}_{m,n}^{t,inter}; \boldsymbol{K}_{m,n}^{t,inter}\}_{n=1,\ldots,N}$ are generated from the same enhanced features $[\boldsymbol{X}_{m+1}^{t-1}; \boldsymbol{X}_{m+1}^{t-2}]$; which means that we only leverage mature enhanced features from previous frames to provide supplementary information and do not need to utilize the time consuming self-attention mechanism to jointly update the current feature and previous features. An illustration of our proposed inter-frame and intra-frame attention (IIAB) block is presented in Fig.3. In addition to the attention block, our proposed transformer layer also utilizes LayerNorm and FFN layers, which have been commonly utilized in other Transformer-based VSR architectures [22, 32].

In our implementation, we also adopt the Shift-window strategy to process the input LR frames and the above self-attention calculation is conduct in $W \times W$ non-overlapping windows. Denote our channel number by C, $\boldsymbol{Q}_{m,n}^t$, $\boldsymbol{K}_{m,n}^t$ and $\boldsymbol{V}_{m,n}^t$ in Eq.5 are with sizes of $N^2 \times C$, $3N^2 \times C$ and $3N^2 \times C$, respectively. Therefore, Eq.5 only suffers the calculation $1/3$ in the self-attention calculation procedure in comparison with the joint self-attention processing strategy. Moreover, as we will validate in the ablation study section, our strategy not only avoids unnecessary computation in jointly updating previous features, but also provides better features for the final VSR task.

3.3. Adaptive Masked Processing in the Recurrent VSR Transformer

In this subsection, we present details of how we generate block-wise masks $\boldsymbol{M}_{m,n}^t$ in Eq.3 to further reduce unimportant computations. Habibian *et al.* [10] first proposed a skip convolution mechanism to reduce redundant computations based on differences between pixel values in adjacent frames. However, for the VSR task, skipping the whole computation process of a certain region according to the intensity differences between adjacent input frames will inevitably affect the accuracy of the model. In this paper, we leverage the temporal continuity to reduce the computation of the VSR model in a subtler way: by exploiting feature differences between adjacent frames to select less important computations for each block.

Since features of different stages have various contributions to the final VSR result, using feature differences between adjacent frames to generate binary masks through a uniform threshold is non-trivial. Therefore, we propose an adaptive masked processing scheme which jointly trains tiny mask predicting networks with VSR feature enhancement blocks. To be more specific, in each stage of processing, the mask predicting network takes the difference between normalized features as input:

$$\Delta \boldsymbol{X}_{m,n}^{t-1 \to t} = \|\text{norm}(\boldsymbol{X}_{m,n}^t) - \text{norm}(\boldsymbol{X}_{m,n}^{t-1})\|_1. \quad (6)$$

Then, we employ a Gumbel-softmax [17] gate to sample masking features:

$$\begin{cases} \text{Mask}(\Delta \boldsymbol{X}_{m,n}^{t-1 \to t})) = \dfrac{\exp\left((\log(\pi_1) + g_1)/\tau\right)}{\sum_{i=1}^2 \exp\left((\log(\pi_i) + g_i)/\tau\right)}, \\ \pi_1 = \text{Sigmoid}(f(\Delta \boldsymbol{X}_{m,n}^{t-1 \to t})), \\ \pi_2 = 1 - \text{Sigmoid}(f(\Delta \boldsymbol{X}_{m,n}^{t-1 \to t})), \end{cases}$$
$$(7)$$

where $f(\cdot)$ is a 1×1 convolution layer to weighted sum feature differences from different channels; π_1 could be interpreted as the probability of whether a position should be preserved; and g_1, g_2 are noise samples drawn from a Gumbel distribution, τ is the temperature coefficient and we set it as 2/3 in all of our experiments. With the masking feature $\text{Mask}(\Delta \boldsymbol{X}_{m,n}^{t-1 \to t})$), we could directly set a threshold value to generate the binary mask:

$$\boldsymbol{M}_{m,n}^t = \begin{cases} 1, & if \quad \text{Mask}(\Delta \boldsymbol{X}_{m,n}^{t-1 \to t}) > 0.5; \\ 0, & else. \end{cases} \quad (8)$$

