MambaOVSR: Multiscale Fusion with Global Motion Modeling for Chinese Opera Video Super-Resolution

Hua Chang[1], **Xin Xu**[1,2*], **Wei Liu**[1], **Wei Wang**[2], **Xin Yuan**[2], **Kui Jiang**[3]

[1] School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065 China
[2] Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System
[3] School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001 China
{changhua,xuxin,liuwei,wangwei8,xinyuan}@wust.edu.cn, jiangkui@hit.edu.cn

Abstract

Chinese opera is celebrated for preserving classical art. However, early filming equipment limitations have degraded videos of last-century performances by renowned artists (*e.g.*, low frame rates and resolution), hindering archival efforts. Although space-time video super-resolution (STVSR) has advanced significantly, applying it directly to opera videos remains challenging. The scarcity of datasets impedes the recovery of high-frequency details, and existing STVSR methods lack global modeling capabilities—compromising visual quality when handling opera's characteristic large motions. To address these challenges, we pioneer a large-scale Chinese Opera Video Clip (COVC) dataset and propose the **Mamba**-based multiscale fusion network for space-time **O**pera **V**ideo **S**uper-**R**esolution (MambaOVSR). Specifically, MambaOVSR involves three novel components: the Global Fusion Module (GFM) for motion modeling through a multiscale alternating scanning mechanism, and the Multiscale Synergistic Mamba Module (MSMM) for alignment across different sequence lengths. Additionally, our MambaVR block resolves feature artifacts and positional information loss during alignment. Experimental results on the COVC dataset show that MambaOVSR significantly outperforms the SOTA STVSR method by an average of 1.86 dB in terms of PSNR.

Introduction

Chinese opera represents a distinctive performing art of significant cultural value. However, limitations in early filming technology and storage media degradation (Jiang et al. 2022) have left many classic recordings with low resolution and frame rates, severely hindering preservation efforts and scholarly study (Chung 2024).

Space-Time Video Super-Resolution (STVSR), first proposed in 2020 (Xiang et al. 2020), enhances both temporal and spatial video resolution. Early approaches combined Video Frame Interpolation (VFI) (Cheng and Chen 2021; Liu et al. 2024a) and Video Super-Resolution (VSR) (Yi et al. 2019; Li et al. 2023) techniques but failed to exploit their intrinsic connections, yielding suboptimal results (Haris, Shakhnarovich, and Ukita 2020; Hu et al. 2023b). Subsequent end-to-end frameworks (Xiang et al. 2020; Xu et al. 2021; Jiang et al. 2021; Wang et al. 2023) and efficiency-focused designs (Geng et al. 2022; Hu et al. 2022, 2023a) improved STVSR for general scenes. However, these methods remain inadequate for opera due to the lack of domain-specific datasets and insufficient global modeling capabilities (Dang et al. 2023, 2024a). As shown in Figure 1(a), models trained on general datasets (e.g., Vimeo90K (Xue et al. 2019)) fail to recover opera-specific high-frequency textures, highlighting the domain gap.

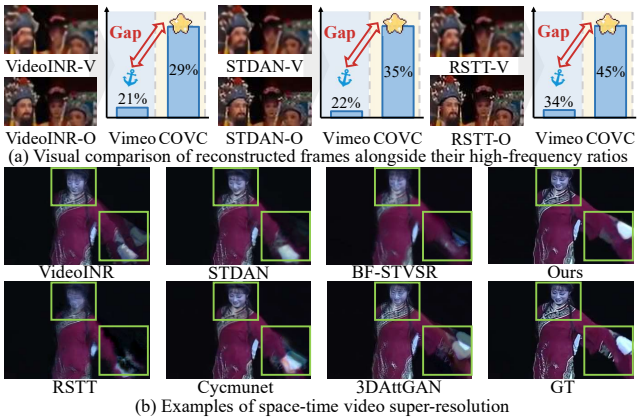

Figure 1: (a) Visual comparison and high-frequency content ratios for the same model trained on Vimeo90K (-V) and COVC (-O). The COVC-trained model recovers more high-frequency details. (b) Presents that existing methods synthesize intermediate frames with blurring artifacts.

To tackle the aforementioned challenges, we introduce the Chinese Opera Video Clip (COVC) dataset, the first large-scale collection for opera restoration. COVC contains 33 distinct opera videos processed into training septuples following the Vimeo90K dataset (Xue et al. 2019), yielding 104,138 training samples. When retraining existing STVSR methods on COVC, the synthesized frames exhibit blurring artifacts (Figure 1(b)), confirming their inability to model the large motions in opera videos.

In this paper, we propose MambaOVSR, a **Mamba**-based multiscale fusion network for space-time **O**pera **V**ideo **S**uper-**R**esolution, which effectively addresses large motion modeling (Figure 1(b)). Specifically, our framework features

*Corresponding author.

three innovations: **Global Fusion Module (GFM)**, **MambaVR Block** and **Multiscale Synergistic Mamba Module (MSMM)**. GFM synthesizes intermediate frames by blending forward/backward predictions. Each direction employs a pyramid structure with a Multiscale Alternate Scanning Mechanism (MASM) for global multiscale modeling of adjacent frames, followed by 3D convolutions to extract temporal features from interpolated short sequences. The MambaVR block is designed to resolve feature artifacts and positional information loss in Vision Mamba alignment. MSMM leverages MambaVR blocks for granular motion alignment across varying sequence lengths.

Our contributions are summarized as follows:

- We pioneer a large-scale Chinese Opera Video Clip (COVC) dataset and propose the **Mamba**-based multiscale fusion network for space-time **O**pera **V**ideo **S**uper-**R**esolution (MambaOVSR).
- We propose the GFM to perform fine-grained holistic modeling of motion between adjacent frames, accurately synthesizing missing intermediate features. Complementarily, a 3D convolution-based module exploits the temporal feature of neighboring frames for refinement.
- We introduce the MambaVR block for global spatial alignment of multi-frame features. Then, MSMM performs multi-scale alignment on sequences of varying lengths, effectively handling large motions.
- We conduct extensive experiments on both the COVC and general Vimeo90K, demonstrating that the proposed MambaOVSR markedly outperforms existing STVSR methods in both quantitative and qualitative evaluations.

