DATA-VSR: Dynamic Trajectory Attention and Texture Adaptive Rooter for Video Super-Resolution

Linfeng He[1,2], Meiqin Liu[1,2*], Qi Tang[1,2], Chao Yao[3], Yao Zhao[1,2],
[1]Institute of Information Science, Beijing Jiaotong University, Beijing, China
[2]Visual Intelligence + X International Cooperation Joint Laboratory of MOE, Beijing Jiaotong University, Beijing, China
[3]School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, China
Emails: {linfenghe, mqliu, qitang, yzhao}@bjtu.edu.cn, yaochao@ustb.edu.cn

***Abstract*—Video Super-Resolution (VSR) is essential for reconstructing high-definition sequences from correlated video frames. While Transformer-based VSR methods have improved reconstruction quality, they require substantial computational resources, limiting deployment on resource-constrained devices. To tackle this issue, we propose a novel framework named Dynamic Trajectory Attention and Texture Adaptive Rooter for Video Super-Resolution (DATA-VSR). There are two key innovations: the Temporal Redundancy-aware Alignment Network (TRAN) and the Spatial Redundancy-aware Refinement Network (SRRN). Specifically, features are aligned by focusing on dynamic temporal trajectories instead of static redundancies in TRAN, and then features are adaptively refined based on the texture complexity of different regions in SRRN. Additionally, the Dual-Domain Enhancement Block (DDEB) is incorporated to effectively capture global dependencies in the frequency domain and enhance the representation of local features in the spatial domain. The experimental results on standard VSR benchmarks show that DATA-VSR achieves competitive performance with fewer parameters, lower FLOPs, and a specific reduction of 17%.**

***Index Terms*—Low-Level Vision, Video Super-Resolution, Efficient Video Process.**

I. Introduction

Video Super-Resolution (VSR) enhances low-resolution (LR) video frames by generating corresponding high-resolution (HR) versions. With applications in live streaming [1], video surveillance [2], and restoring old films [3], VSR has gained significant attention. Effective frame enhancement in VSR relies on utilizing information from adjacent frames. According to the propagation of temporal features, existing VSR methods can be classified into three distinct paradigms [4]: sliding-window structures [5], [6], recurrent-based structures [7]–[10] and transformer-based structures [11]–[14].

Sliding-window methods use multiple frames within a small window and techniques like optical flow to align and fuse temporal features. However, these methods are constrained by a limited temporal receptive field and their efficiency is often hindered due to the repeated computations required for overlapping video frames. Recurrent-based methods propagate temporal information, offering a broader temporal receptive field and lower computation loads but struggle with low re-usability and long-term modeling due to the vanishing gradient problem [15]. Transformer-based methods employ the self-attention mechanism to capture distant information [16], significantly improving super-resolution performance, but their high computational costs hinder deployment on edge devices.

To promote the employment of VSR models in efficient video processing, we identify spatial-temporal redundancies in natural videos and leverage them to minimize unnecessary computation and error information. As evident in the regions with minimal displacement highlighted by the blue box in Fig. 1, the temporal redundancy will hinder information propagation and cause a performance drop. Additionally, the heatmap in Fig. 1 demonstrates that areas with smooth texture, indicating spatial redundancies, tend to achieve higher PSNR values more easily than complex textures within video frames. To better discriminate and utilize these redundancies, we propose a novel architectural framework: Dynamic Trajectory Attention and Texture Adaptive Rooter for Video Super-Resolution (DATA-VSR).