The above Gumbel-softmax trick enables us to train mask predicting networks jointly with the VSR network, for learning diverse masking criterion for different layers from the training data. In the inference phase, we do not add Gumbel noise to generate the masking feature and directly generate binary masks with $f(\Delta \boldsymbol{X}_{m,n}^{t-1 \to t})$.

We apply the same mask to save computations in the $\mathtt{IIAB_{Attention}}$ part and the following $\mathtt{FFNs}$ in the feature projection part. The above masked processing strategy allows us to generate less Query tokens in the $\mathtt{IIAB_{Attention}}$ part and skip projections in the feature projection part. Let's denote $\alpha \in [0, 1]$ as the percentage of non-zeros in $\boldsymbol{M}_{m,n}^t$, our Masked Intra&Inter-frame Attention Block is able to reduce the computational complexity of each IIAB from $12HWC^2 + (4 + 12N^2)HWC$ to $(6 + 6\alpha)HWC^2 + (3\alpha N^2 + 9N^2 + 4)HWC$. The saved computations overwhelm the extra computations introduced by the tiny mask predicting networks, which only use one layer of 1×1 convolution to reduce the channel number of the feature difference tensor from C to 1.

3.4. Training Objectives

We train our network in a supervised manner. Following recent state-of-the-art approaches, we utilize the Charbonnier loss [6] $\mathcal{L}_{sr} = \sqrt{\| \hat{\boldsymbol{I}}^{HQ} - \boldsymbol{I}^{HQ} \|^2 + \varepsilon^2}$ between the estimated HR image $\hat{\boldsymbol{I}}^{HQ}$ and the ground truth image $\boldsymbol{I}^{HQ}$ to train our network; where ϵ is a constant and we set it as 10^{-3} in all of our experiments. Moreover, In order to push our mask predicting networks to mask out more positions, we apply a ℓ_1 loss on the masking features:

$$\mathcal{L}_{mask} = \frac{1}{MNZ} \sum_{m=1}^{M} \sum_{n=1}^{N} \|\mathtt{Mask}(\Delta \boldsymbol{X}_{m,n}^{t-1 \to t}))\|_1, \quad (9)$$

where Z is the number of pixels for each masking feature. Our network is trained by a combination of $\mathcal{L}_{sr}$ and $\mathcal{L}_{mask}$:

$$\mathcal{L} = \mathcal{L}_{sr} + \lambda \mathcal{L}_{mask}, \quad (10)$$

λ is a factor for masking ratio adjustment. More details can be found in our experimental section.

4. Experiments

4.1. Experimental Settings

Following the experimental settings of recently proposed VSR methods [2, 5, 21, 22, 32], we evaluate our MIA-VSR model on the REDS [28], Vimeo90K [40] and the commonly used Vid4 [24] datasets. We train two models on the REDS dateset and the Vimeo90K dataset, respectively. The model trained on the REDS dataset is used for evaluating the REDS testing data and the Viemo90K model is used for evaluating the Vimeo90K and Vid4 testing data. For our proposed adaptive mask prediction module, we fine-tune it with the pre-trained VSR model for another 100K iterations. We implement our model with Pytorch and train our models with RTX 4090 GPUs. The respective hyper-parameters used for ablation study and comparison with state-of-the-art methods will be introduced in the following subsections.

Table 1. Ablation studies on the processing blocks and the proposed adaptive mask prediction module. More details can be found in our ablation study section.