Related Work

Space-Time Video Super-Resolution

The Space-Time Video Super-Resolution (STVSR) aims to enhance both the spatial and temporal resolution of videos. Compared to the sequentially combined Video Super-Resolution (VSR) and Video Frame Interpolation (VFI) methods (Zhou et al. 2021), the jointly optimized framework has smaller parameters and better results (Xiang et al. 2020). STARnet (Haris, Shakhnarovich, and Ukita 2020) used high- and low-resolution features to synthesize missing intermediate frames. ZSM (Xiang et al. 2020) combined deformable convolution with ConvLSTM to propagate frame information. Based on this, TMNet (Xu et al. 2021) implemented arbitrary time-step frame interpolation. Very recently, Cycmunet (Hu et al. 2023b) and STDAN (Wang et al. 2023) proposed innovative up-and-down projection units (UPU&DPU) and deformable feature aggregation (DFA) to achieve frame alignment. Furthermore, to improve the inference speed, RSTT (Geng et al. 2022) proposed an overall model based on Swin Transformer (Liu et al. 2021). Although these methods perform well on general scene videos, their performance on opera videos is suboptimal due to richer texture details and larger motions.

Visual Mamba

Due to its linear complexity and efficient selection mechanism, Mamba (Gu and Dao 2023) has achieved impressive results in natural language processing (NLP). VisionMamba (Zhu et al. 2024) and VMamba (Liu et al. 2024b) pioneered the application of Mamba in computer vision by using distinct scanning methods to process images. VideoMamba (Li et al. 2025) extended Mamba to video understanding by incorporating spatial and temporal position embedding. Furthermore, Video Mamba Suite (Chen et al. 2024) explored the role of Mamba in the four phases of video understanding, highlighting its advantages in video handling. VFIMamba (Zhang et al. 2024) achieves SOTA performance in video frame interpolation (VFI) by modeling adjacent frames through an alternating scanning mechanism (ASM). However, ASM focuses only on global motion and cannot model local motion variations, and we propose the Multiscale Alternate Scanning Mechanism (MASM) to model adjacent frame features. For alignment, the original VideoMamba block (Li et al. 2025) is limited by feature artifacts and flexibility. To address this problem, we propose MambaVR for global alignment of frames.

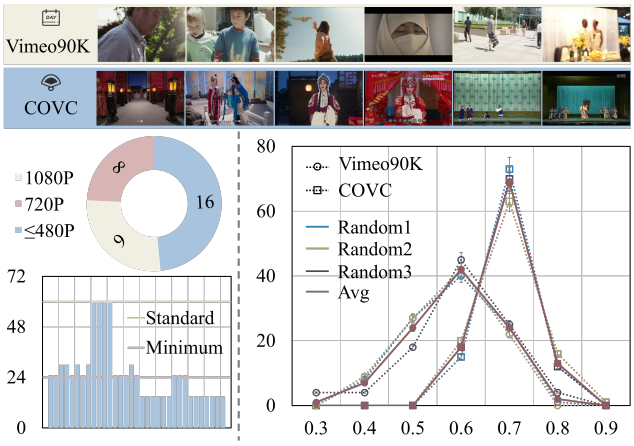

Figure 2: Comparison of samples and statistics.

Proposed Method

Chinese Opera Video Clip Dataset

Low-quality opera videos hinder the art's preservation and evolution, with their elaborate costumes, sets, and props producing far richer textures than general benchmarks (e.g., Vimeo90K (Xue et al. 2019)), causing existing models to fail in this domain (see Figure 1(a)). To address this, we introduce COVC: a large-scale Chinese opera video clip dataset.

To ensure dataset quality, we curated 33 high-quality opera videos based on bitrate, resolution, and subjective clarity, and extracted continuous frames while omitting all-black boundary frames to avoid invalid PSNR measurements. Following the Vimeo90K protocol (Xue et al. 2019), every seven consecutive frames form one clip, yielding 115,548 clips (11,410 for testing and the remainder for training). The test set is stratified by visual quality into High (5,120 clips), Medium (3,150 clips), and Low (3,140 clips).

The COVC dataset comprises 33 opera videos: 9 at 1080p, 8 at 720p, and 16 at $\leq$480p (see Figure 2). The frame rates cluster predominantly around 24 fps, meeting the minimum