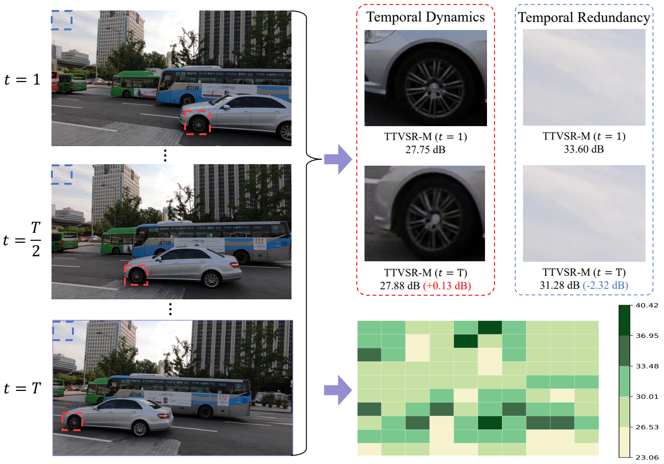


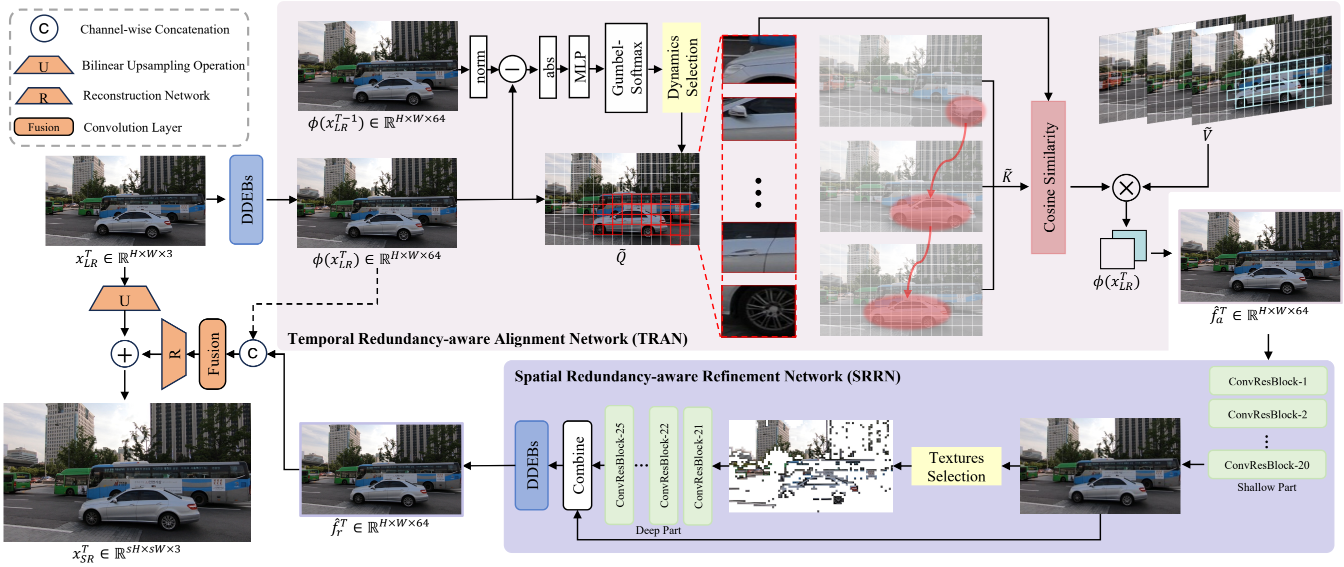
Fig. 1: Illustration of the impact of spatial-temporal redundancy in the VSR network. Using temporal trajectory attention, the VSR network extracts complementary information from dynamic regions of past frames to enhance reconstruction. Temporal redundancy in static regions hinders propagation, leading to a performance drop. Dark green areas in the heatmap indicate higher PSNR values and easier restoration. Results are obtained with the pre-trained TTVSR-M [17].

DATA-VSR achieves this through the Temporal Redundancy-aware Alignment Network (TRAN) and the Spatial Redundancy-aware Refinement Network (SRRN). Both TRAN and SRRN leverage redundancy-aware mechanisms to efficiently exploit inter-frame and intra-frame information. TRAN employs an effective self-attention operation inspired by the trajectory-aware transformer [17], focusing on trajectory segments that contain supplementary temporal information while ignoring those with temporal redundancy. Subsequently, SRRN adaptively adjusts its depth based on texture complexity in various frame regions. Moreover, DATA-VSR integrates several Dual Domain Enhancement Blocks (DDEB), which employ a three-branch structure to achieve a larger spatial receptive field and better reconstruction performance with fewer parameters and FLOPs.