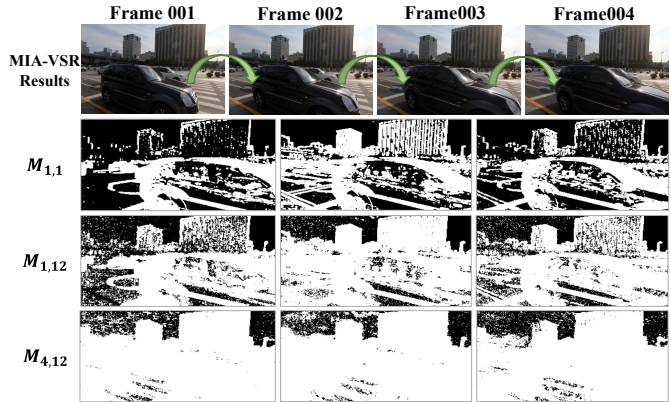| Model	λ	Params (M)	REDS4		
			PSNR	SSIM	FLOPs(G)
MFSAB-VSR	-	6.41	31.03	0.8965	871.59
IIAB-VSR	-	6.34	31.12	0.8979	518.92
IIAB-VSR + HM	-	6.34	29.83	0.8390	420.53
MIA-VSR	1e-4	6.35	31.11	0.8978	506.28
MIA-VSR	3e-4	6.35	31.07	0.8972	469.78
MIA-VSR	5e-4	6.35	31.01	0.8966	442.32
MIA-VSR	1e-3	6.35	30.76	0.8773	426.84

Figure 4. Visualization of predicted masks for a sequence in the REDS dataset.

4.2. Ablation Study

The effectiveness of IIAB In order to show the advantages of the proposed IIAB, we firstly compare the proposed intra&inter-frame attention block (IIAB) with the multi-frame self-attention block (MFSAB) which was adopted in PSRT [32]. We use 6 IIAB or MFSAB blocks to build feature propagation modules and instantialize VSR models with 4 feature propagation modules. The channel number, window size and the head number for the two models are set as 120, 8 and 6, respectively. We denote the two VSR models as IIA-VSR, MFSA-VSR, and train the two models with 6 frame training data for 300K iterations. The super-resolution results of the two models are shown in Table3. The number of parameters and computational consumption (FLOPs) of the two models are also reported for reference. The number of FLOPs is calculated on the REDS dataset, which super-resolve 180×320 video sequences to a resolution of 720×1280, we report the average FLOPs for each frame of processing. As shown in Table 1, it is clear that the proposed IIAB could generate better VSR results with less computational consumption than the MFSAB block.

Table 2. Quantitative comparison (PSNR/SSIM) on the REDS4[28], Vimeo90K-T[40] and Vid4[24] dataset for 4× video super-resolution task. For each group of experiments, we color the best and second best performance with red and blue, respectively.