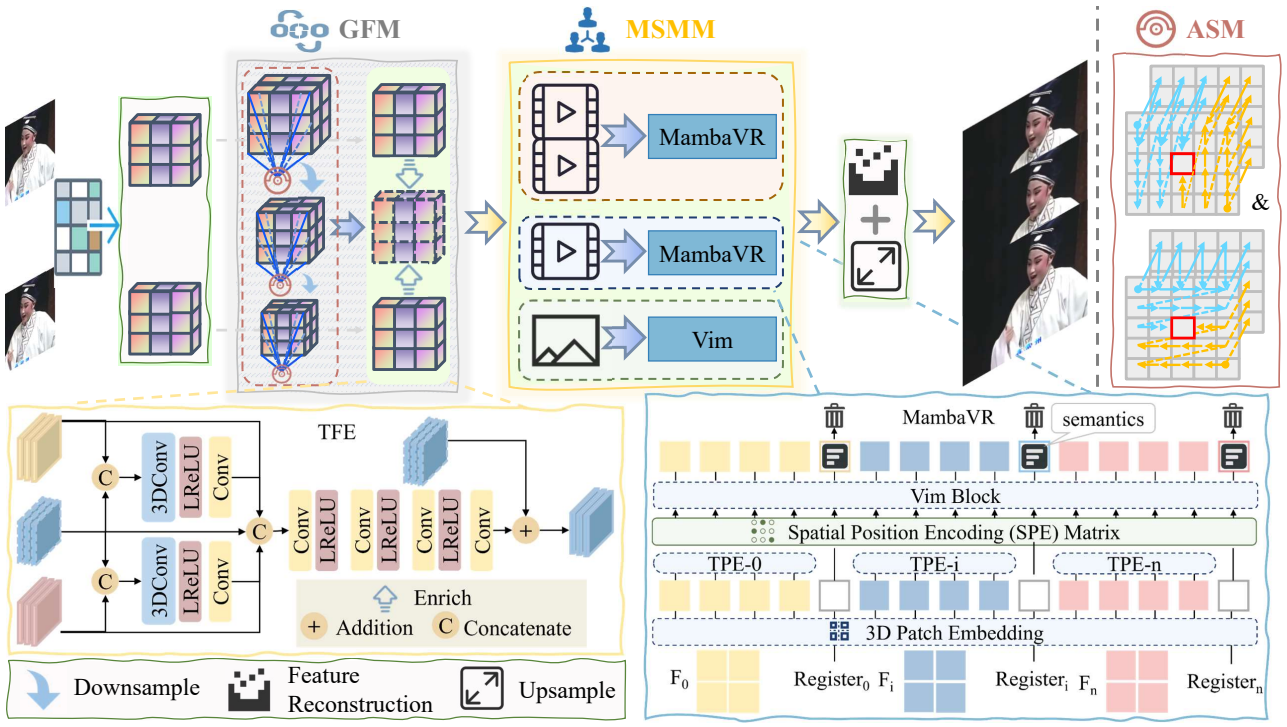

Figure 3: Architecture of the proposed Mamba-Based multiscale fusion network. Firstly, the features are extracted, and the missing intermediate frame features are obtained by the Global Fusion Module (GFM) with a Multiscale Alternate Scanning Mechanism (MASM). Next, each frame feature is enhanced by aligning sequences of different lengths using the Multiscale Synergistic Mamba Module (MSMM). Finally, high-quality video is obtained by feature reconstruction and PixelShuffle.

cinematic standard (Tag et al. 2016), while four clips reach 60 fps (Mackin et al. 2017). The top panel of Figure 2 presents representative frames from COVC and Vimeo90K (Xue et al. 2019). Whereas Vimeo90K videos primarily depict general scenes, opera clips with their elaborate makeup, costumes, and stage settings exhibit richer high-frequency textures. This distinction is illustrated by the line chart in Figure 2 (bottom right). Since COVC contains 1.6 times more clips than Vimeo90K, we randomly sampled three sets of 100-frame sequences from both datasets to quantify this difference. For each set, we computed the per-frame proportion of high-frequency information and plotted the results together with their three-trial average. Vimeo90K frames contain approximately 65% high-frequency content, while COVC frames contain around 75%., with even the lowest value exceeding 55%. These findings demonstrate that Chinese opera videos exhibit higher texture complexity. Furthermore, since existing methods fail to model the large motions in opera videos (see Figure 1(b)), we propose a Mamba-based multiscale fusion network.

Network Overview

The proposed Mamba-based multiscale fusion network, shown in Figure 3, aims to generate high-resolution (HR) and high-frame-rate (HFR) opera frames $I^H = \{I_t^H\}_{t=1}^{2n+1}$ of size $3 \times nH \times nW$, where n is the spatial upsampling factor, from a low-resolution (LR) and low-frame-rate (LFR) sequence $I^L = \{I_{2t-1}^L\}_{t=1}^{n+1}$. First, the feature extractor, composed of a convolution layer and five residual blocks, extracts frame features $F^L = \{F_{2t-1}^L\}_{t=1}^{n+1}$. These are passed to the Global Fusion Module (GFM) to generate missing intermediate features $F^L = \{F_{2t}^L\}_{t=1}^{n}$. Next, the complete sequence is fed into the Multiscale Synergistic Mamba Module (MSMM) to obtain enhanced high-quality (HQ) features $F^H = \{F_t^H\}_{t=1}^{2n+1}$. Finally, the reconstruction and PixelShuffle (Shi et al. 2016) modules produce the HQ output frames $I^H = \{I_t^H\}_{t=1}^{2n+1}$.

Global Fusion Module

Deformable Convolution (DConv) (Dai et al. 2017) enables effective alignment by dynamically sampling spatial locations in a feature map. ZSM (Xiang et al. 2020) first leveraged DConv for synthesizing missing intermediate frames, yielding promising results, and subsequent methods have widely adopted DConv for this purpose (Xu et al. 2021; Jiang et al. 2024; Wang et al. 2023). However, the fixed kernel size of DConv limits its receptive field, degrading performance on sequences with large motions. More recently, Mamba (Gu and Dao 2023) has emerged as an efficient global modeling framework in computer vision; VFIMamba (Zhang et al. 2024) applied Mamba to VFI, but it still struggles to capture fine-grained motion variations.

Global Fusion Module. Inspired by this, we propose

the Global Fusion Module (GFM), which employs a Multiscale Alternate Scanning Mechanism (MASM) to globally model adjacent frame features and accurately capture large inter-frame motions. To synthesize intermediate frames, we fuse predictions from both forward and backward directions, wherein each direction learns global-to-local motion offsets via a multiscale pyramid architecture. Taking the forward synthesis direction ($0 \rightarrow t$) as an example, we first downsample the neighboring frame features to multiple scales. At each scale, we merge the preceding and succeeding frame features into a single large feature map and globally arrange corresponding pixels along four directions (as shown in Figure 3, top-right). We then compute the motion offset $H^{(N)}_{i,0\rightarrow t}$ by modeling pixel displacements between F^L_{i-1} and F^L_{i+1}, and fuse this offset with the succeeding frame features to obtain the predicted intermediate frame at the current scale, $F^{(N)}_{i,0\rightarrow t}$. Finally, both the small-scale offset and its predicted intermediate frame are upsampled and integrated with larger-scale predictions to yield the final multiscale fusion result, $F^L_{i,0\rightarrow t}$.

$$H^{(N)}_{i,0\rightarrow t} \;=\; MASM^{(N)}_{0\rightarrow t}\big(\downarrow^{(N)}(F^L_{i-1}),\ \downarrow^{(N)}(F^L_{i+1})\big), \quad (1)$$

$$F^{(N)}_{i,0\rightarrow t} = Fuse\left(H^{(N)}_{i,0\rightarrow t}, \downarrow^{(N)}(F^L_{i+1})\right), \quad (2)$$

$$F^L_{i,0\rightarrow t} = Fuse\big(\uparrow^{(N\rightarrow N-1)}\big(F^{(N)}_{i,0\rightarrow t}\big),\ F^{(N-1)}_{i,0\rightarrow t}\big), \quad (3)$$

where $i-1$, i, $i+1$ denote three consecutive frames; N is the number of layers in the multiscale pyramid. $\downarrow$ indicates down-sampling and $\uparrow$ indicates up-sampling.