In summary, our contributions are shown as follows:

- We propose a novel architectural framework named DATA-VSR. It efficiently aggregates inter-frame and intra-frame information by enhancing the identification and utilization of redundancy.
- We design TRAN and SRRN. TRAN aligns dynamic features

[*] Corresponding author.



Channel-wise Concatenation
Bilinear Upsampling Operation
Reconstruction Network
Convolution Layer
C
U
R
Fusion
$\phi(x_{LR}^{T-1}) \in \mathbb{R}^{H\times W\times 64}$
norm
abs
MLP
Gumbel-Softmax
Dynamics Selection
$x_{LR}^{T} \in \mathbb{R}^{H\times W\times 3}$
DDEBs
$\phi(x_{LR}^{T}) \in \mathbb{R}^{H\times W\times 64}$
$\tilde{Q}$
$\tilde{K}$
Cosine Similarity
$\tilde{V}$
$\phi(x_{LR}^{T})$
$f_a^T \in \mathbb{R}^{H\times W\times 64}$
Temporal Redundancy-aware Alignment Network (TRAN)
Spatial Redundancy-aware Refinement Network (SRRN)
ConvResBlock-1
ConvResBlock-2
ConvResBlock-20
Shallow Part
Textures Selection
ConvResBlock-21
ConvResBlock-22
ConvResBlock-25
Combine
Deep Part
$f_r^T \in \mathbb{R}^{H\times W\times 64}$
$x_{SR}^{T} \in \mathbb{R}^{sH\times sW\times 3}$

Fig. 2: The overall architecture of our DATA-VSR. It primarily consists of feature extraction, propagation, alignment, refinement, and reconstruction. More details of our proposed DATA-VSR can be found in Section II

along trajectories to enhance information propagation, while SRRN adaptively refines textures to reduce computations.

- We propose DDEB to enhance the performance of feature extraction and refinement while reducing computational overhead.

II. The Proposed Method

A. *Method Overview*

In the VSR task, the goal is to reconstruct an HR version from a given LR sequence. DATA-VSR utilizes a recurrent bi-directional propagation structure, and the forward reconstruction process is depicted in Fig. 2. To obtain the reconstructed T^{th} frame $x_{SR}^T \in \mathbb{R}^{sH\times sW\times 3}$, we use the current LR frame $x_{LR}^T \in \mathbb{R}^{H\times W\times 3}$ and the previous LR frames from 1 to $T-1$ as supporting data, where s is the scaling factor and H, W, 3 are the height, width, and number of channels of the input frames, respectively. DATA-VSR first extracts shallow features from x_{LR}^T using DDEB and integrates temporal information along motion trajectories. TRAN enhances inter-frame alignment by fusing dynamic content instead of temporal redundancy. SRRN then refines the aligned feature by adapting the network depth according to texture complexity. Finally, the refined and shallow features are combined to produce the HR video frame $\boldsymbol{x}_{SR}^T$.

B. *Temporal Redundancy-aware Alignment Network*

The alignment process involves both forward and backward branches, constructed based on motion trajectories in the temporal dimension. Taking the forward branch of TRAN in Fig. 2 as an example, the trajectory set $\mathcal{T}$ and each trajectory are formulated as:

$$\begin{aligned} \mathcal{T} &= \{\tau_i, i \in [1, \mathrm{N}]\}, \\ \tau_i &= \left\{\left(x_i^t, y_i^t\right), t \in [1, T]\right\}, \end{aligned} \tag{1}$$

where $x_i^t \in [1, H]$, $y_i^t \in [1, W]$. (x_i^t, y_i^t) denotes the coordinates of trajectory τ_i at time t. N is the number of tokens per frame, and i is the corresponding token index.