| Method	Frames REDS/Vimeo	REDS4			Vimeo-90K-T			Vid4		
		PSNR	SSIM	FLOPs	PSNR	SSIM	FLOPs	PSNR	SSIM	FLOPs
TOFlow[40]	5/7	27.98	0.7990	-	33.08	0.9054	-	25.89	0.7651	-
EDVR[36]	5/7	31.09	0.8800	2.95	37.61	0.9489	0.367	27.35	0.8264	1.197
MuCAN[19]	5/7	30.88	0.8750	1.07	37.32	0.9465	0.135	-	-	-
VSR-T[2]	5/7	31.19	0.8815	-	37.71	0.9494	-	27.36	0.8258	-
PSRT-sliding[32]	5/-	31.32	0.8834	1.66	-	-	-	-	-	-
VRT[21]	6/-	31.60	0.8888	1.37	-	-	-	-	-	-
PSRT-recurrent[32]	6/-	31.88	0.8964	2.39	-	-	-	-	-	-
MIA-VSR(ours)	6/-	32.01	0.8997	1.50	-	-	-	-	-	-
BasicVSR[3]	15/14	31.42	0.8909	0.33	37.18	0.9450	0.041	27.24	0.8251	0.134
IconVSR[3]	15/14	31.67	0.8948	0.51	37.47	0.9476	0.063	27.39	0.8279	0.207
TTVSR[25]	**50/-**	32.12	0.9021	0.61	-	-	-	-	-	-
BasicVSR++[5]	**30**/14	32.39	0.9069	0.39	37.79	0.9500	0.049	27.79	0.8400	0.158
VRT[21]	16/7	32.19	0.9006	1.37	38.20	0.9530	0.170	27.93	0.8425	0.556
RVRT[22]	**30**/14	32.75	0.9113	2.21	38.15	0.9527	0.275	27.99	0.8462	0.913
PSRT-recurrent[32]	16/14	32.72	0.9106	2.39	38.27	0.9536	0.297	28.07	0.8485	0.970
MIA-VSR(ours)	16/14	32.78	0.9220	1.61	38.22	0.9532	0.204	28.20	0.8507	0.624
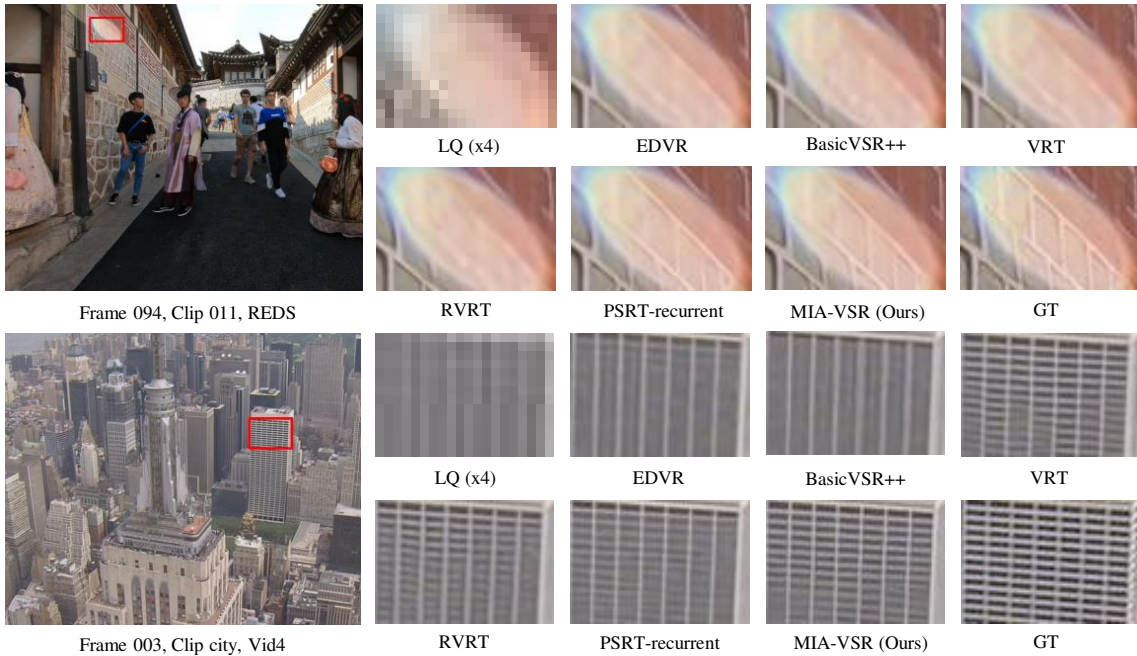
Figure 5. Visual comparison for 4× VSR on REDS4 dataset and Vid4 dataset.

Masked processing strategy In this part, we validate the effectiveness of the proposed masked processing strategy. Our basic experimental settings are the same as our experiment to validate IIAB. We firstly present experimental results to test whether we could get good experimental results by setting a uniform threshold to generate masks based on feature differences. We denote the model as IIAB-VSR + HM and set the threshold value as 0.2, where HM is the abbreviation for the handcrafted mask.