Finally, we fuse the forward and backward predictions to produce the final intermediate frame features. While the GFM module generates a complete frame sequence, the initially synthesized intermediate frames may exhibit minor artifacts under large-motion conditions.

Temporal Feature Enhancement. To refine the initially synthesized intermediate-frame features, we introduce the Temporal Feature Enhancement (TFE) module (see Figure 3, bottom-left). TFE concatenates the preceding, current, and succeeding frame features and processes them through a 3DConv–ReLU block to extract bidirectional motion offsets. These offsets are then concatenated with the original frame trio and passed through a multi-layer convolutional blending network to yield refined intermediate features. Finally, we add the original intermediate features to the refined output to produce the enhanced frame representation. By leveraging adjacent frames for local convolutional refinement, TFE recovers fine-grained details and improves the alignment information available to subsequent modules.

Multiscale Synergistic Mamba Module

Using the GFM, we generate high-frame-rate video sequence features. Most existing methods rely on either the pass-through or sliding-window approach for global frame sequence alignment (Xiang et al. 2020; Xu et al. 2021; Wang et al. 2023). The pass-through method accumulates alignment errors (Dang et al. 2025), which progressively affect subsequent frames, while sliding-window approaches are limited to a fixed temporal neighborhood and cannot capture long-range dependencies. Although 3D convolution can achieve global alignment by concatenating multiple frames (Fu et al. 2024; Jiang et al. 2019), it is applicable only to short sequences and captures limited temporal information. VideoMamba (Li et al. 2025) has emerged as a potential alternative, but we found that its inherent feature artifacts and fixed-position encoding render it unsuitable.

Mamba for Video Restoration (MambaVR). To address these challenges, we propose the MambaVR block specifically for video restoration (see Figure 3, bottom right). First, in each frame's feature map, we uniformly insert a fixed number of blank register tokens to buffer high-norm semantic activations that could introduce feature artifacts. Video restoration demands strict local consistency, unlike Mamba-R's (Wang et al. 2024) use of VisionMamba (Zhu et al. 2024) to encode high-level semantics into background regions for classification; MambaVR isolates those semantics in removable tokens and discards them during reconstruction to preserve fine-grained structure.

Second, we introduce Flexible Rotary Position Embedding (F-RoPE) to overcome VisionMamba's fixed, depth-attenuating embeddings (Zhu et al. 2024). F-RoPE extends RoPE (Su et al. 2024) by generating relative spatial encodings on-the-fly for any input resolution. It constructs base frequency tensors for the input dimensions ($D \times H \times W$), transforms them into a Spatial Position Encoding (SPE) matrix, and injects precise positional cues into the self-attention mechanism via element-wise multiplication.

$$\omega_i = [\pi \cdot \frac{i}{2}], \quad i = 1, 2, \ldots, \frac{D}{2}, \quad (4)$$

$$\mathbf{f}_h(u) = [\, u\,\omega_i \,]^{D/2}_{i=1},\ \mathbf{f}_w(v) = [\, v\,\omega_i \,]^{D/2}_{i=1}, \quad (5)$$

$$\mathrm{SPE}(u,v) = \mathrm{broadcast}\big(\mathbf{f}_h(u),\ \mathbf{f}_w(v)\big) \in \mathbb{R}^D, \quad (6)$$

where $u = 0, \ldots, H-1,\ v = 0, \ldots, W-1$. broadcast is the original broadcasting mechanism.

Multiscale Synergistic Mamba Module. To achieve multi-granularity motion alignment (Dang et al. 2024c,b), we propose the Multiscale Synergistic Mamba Module (MSMM), built upon our MambaVR block and Vim (Zhu et al. 2024) (see Figure 3). **Global implicit alignment:** Feed the full sequence into MambaVR for holistic feature interaction. **Short-term temporal consistency:** Apply a sliding window over segments to preserve local motion coherence. **Global guidance enhancement:** Use MambaVR's hidden state to update Vim's, enriching each frame with global context. As an example of global alignment, the full sequence features are concatenated and passed through a 3D convolution to generate successive temporal patches of length L. Then, we uniformly insert n register tokens (r) into the sequence, and the temporal position encoding (TPE) $P_t \in R^{T\times C}$ is added.

$$X \;=\; [\cdots x_i, r_1, \cdots x_{2i}, r_2, \cdots x_{ni}, r_n, \cdots x_L] + P_t, \quad (7)$$

where T denotes the sequence length, C the channel dimension. Next, we apply element-wise multiplication with the Spatial Position Encoding $\mathrm{SPE}(u,v)$ and feed the result into the MambaVR block to obtain globally aligned features E_g:

$$X \;= X \otimes \mathrm{SPE}(u,v),\ E_g = \mathrm{MambaVR}\,(X)\,. \quad (8)$$

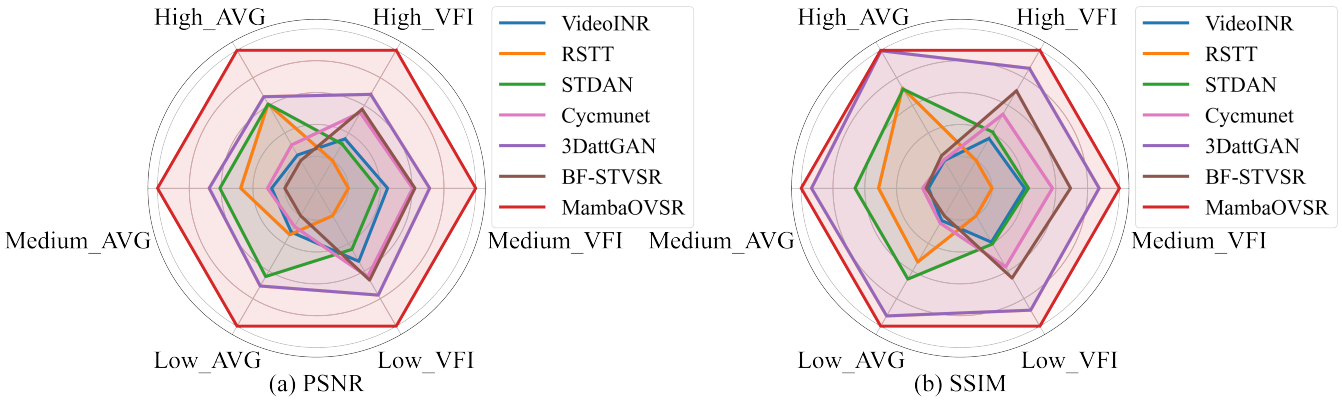

Figure 4: Quantitative comparison with the Other Space-Time Video Super-Resolution (STVSR) methods on COVC. (a) depicts a radar plot for PSNR comparisons between all generated frames (AVG) and for interpolated frames (VFI) on the three test sets, High, Medium and Low, while (b) depicts a radar plot for SSIM. Note that all metrics have been normalized.

To preserve short-term consistency, we feed three consecutive frames sequentially into distinct MambaVR blocks, yielding short-term aligned sequences E_l^j. Additionally, we initialize Vim's hidden state with that of the global MambaVR, thereby leveraging global context to guide per-frame feature enhancement.

$$L_1^{'}, \cdots, L_l^{'} = vim\left(F_1^L, \cdots, F_l^L\right). \tag{9}$$

Frame-specific supplementary information is obtained by concatenating features across multiple scales:

$$F_i^E = concat\left(E_g, E_l^j, L_i^{'}\right), \tag{10}$$

where i indexes the current frame and j denotes distinct short-term contexts.

Leveraging residual connections, we integrate MSMM-extracted features, refined via channel attention and projected back to the original dimensionality through a 1×1 convolution ($conv1D$), with the original frame features preserved by an initial convolution.

$$F_i^H = conv\left(F_i^L\right) + conv1D\left(attn_i\left(F_i^E\right)\right), \tag{11}$$

where $attn_i$ denotes the channel attention mechanism for the i-th frame and $conv$ a convolution layer.

Experiment

Implement Details. Even-indexed frames are downsampled 2× for input, with the seven-frame sequence as supervision. Frames are randomly cropped to 128×128, downscaled to 64×64, and augmented via flips and rotations. We train with batch size 8, using an initial learning rate of 0.01 decayed to 1×10^{-7} via cosine annealing (Gotmare et al. 2018), and optimize with AdaMax ($\beta_1 = 0.9$ and $\beta_2 = 0.999$). All experiments are implemented in PyTorch 2.1.

Datasets and Metrics. We retrained other STVSR methods on the introduced COVC and general Vimeo90K (Xue et al. 2019) dataset and quantitatively evaluated the performance of the different models using Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) (Wang et al. 2004) as evaluation metrics.

Methods	Venue	PSNR↑	SSIM↑
VideoINR	CVPR'22	20.41	0.6518
RSTT	CVPR'22	29.09	0.7996
Cycmunet	TPAMI'23	21.30	0.6532
3DAttGAN	TETCI'24	<u>30.65</u>	<u>0.8371</u>
BF-STVSR	CVPR'25	20.67	0.6762
MambaOVSR	-	**35.61**	**0.8794**

Table 1: Quantitative comparison on Vimeo90K Fast subset.

Comparison of Methods

We present a comprehensive comparison of our framework with existing STVSR methods, including VideoINR (Chen et al. 2022), RSTT (Geng et al. 2022), STDAN (Wang et al. 2023), Cycmunet (Hu et al. 2023b), 3DAttGAN (Fu et al. 2024), and BF-STVSR (Kim et al. 2025). To ensure a fair comparison, we retrained these models on the COVC.

As shown in Figure 4, we present radar plots of PSNR (a) and SSIM (b) for all generated frames (AVG) and interpolated frames (VFI) across the High, Medium, and Low test subsets. MambaOVSR achieves significant improvements in both metrics, particularly PSNR, with relative gains of 6.51%, 6.24%, and 5.24% over the SOTA 3DAttGAN on the three subsets. These results confirm the method's effectiveness in modeling large motions.

The visual comparison in Figure 5 shows that existing methods struggle with large motion, resulting in significant blurring artifacts. In contrast, our method produces fewer blurs and recovers finer details, demonstrating its effectiveness while maintaining moderate computational complexity.

To validate MambaOVSR's ability to model large motions, we compared methods on the Vimeo90K Fast test set, which is known for large motions. As shown in Table 1, MambaOVSR achieves SOTA.