To avoid unnecessary computations and propagation of potential errors when capturing long-term temporal information through self-attention, DATA-VSR selects complementary dynamic tokens for precise trajectory attention alignment at the current time T.

1) Temporal Dynamics Selection. TRAN first extracts the queries $\boldsymbol{Q} = \phi(x_{LR}^T)$, keys $\boldsymbol{K} = \phi(x_{LR}^{1:T-1})$, and values $\boldsymbol{V} = \varphi(x_{LR}^{1:T-1})$ used for self-attention along the trajectory from the video frames, where $\phi(\cdot)$ denotes a shallow feature extraction network built from stacked DDEB. $\phi(\cdot)$ is integrated with TRAN and SRRN to construct a deep feature refinement network $\varphi(\cdot)$.

To discriminate the temporal redundancy in the current frame, TRAN uses feature differences between adjacent frames to generate a binary mask with a uniform threshold and employs a Gumbel-Softmax gate for sampling [18]. Specifically, it facilitates joint training of the mask prediction network with the VSR network, where distinct masking features $M^{T-1\to T}$ for various videos are formulated as:

$$M^{T-1\to T} = \frac{\exp\left(\left(\log\left(\pi_1\right) + g_1\right)/\sigma\right)}{\sum_{i=1}^{2} \exp\left(\left(\log\left(\pi_i\right) + g_i\right)/\sigma\right)}, \tag{2}$$

where g_1 and g_2 are Gumbel noise samples. π_1 and π_2 represent the probabilities of the presence and absence of complementary information, respectively, and are formulated as:

$$\begin{aligned} \pi_1 &= \mathrm{Sigmoid}(f(\Delta Q^{T-1\to T})), \\ \pi_2 &= 1 - \pi_1, \end{aligned} \tag{3}$$

where $f(\cdot)$ denotes an MLP layer for weighted sums of feature differences. σ is the temperature coefficient. $\Delta Q^{T-1\to T}$ is the difference between normalized features. The masking feature enables TRAN to generate the alignment binary mask M_a and dynamic trajectories set $\tilde{\mathcal{T}}$, which are formulated as:

$$\begin{aligned} M_a &= \begin{cases} 1, & if \quad M^{T-1\to T} > 0.5, \\ 0, & else, \end{cases} \\ \tilde{\mathcal{T}} &= \mathcal{T} \times M_a = \{\tilde{\tau}_i, \ i \in [1, \mathrm{M}]\}, \end{aligned} \tag{4}$$

where M is the number of selected dynamic trajectories with M far less than N and 0.5 is the threshold. Based on these selected trajectories, the corresponding tokens $\tilde{\boldsymbol{Q}}$, $\tilde{\boldsymbol{K}}$, and $\tilde{\boldsymbol{V}}$ for long trajectory attention alignment are formulated as:

$$\begin{aligned} \tilde{\boldsymbol{Q}} &= \{q_{\tilde{\tau}_i^T}, i \in [1, \mathrm{M}]\}, \\ \tilde{\boldsymbol{K}} &= \{k_{\tilde{\tau}_i^t}, i \in [1, \mathrm{M}], t \in [1, T-1]\}, \\ \tilde{\boldsymbol{V}} &= \{v_{\tilde{\tau}_i^t}, i \in [1, \mathrm{M}], t \in [1, T-1]\}, \end{aligned} \tag{5}$$

2) Trajectory-based Attention Alignment. Once the trajectories with complementary information are selected, the self-attention operations are applied to the dynamic features along these trajectories to achieve alignment. The aligned feature f_a^T can be formulated as:

$$f_a^T = \mathrm{A}_{traj}(q_{\tilde{\tau}_i^T}, k_{\tilde{\tau}_i^t}, v_{\tilde{\tau}_i^t}), \tag{6}$$

where A_{traj} represents the trajectory attention alignment operation and is formulated as:

$$\mathrm{A}_{traj}(q_{\tilde{\tau}_i^T}, k_{\tilde{\tau}_i^t}, v_{\tilde{\tau}_i^t}) = \mathrm{F}(q_{\tilde{\tau}_i^T}\ ,\ s \odot v_{\tilde{\tau}_i^h}), \tag{7}$$