As can be found in Table 1, adopting a uniform threshold

for different layers will lead to a significant accuracy drop of the VSR model. Then we validate the effectiveness of our proposed adaptive mask prediction module and analyze the effects of different λ values in Eq.10. Concretely, we set λ as $1e-4$, $3e-4$, $5e-4$ and $1e-3$ and train four different models. Generally, setting a larger weight for the sparsity loss could push the network to mask our more computations but also results in less accurate VSR results. By setting λ as $5e-4$, we could further save approximately 20% computations from the highly efficient IIAB-VSR model without a

significant performance drop. The much better trade-off between accuracy and computation by our MIA-VSR model over IIA-VSR + Handcraft Mask clearly validated the superiority of our adaptive mask predicting network. Some visual examples of the masks generated in our MIA-VSR model can be found in Fig.4. The network tends to skip a large portion of computations in the early stage and could mask fewer positions for deeper layers.

4.3. Comparison with State-of-the-art Methods

In this subsection, we compare the proposed MIA-VSR model with state-of-the-art VSR methods. We compare the MIA-VSR with representative sliding-window based methods TOFlow [40], EDVR [36], MuCAN [19], VSR-T [2], VRT [21], RVRT [22] and representative recurrent-based methods BasicVSR [3], BasicVSR++ [5], TTVSR [25], PSRT-recurrent [32]; among which VRT, RVRT, TTVSR and PSRT-recurrent are Transformer-based approaches and the other approaches are CNN-based models.

In order to achieve comparable VSR results with state-of-the-art methods, we instantiate our MIA-VSR model with 4 feature propagation modules and each feature propagation module contains 24 MIIA blocks. Among them, we set the interval of skip connections to [6,6,6,6]. The spatial window size, head size and channel size are set to 8×8, 6 and 120 accordingly. The number of parameters in our model is on par with the recent state-of-the-art methods RVRT [22] and PSRT-recurrent [32]. Previous works [21, 22, 25, 30, 32] have figured out that models trained with longer sequences could achieve better VSR results. For a fair comparison, we follow the experimental setting of [21, 22, 32] and train the VSR model with short sequences (6 frames from the REDS dataset) and long sequences (16/14 frames from the REDS dataset and the Vimeo-90K dataset).

The VSR results by different methods can be found in Table 2. As the number of FLOPs(T) for our MIA-VSR model is related to the content of video sequences, we therefore report the average per frame FLOPs on different datasets. In terms of the results trained with short sequences, the proposed MIA-VSR outperforms the compared methods by a large margin. Our model improves the PSNR of the state-of-the-art PSRT-recurrent model by 0.13 dB with a reduction in the number of FLOPs of almost 40%. As for the models trained with longer training sequences, our MIA-VSR still achieves a better trade-off between VSR accuracy and efficiency over recent state-of-the-art approaches. With more than 40% less computations on the RVRT and PSRT-recurrent approaches, our model achieved the best VSR results on the REDS and Vid4 datasets and the second-best results on the Vimeo-90K dataset. Some visual results from different VSR results can be found in Fig.5, our proposed MIA-VSR method is able

Table 3. Comparison of model size, testing memory and complexity of different VSR models on the REDS[28] dataset.

Model	Params(M)	FLOPs(T)	Memory(M)	Runtime(ms)	PSNR(dB)
VRT	35.6	1.37	2149	888	32.19
RVRT*	10.8	2.21	1056	473	32.75
PSRT	13.4	2.39	190	1041	32.72
MIA-VSR	16.5	1.61	206	822	32.78

* means that uses customized CUDA kernels for better performance.

to recover more natural and sharp textures from the input LR video sequences.

In our supplementary file, we also provide other instantializations of our proposed MIA-VSR model to show the scalability of our approach and compare our method with efficient VSR models. More visual examples are also provided for different datasets.