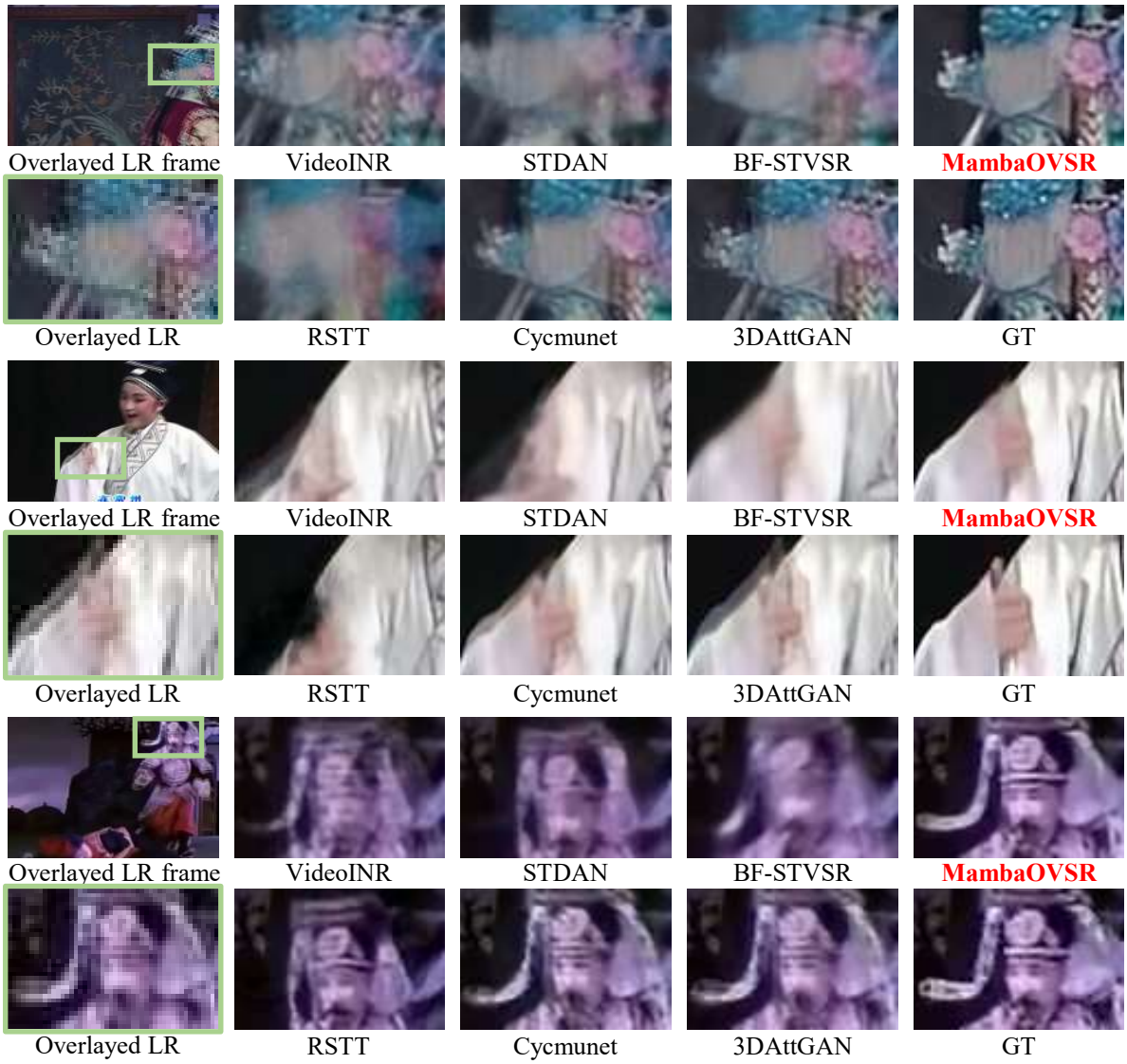

Figure 5: Qualitative Comparisons of the different approaches on three qualities of Chinese opera videos, from top to bottom, for the High, Medium and Low test sets. Our framework can recover more details while producing fewer artifacts.

Ablation Studies

To validate the effectiveness of each proposed module, we further conducted the following comprehensive ablation studies on the Medium test set.

Effectiveness of COVC. To validate the effectiveness of the COVC dataset, we trained all methods on Vimeo90K and COVC with the same configuration and compared their average performance on the Medium test set (see Figure 6). The results show that the model trained on COVC outperforms Vimeo90K in both PSNR and SSIM. MambaOVSR performs the best among all methods, with the line at the top, proving its excellent generalization ability.

Effectiveness of MSMM. To assess the effectiveness of the proposed MSMM, we designed three models: Ω_1, Ω_3, and Ω_4. Each model incorporates a deformable convolution-based module for intermediate feature interpolation. Ω_1 utilizes space-time correlation through up-and-down projections (Hu et al. 2023b), while Ω_4 employs the MSMM module to implicitly align sequences of varying lengths. To evaluate the effectiveness and efficiency of the Mamba framework in video modeling, we replaced the MambaVR and Vim blocks of Ω_3 with Motionformer (Patrick et al. 2021).

Table 2 shows that both Ω_3 and Ω_4 significantly outperform Ω_1 in terms of PSNR and SSIM. Incorporating the MambaVR block results in improvements of 3.71 dB for AVG and 2.74 dB for VFI. A visual comparison in Figure 7 (a) illustrates that both Ω_3 and Ω_4 exhibit greater clarity and more detailed textures than Ω_1, while the MSTM shows slight blurring of edge structures compared to MSMM.

Effectiveness of GFM. To verify the effectiveness of the

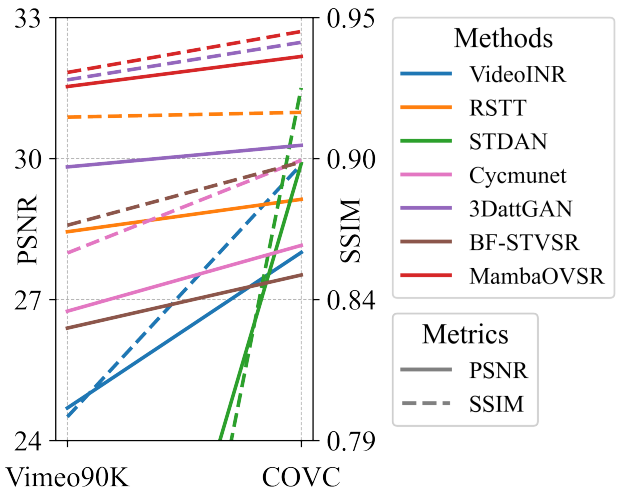

Figure 6: Quantitative comparison of methods trained on the Vimeo90K and COVC datasets.

Methods	Ω_1	Ω_2	Ω_3	Ω_4	Ω_5
DConv	✓	✗	✓	✓	✗
GFM	✗	✓	✗	✗	✓
Cyc	✓	✓	✗	✗	✗
MSTM	✗	✗	✓	✗	✗
MSMM	✗	✗	✗	✓	✓
AVG	28.15	28.37	31.67	31.86	**32.17**
VFI	28.00	28.31	30.53	30.74	**31.05**

Table 2: Ablation study results for GFM and MSMM are presented via PSNR comparisons. Cyc denotes the baseline, while MSTM refers to the MSMM module incorporating a Motionformer transformer block.

proposed GFM, we conducted two ablation studies on distinct baseline architectures: (Ω_1, Ω_2) and (Ω_4, Ω_5). Ω_1 and Ω_4 use a deformable convolution (DConv)-based feature interpolation module, while Ω_2 and Ω_5 replace it with GFM. Quantitative results (see Table 2) show a consistent PSNR improvement of 0.2–0.3 dB with GFM over DConv. Visual comparisons in Figure 7 (b) further confirm that GFM produces sharper edges and more accurate fine structures, demonstrating its effectiveness in enhancing detail recovery and reconstruction fidelity in video super-resolution.