where $\mathrm{F}(\cdot)$ represents feature concatenation and fusion operation. h and s, representing the outcomes of hard and soft attention operations, respectively, are formulated as:

$$\begin{aligned} h &= \arg\max_t \langle \frac{q_{\tilde{\tau}_i^T}}{\| q_{\tilde{\tau}_i^T} \|_2}, \frac{k_{\tilde{\tau}_i^t}}{\| k_{\tilde{\tau}_i^t} \|_2} \rangle, \\ s &= \max_t \langle \frac{q_{\tilde{\tau}_i^T}}{\| q_{\tilde{\tau}_i^T} \|_2}, \frac{k_{\tilde{\tau}_i^t}}{\| k_{\tilde{\tau}_i^t} \|_2} \rangle, \end{aligned} \tag{8}$$

where $\langle \cdot \rangle$ and $\| \cdot \|_2$ represent the dot product and the Euclidean norm operations, respectively.

3) Propagated Tokens Update. Following the alignment of the selected dynamic features, the updated features in forward propagation $\hat{f}_a^T$ can be formulated as:

$$\hat{f}_a^T = f_a^T + \hat{f}_r^{T-1} \times (1 - M_a), \tag{9}$$

where $\hat{f}_r^{T-1}$ denotes the propagated features after being fully refined by network $\varphi(\cdot)$ from the previous step.

C. Spatial Redundancy-aware Refinement Network

1) DCT-based Texture Selection. Inspired by temporal dynamics selection, SRRN is designed to adaptively allocate computational resources within a frame based on texture complexity. By leveraging the Discrete Cosine Transform (DCT) [19], SRRN takes advantage of the low-frequency energy concentration property, where complex textures with high-frequency components correspond to lower energy. This property is used to derive $M_{i,j}^{texture}$ for selective refinement, which can be formulated as:

$$M_{i,j}^{texture} = \begin{cases} 1, & E_{i,j}^2 \leq \text{threshold}, \\ 0, & E_{i,j}^2 > \text{threshold}, \end{cases} \tag{10}$$

where $E_{i,j}^2$ denotes the average energy of the corresponding patch obtained from the DCT transform and (i, j) denotes the coordinate of the patch. Therefore, SRRN avoids causing unnecessary computational costs in spatially redundant areas like regions of the sky and applies additional refinement only to complex textures like tiled surfaces.

2) Dual-Domain Enhancement Block. As depicted in Fig. 3, the DDEB is proposed to enhance the feature representation capability of DATA-VSR, replacing the conventional convolutional residual blocks in both shallow feature extraction and deep feature refinement layers, thereby overcoming the limitations in representational capacity.

DDEB adopts a three-branch structure: the frequency branch leverages fast Fourier transform [20], [21] convolutions to capture global spatial dependencies, the spatial branch extracts local features through conventional convolutions, and the channel branch employs a channel attention mechanism [22] to enhance critical channels. The interactive fusion of features from these branches balances global and local information, improving reconstruction performance while reducing parameter count and FLOPs.

III. Experiment

A. Training Details

During the training process, we use a Cosine Annealing scheme [35] and an Adam optimizer [36] with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The learning rates for the motion estimation and other model components are set as 1.25×10^{-5} and 1×10^{-4}, respectively. The Charbonnier penalty loss [37] $\mathcal{L}_{sr} = \sqrt{\|x_{HR} - x_{SR}\|^2 + \varepsilon^2}$ is applied on the reconstructed image x_{SR} and the ground truth image x_{HR}, where ε is a constant and is set as 10^{-3}. We also impose an ℓ_1 loss $\mathcal{L}_{mask} = \frac{1}{T}\sum_{t=1}^{T} \|\mathrm{M}^{T-1\rightarrow T}\|$ on the mask for alignment to promote more effective masking. Thus, the final loss $\mathcal{L}$ can be formulated as:

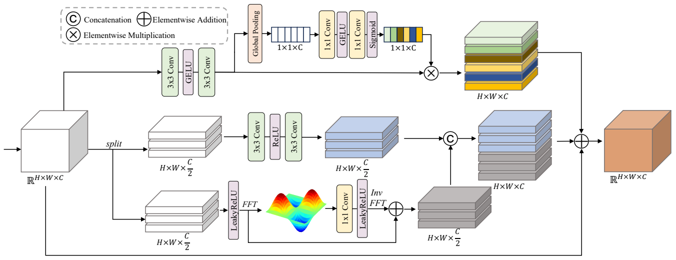

Concatenation
Elementwise Addition
Elementwise Multiplication
3x3 Conv
GELU
3x3 Conv
Global Pooling
1×1×C
1x1 Conv
GELU
1x1 Conv
Sigmoid
1×1×C
H×W×C
$\mathbb{R}^{H\times W\times C}$
split
$H\times W\times \frac{C}{2}$
3x3 Conv
ReLU
3x3 Conv
$H\times W\times \frac{C}{2}$
H×W×C
$\mathbb{R}^{H\times W\times C}$
$H\times W\times \frac{C}{2}$
LeakyReLU
FFT
1x1 Conv
LeakyReLU
Inv FFT
$H\times W\times \frac{C}{2}$

Fig. 3: The overall architecture of our DDEB.

$$\mathcal{L} = \mathcal{L}_{sr} + \lambda \mathcal{L}_{mask}. \tag{11}$$

where λ is used to adjust the masking ratio.

Referenced as BasicVSR [7], we also adopt two widely-used datasets for training: REDS [23] and Vimeo-90K [24]. The training process contains two stages. A base model without redundancy-aware capability is trained for 400K iterations, and the model is fine-tuned with the redundancy selection module for an additional 100K iterations. During the fine-tuning stage, a DCT threshold of 150 and a patch size of 4 are empirically set to identify complex textures.

B. Evaluation Metrics

Generated video quality is evaluated with Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) [38]. Model efficiency is evaluated based on the number of learnable parameters, floating-point operations per second (FLOPs) for an LR frame size of 180 × 320, and average runtimes on the REDS4 [23].

C. Comparison with Other Methods

We conduct a comparative analysis of DATA-VSR's performance against leading VSR methods under limited computational resources.

1) Quantitative Comparison. We evaluate the performance of our DATA-VSR model against state-of-the-art (SOTA) methods using datasets REDS4 [23], Vimeo90K [24], Vid4 [25] and UDM10 [26] as shown in TABLE I. Compared to the representative sliding window-based model EDVR [6], DATA-VSR achieves notable performance gains of 0.11 dB to 0.46 dB on longer sequence datasets and a nearly tenfold reduction in FLOPs. Compared to recurrent-based methods such as BasicVSR [7] and TTVSR [17], with TTVSR-M retrained to match BasicVSR's residual blocks for fair comparison, DATA-VSR achieves a 17% reduction in FLOPs, a 33% decrease in parameters, and a 0.14 dB PSNR improvement on the Vid4 dataset. Furthermore, compared to the recent efficient VSR model like SKipVSR [34], DATA-VSR achieves up to 0.95 dB PSNR gains with a 15% reduction in parameters.