4.4. Complexity and Memory Analysis

In this part, we compare the complexity and memory footprint of different Transformer-based state-of-the-art VSR methods. In Table 3, we report the number of parameters, the peak GPU memory consumption, the number of FLOPs, the Runtime and the PSNR by different VSR methods. In comparison with other Transformer-based VSR methods, our MIA-VSR method has a similar number of parameters and requires much less number of FLOPs for processing the video sequence. In addition, the peak GPU memory consumption, which is critical factor for deploying model on terminal equipment, by our model is much less than the VRT and RVRT approaches. As for the runtime, our model is not as fast as RVRT, because the authors of RVRT have implemented the key components of RVRT with customized CUDA kernels. As the acceleration and optimization of Transformers are still to be studied, there is room for further optimization of the runtime of our method by our relatively small FLOPs.

5. Conclusion

In this paper, we proposed a novel Transformer-based recurrent video super-resolution model, named MIA-VSR. We proposed a masked processing framework to leverage the temporal continuity between adjacent frames to save computations for the video super-resolution model. An Intra-frame and Inter-frame attention block is proposed to make better use of previously enhanced features to provide supplementary information; and an adaptive mask prediction module is developed to generate block-wise masks for each stage of processing. Furthermore, we evaluated our MIA-VSR model on various benchmark datasets. Our model is able to achieve state-of-the-art video super-resolution results with less computational resources.

References

[1] Jose Caballero, Christian Ledig, Andrew Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4778–4787, 2017. 2

[2] Jiezhang Cao, Yawei Li, Kai Zhang, Jingyun Liang, and Luc Van Gool. Video super-resolution transformer. *arXiv preprint arXiv:2106.06847*, 2021. 2, 6, 7, 8

[3] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. *arXiv preprint arXiv:2012.02181*, 2020. 3, 7, 8

[4] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. 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*, pages 4947–4956, 2021. 1

[5] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5972–5981, 2022. 1, 3, 4, 6, 7, 8

[6] Pierre Charbonnier, Laure Blanc-Feraud, Gilles Aubert, and Michel Barlaud. Two deterministic half-quadratic regularization algorithms for computed imaging. In *Proceedings of 1st international conference on image processing*, pages 168–172. IEEE, 1994. 6

[7] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, pages 764–773, 2017. 2

[8] Dario Fuoli, Shuhang Gu, and Radu Timofte. Efficient video super-resolution through recurrent latent space propagation. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3476–3485. IEEE, 2019. 1, 2, 3

[9] Shuhang Gu, Yawei Li, Luc Van Gool, and Radu Timofte. Self-guided network for fast image denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2511–2520, 2019. 4

[10] Amirhossein Habibian, Davide Abati, Taco S Cohen, and Babak Ehteshami Bejnordi. Skip-convolutions for efficient video processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2695–2704, 2021. 2, 3, 5

[11] Amirhossein Habibian, Haitam Ben Yahia, Davide Abati, Efstratios Gavves, and Fatih Porikli. Delta distillation for efficient video processing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 213–229. Springer, 2022. 3

[12] Ping Hu, Fabian Caba, Oliver Wang, Zhe Lin, Stan Sclaroff, and Federico Perazzi. Temporally distributed networks for fast video semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8818–8827, 2020. 3

[13] Takashi Isobe, Xu Jia, Shuhang Gu, Songjiang Li, Shengjin Wang, and Qi Tian. Video super-resolution with recurrent structure-detail network. *arXiv preprint arXiv:2008.00455*, 2020. 1, 3

[14] Takashi Isobe, Songjiang Li, Xu Jia, Shanxin Yuan, Gregory Slabaugh, Chunjing Xu, Ya-Li Li, Shengjin Wang, and Qi Tian. Video super-resolution with temporal group attention. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[15] Takashi Isobe, Xu Jia, Xin Tao, Changlin Li, Ruihuang Li, Yongjie Shi, Jing Mu, Huchuan Lu, and Yu-Wing Tai. Look back and forth: Video super-resolution with explicit temporal difference modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17411–17420, 2022. 1