Effectiveness of MambaVR.

The ablation study results for our MambaVR block are shown in Table 3. All MambaVR variants outperform the vanilla VideoMamba block (Li et al. 2025) in both PSNR and SSIM. Notably, Registers and F-RoPE work synergistically, with their combination providing the greatest improvement in reconstruction quality. Feature-map visualizations in Figure 8 show that VideoMamba alone generates overly blurred facial regions, while incorporating Registers reduces

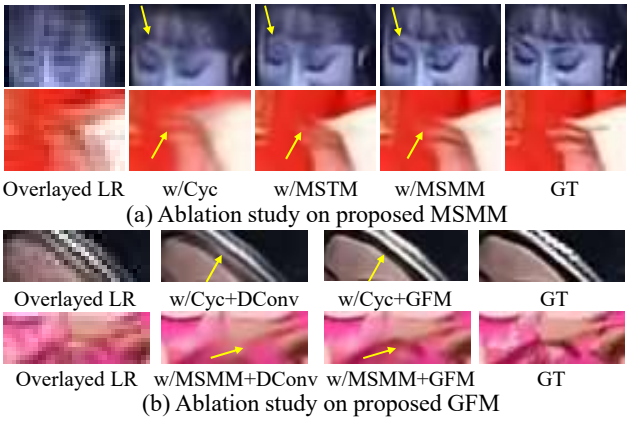

Figure 7: Qualitative Comparisons of proposed GFM and MSMM for ablation studies. DConv denotes the deformable convolution-based module.

Methods	PSNR		SSIM	
	AVG	VFI	AVG	VFI
VideoMamba	31.72	30.65	0.9368	0.9259
w/R	31.82	**30.79**	0.9422	0.9328
w/F-RoPE	31.80	30.71	0.9419	0.9320
MambaVR	**31.86**	30.74	**0.9438**	**0.9336**

Table 3: Ablation study on the proposed MambaVR. w/ denotes inclusion of each enhancement.

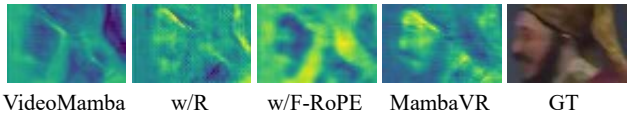

Figure 8: Feature map of MambaVR ablation study.

the blur, and F-RoPE sharpens facial contours. Combining both Registers and F-RoPE enables MambaVR to achieve the most detailed and accurate facial reconstructions.

Conclusion

In this work, we built a large-scale Chinese opera video clip (COVC) dataset and introduced the Mamba-Based multiscale fusion network for space-time Opera Video Super-Resolution (MambaOVSR). Specifically, we designed a global fusion module (GFM) for fine-grained holistic motion modeling between adjacent frames. Additionally, we proposed a MambaVR block to achieve global alignment. Based on this, our Multiscale Synergistic Mamba Module (MSMM) implemented granular motion alignment across varying sequence lengths. Experimental results on the COVC and Vimeo90K dataset showed that our method significantly outperforms existing STVSR methods. Future work will focus on optimising computational efficiency.

Acknowledgments

This work was supported by the Natural Science Foundation of China (62376201, and 62501189), Hubei Provincial Science & Technology Talent Enterprise Services Program (2025DJB059), Hubei Provincial Special Fund for Central-Guided Local S&T Development (2025CSA017), and the Natural Science Foundation of Heilongjiang Province of China for Excellent Youth Project (YQ2024F006).

References

Chen, G.; Huang, Y.; Xu, J.; Pei, B.; Chen, Z.; Li, Z.; Wang, J.; Li, K.; Lu, T.; and Wang, L. 2024. Video mamba suite: State space model as a versatile alternative for video understanding. *arXiv preprint arXiv:2403.09626*.

Chen, Z.; Chen, Y.; Liu, J.; Xu, X.; Goel, V.; Wang, Z.; Shi, H.; and Wang, X. 2022. Videoinr: Learning video implicit neural representation for continuous space-time super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2047–2057.

Cheng, X.; and Chen, Z. 2021. Multiple video frame interpolation via enhanced deformable separable convolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 7029–7045.

Chung, F. M.-Y. 2024. Utilising technology as a transmission strategy in intangible cultural heritage: the case of Cantonese opera performances. *International Journal of Heritage Studies*, 30(2): 210–225.

Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 764–773.

Dang, J.; Zheng, H.; Lai, J.; Yan, X.; and Guo, Y. 2023. Efficient and robust video object segmentation through isogenous memory sampling and frame relation mining. *IEEE Transactions on Image Processing*, 32: 3924–3938.

Dang, J.; Zheng, H.; Wang, B.; Wang, L.; and Guo, Y. 2024a. Temporo-spatial parallel sparse memory networks for efficient video object segmentation. *IEEE Transactions on Intelligent Transportation Systems*.

Dang, J.; Zheng, H.; Wu, X.; Jiao, J.; Wang, B.; Yang, J.; Hu, B.; Lai, J.; and Chua, T. S. 2025. External Memory Matters: Generalizable Object-Action Memory for Retrieval-Augmented Long-Term Video Understanding.

Dang, J.; Zheng, H.; Xu, X.; Wang, L.; and Guo, Y. 2024b. Beyond appearance: Multi-frame spatio-temporal context memory networks for efficient and robust video object segmentation. *IEEE Transactions on Image Processing*.

Dang, J.; Zheng, H.; Xu, X.; Wang, L.; Hu, Q.; and Guo, Y. 2024c. Adaptive sparse memory networks for efficient and robust video object segmentation. *IEEE Transactions on Neural Networks and Learning Systems*.