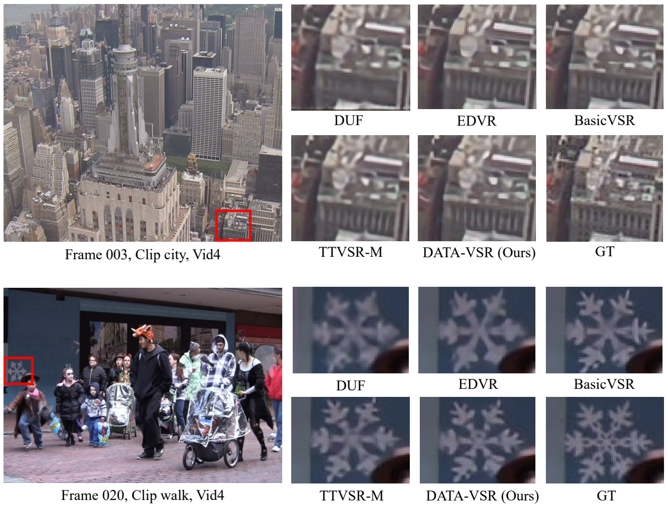
2) Qualitative Comparison. We conduct a qualitative comparison of DATA-VSR and SOTA methods on the Vid4 dataset. As shown in Fig. 4a, DATA-VSR significantly improves visual quality, especially in complex textured areas. Additionally, the result of the Local Attention Map (LAM) [39] in Fig. 4b reveals that DATA-VSR has a much larger receptive field than TTVSR-M and integrates more relevant information from the surrounding regions.

D. Ablation Study

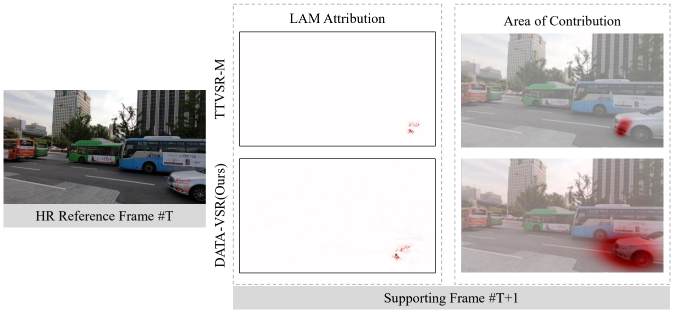
To evaluate the effectiveness of each component in DATA-VSR, we conduct the following ablation studies on the REDS4 dataset [23].

TABLE I: Quantitative comparison (average PSNR/SSIM) with state-of-the-art methods for video super-resolution (×4).

Method				BI degradation			BD degradation		
	Params (M)	FLOPs (T)	Runtime (ms)	REDS4 [23]	Vimeo-90K-T [24]	Vid4 [25]	UDM10 [26]	Vimeo-90K-T [24]	Vid4 [25]
Bicubic	-	-	-	26.14/0.7292	31.32/0.8684	23.78/0.6347	28.47/0.8253	31.30/0.8687	21.80/0.5246
DUF [27]	5.8	2.34	974	28.63/0.8251	-	-	38.48/0.9605	36.87/0.9447	27.38/0.8329
RBPN [28]	12.2	8.51	1507	30.09/0.8590	37.07/0.9435	27.12/0.8180	38.66/0.9596	37.20/0.9458	-
EDVR-M [6]	3.3	0.46	118	30.53/0.8699	37.09/0.9446	27.10/0.8186	39.40/0.9663	37.33/0.9484	27.45/0.8406
EDVR [6]	20.6	2.95	378	31.09/0.8800	37.61/0.9489	27.35/0.8264	39.89/0.9686	**37.81/0.9523**	27.85/0.8503
BoostedEDVR [29]	**3.3**	0.31	260	30.53/0.8699	-	25.32/0.7950	-	-	-
MuCAN [30]	13.6	1.07	-	30.88/0.8750	37.32/0.9465	-	-	-	-
BasicVSR [7]	6.3	0.34	63	31.42/0.8909	37.18/0.9450	27.24/0.8251	39.96/0.9694	37.53/0.9498	27.96/0.8553
BoostedBasicVSR [29]	6.3	0.51	93	31.42/0.8917	-	-	-	-	-
VSRT [12]	32.6	2.91	367	31.19/0.8815	**37.71/0.9494**	27.36/0.8258	-	-	-
TTVSR-M [17]	4.45	0.36	91	31.50/0.8927	37.09/0.9442	27.43/0.8283	40.14/0.9699	37.45/0.9493	28.08/0.8588
R2D2 [31]	8.25	-	-	-	-	-	39.40/0.9662	-	28.07/0.8537
DPR [32]	6.3	0.36	**48**	31.38/0.8907	37.11/0.9446	27.19/0.8243	39.72/0.9684	37.24/0.9461	27.89/0.8539
LGDFNet [33]	9.8	-	92	31.54/0.8911	-	27.44/**0.8395**	**40.23**/0.9703	-	**28.40**/0.8556
SkipVSR [34]	4.9	0.34	58	30.60/0.8726	36.39/0.9365	26.54/0.7924	39.05/0.9645	36.73/0.9398	26.98/0.8434
DATA-VSR (Ours)	4.24	**0.30**	88	**31.55/0.8934**	37.19/0.9452	**27.46**/0.8303	40.18/**0.9703**	37.55/0.9501	28.22/**0.8603**