[16] Samvit Jain, Xin Wang, and Joseph E Gonzalez. Accel: A corrective fusion network for efficient semantic segmentation on video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8866–8875, 2019. 3

[17] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 5

[18] Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim. 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 (CVPR)*, pages 3224–3232, 2018. 2

[19] Wenbo Li, Xin Tao, Taian Guo, Lu Qi, Jiangbo Lu, and Jiaya Jia. Mucan: Multi-correspondence aggregation network for video super-resolution. *arXiv preprint arXiv:2007.11803*, 2020. 1, 2, 7, 8

[20] Yule Li, Jianping Shi, and Dahua Lin. Low-latency video semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5997–6005, 2018. 3

[21] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer. *arXiv preprint arXiv:2201.12288*, 2022. 1, 2, 6, 7, 8

[22] Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc V Gool. Recurrent video restoration transformer with guided deformable attention. *Advances in Neural Information Processing Systems*, 35:378–393, 2022. 1, 2, 4, 5, 6, 7, 8

[23] Lijian Lin, Xintao Wang, Zhongang Qi, and Ying Shan. Accelerating the training of video super-resolution models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1595–1603, 2023. 1

[24] Ce Liu and Deqing Sun. On bayesian adaptive video super resolution. *IEEE transactions on pattern analysis and machine intelligence*, 36(2):346–360, 2013. 1, 6, 7

[25] Chengxu Liu, Huan Yang, Jianlong Fu, and Xueming Qian. Learning trajectory-aware transformer for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5687–5696, 2022. 1, 3, 7, 8

[26] Ding Liu, Zhaowen Wang, Yuchen Fan, Xianming Liu, Zhangyang Wang, Shiyu Chang, and Thomas Huang. Robust video super-resolution with learned temporal dynamics. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2507–2515, 2017. 2

[27] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[28] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. 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*, pages 0–0, 2019. 6, 7, 8

[29] Xuecheng Nie, Yuncheng Li, Linjie Luo, Ning Zhang, and Jiashi Feng. Dynamic kernel distillation for efficient pose estimation in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6942–6950, 2019. 3

[30] Zhongwei Qiu, Huan Yang, Jianlong Fu, and Dongmei Fu. Learning spatiotemporal frequency-transformer for compressed video super-resolution. In *European Conference on Computer Vision*, pages 257–273. Springer, 2022. 1, 8

[31] Mehdi SM Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-recurrent video super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6626–6634, 2018. 2

[32] Shuwei Shi, Jinjin Gu, Liangbin Xie, Xintao Wang, Yujiu Yang, and Chao Dong. Rethinking alignment in video super-resolution transformers. *arXiv preprint arXiv:2207.08494*, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[33] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 4

[34] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4472–4480, 2017. 2

[35] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3360–3369, 2020. 1, 2

[36] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 0–0, 2019. 1, 2, 7, 8

[37] Bin Xia, Jingwen He, Yulun Zhang, Yucheng Hang, Wenming Yang, and Luc Van Gool. Residual sparsity connection learning for efficient video super-resolution. *arXiv preprint arXiv:2206.07687*, 2022. 3

[38] Zeyu Xiao, Xueyang Fu, Jie Huang, Zhen Cheng, and Zhiwei Xiong. Space-time distillation for video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2113–2122, 2021. 1

[39] Kai Xu, Ziwei Yu, Xin Wang, Michael Bi Mi, and Angela Yao. An implicit alignment for video super-resolution. *arXiv preprint arXiv:2305.00163*, 2023. 1, 2

[40] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127(8): 1106–1125, 2019. 6, 7, 8

[41] Xizhou Zhu, Yuwen Xiong, Jifeng Dai, Lu Yuan, and Yichen Wei. Deep feature flow for video recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2349–2358, 2017. 3

[42] Xizhou Zhu, Jifeng Dai, Lu Yuan, and Yichen Wei. Towards high performance video object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7210–7218, 2018. 3