Fu, C.; Yuan, H.; Shen, L.; Hamzaoui, R.; and Zhang, H. 2024. 3DAttGAN: A 3D Attention-Based Generative Adversarial Network for Joint Space-Time Video Super-Resolution. *IEEE Transactions on Emerging Topics in Computational Intelligence*.

Geng, Z.; Liang, L.; Ding, T.; and Zharkov, I. 2022. 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*, 17441–17451.

Gotmare, A.; Keskar, N. S.; Xiong, C.; and Socher, R. 2018. A closer look at deep learning heuristics: Learning rate restarts, warmup and distillation. *arXiv preprint arXiv:1810.13243*.

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Haris, M.; Shakhnarovich, G.; and Ukita, N. 2020. Space-time-aware multi-resolution video enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2859–2868.

Hu, M.; Jiang, K.; Nie, Z.; and Wang, Z. 2022. You only align once: Bidirectional interaction for spatial-temporal video super-resolution. In *Proceedings of the 30th ACM International Conference on Multimedia*, 847–855.

Hu, M.; Jiang, K.; Nie, Z.; Zhou, J.; and Wang, Z. 2023a. Store and fetch immediately: Everything is all you need for space-time video super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 863–871.

Hu, M.; Jiang, K.; Wang, Z.; Bai, X.; and Hu, R. 2023b. Cycmunet+: Cycle-projected mutual learning for spatial-temporal video super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Jiang, K.; Wang, Q.; An, Z.; Wang, Z.; Zhang, C.; and Lin, C.-W. 2024. Mutual retinex: Combining transformer and cnn for image enhancement. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 8(3): 2240–2252.

Jiang, K.; Wang, Z.; Wang, Z.; Chen, C.; Yi, P.; Lu, T.; and Lin, C.-W. 2022. Degrade is upgrade: Learning degradation for low-light image enhancement. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 1078–1086.

Jiang, K.; Wang, Z.; Yi, P.; Chen, C.; Wang, Z.; Wang, X.; Jiang, J.; and Lin, C.-W. 2021. Rain-free and residue hand-in-hand: A progressive coupled network for real-time image deraining. *IEEE Transactions on Image Processing*, 30: 7404–7418.

Jiang, K.; Wang, Z.; Yi, P.; Wang, G.; Gu, K.; and Jiang, J. 2019. ATMFN: Adaptive-threshold-based multi-model fusion network for compressed face hallucination. *IEEE Transactions on Multimedia*, 22(10): 2734–2747.

Kim, E.; Kim, H.; Jin, K. H.; and Yoo, J. 2025. BF-STVSR: B-Splines and Fourier-Best Friends for High Fidelity Spatial-Temporal Video Super-Resolution. *arXiv preprint arXiv:2501.11043*.

Li, F.; Zhang, L.; Liu, Z.; Lei, J.; and Li, Z. 2023. Multi-frequency representation enhancement with privilege information for video super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12814–12825.

Li, K.; Li, X.; Wang, Y.; He, Y.; Wang, Y.; Wang, L.; and Qiao, Y. 2025. Videomamba: State space model for efficient video understanding. In *European Conference on Computer Vision*, 237–255. Springer.

Liu, Y.; Deng, Y.; Chen, H.; and Yang, Z. 2024a. Video Frame Interpolation via Direct Synthesis with the Event-based Reference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8477–8487.

Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; and Liu, Y. 2024b. VMamba: Visual State Space Model. *arXiv e-prints*, arXiv–2401.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Mackin, A.; Zhang, F.; Papadopoulos, M. A.; and Bull, D. 2017. Investigating the impact of high frame rates on video compression. In *2017 IEEE International Conference on Image Processing (ICIP)*, 295–299. IEEE.

Patrick, M.; Campbell, D.; Asano, Y.; Misra, I.; Metze, F.; Feichtenhofer, C.; Vedaldi, A.; and Henriques, J. F. 2021. Keeping your eye on the ball: Trajectory attention in video transformers. *Advances in neural information processing systems*, 34: 12493–12506.

Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A. P.; Bishop, R.; Rueckert, D.; and Wang, Z. 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1874–1883.

Su, J.; Ahmed, M.; Lu, Y.; Pan, S.; Bo, W.; and Liu, Y. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568: 127063.

Tag, B.; Shimizu, J.; Zhang, C.; Kunze, K.; Ohta, N.; and Sugiura, K. 2016. In the eye of the beholder: The impact of frame rate on human eye blink. In *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*, 2321–2327.

Wang, F.; Wang, J.; Ren, S.; Wei, G.; Mei, J.; Shao, W.; Zhou, Y.; Yuille, A.; and Xie, C. 2024. Mamba-R: Vision Mamba ALSO Needs Registers. *arXiv e-prints*, arXiv–2405.

Wang, H.; Xiang, X.; Tian, Y.; Yang, W.; and Liao, Q. 2023. Stdan: deformable attention network for space-time video super-resolution. *IEEE Transactions on Neural Networks and Learning Systems*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Xiang, X.; Tian, Y.; Zhang, Y.; Fu, Y.; Allebach, J. P.; and Xu, C. 2020. 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*, 3370–3379.

Xu, G.; Xu, J.; Li, Z.; Wang, L.; Sun, X.; and Cheng, M.-M. 2021. Temporal modulation network for controllable space-time video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6388–6397.

Xue, T.; Chen, B.; Wu, J.; Wei, D.; and Freeman, W. T. 2019. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127: 1106–1125.

Yi, P.; Wang, Z.; Jiang, K.; Jiang, J.; and Ma, J. 2019. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3106–3115.

Zhang, G.; Liu, C.; Cui, Y.; Zhao, X.; Ma, K.; and Wang, L. 2024. Vfimamba: Video frame interpolation with state space models. *arXiv preprint arXiv:2407.02315*.

Zhou, C.; Lu, Z.; Li, L.; Yan, Q.; and Xue, J.-H. 2021. How video super-resolution and frame interpolation mutually benefit. In *Proceedings of the 29th ACM International Conference on Multimedia*, 5445–5453.

Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; and Wang, X. 2024. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model. *arXiv e-prints*, arXiv–2401.