(a) Visual comparison for 4× VSR on the Vid4 dataset



(b) The comparative result of the LAM for adjacent frames obtained by TTVSR-M and DATA-VSR

Fig. 4: Qualitative comparison on the REDS4 [23] and Vid4 [25] dataset.

TABLE II: Ablation study result of Redundancy-aware Network (RN) and DDEB in our DATA-VSR.

Module	DDEB	TRAN	SRRN	PSNR/SSIM ↑	FLOPs (G) ↓	Param (M) ↓
Base	✗	✗	✗	31.50/0.8927	363.40	4.447
Base+RN	✗	✓	✓	31.39/0.8909	320.94	4.447
Base+DDEB	✓	✗	✗	31.53/0.8930	341.71	4.244
Base+RN+DDEB	✓	✓	✓	**31.55/0.8934**	**301.21**	**4.244**

1) Validity of Redundancy-aware Network (RN) and DDEB. We use a base model without redundancy-aware alignment and refinement, replacing the DDEB with a conventional convolutional residual block. As shown in TABLE II, the proposed RN effectively discriminates and utilizes redundancy, leading to a 12% reduction in FLOPs. Additionally, the proposed DDEB compensates for the RN's performance drop while reducing parameters by 0.2 M. By integrating both RN and DDEB, the proposed DATA-VSR not only achieves 0.05 dB gains in PSNR but also results in a 17% reduction in FLOPs compared to the base model.

TABLE III: Ablation study result of the temporal redundancy masking strategy

TRM strategy	PSNR/SSIM↑	FLOPs (G) ↓	Param (M) ↓	Runtime (ms) ↓
Uniform	31.50/0.8923	**301.20**	**4.243**	101
Random	31.50/0.8923	301.20	4.243	105
Gumbel	**31.55/0.8934**	301.21	4.244	**88**

TABLE IV: Ablation study result of the spatial redundancy masking strategy

SRM strategy	PSNR/SSIM↑	FLOPs (G) ↓	Param (M) ↓	Runtime (ms) ↓
Uniform	31.50/0.8925	322.44	4.244	**75**
Random	31.49/0.8925	322.07	4.244	80
DCT-Based	**31.55/0.8934**	**301.21**	**4.244**	88

2) Effectiveness of the Redundancy Masking Strategy. Our proposed Gumbel-based and DCT-based masks achieve a 16% improvement in inference speed and a 7% reduction in FLOPs, with at least 0.05 dB gains in PSNR, outperforming random and uniform masks, as shown in TABLE III and IV.

IV. Conclusion

In this paper, we propose DATA-VSR to efficiently aggregate inter-frame information through TRAN and intra-frame information through SRRN. TRAN aligns dynamic features along trajectories, skipping temporal redundancies, while SRRN adaptively refines complex textures to minimize computations on spatial redundancies. Furthermore, the proposed three-branch structure DDEB further enhances the receptive field and improves the reconstruction quality.

Acknowledgment

This work is supported by the Fundamental Research Funds for the Central Universities (2024JBZX001), the National Natural Science Foundation of China (62372036, 62120106009, 62332017) and the Talent Fund of Beijing Jiaotong University (2024XKRC023).

References

[1] Z. Lu, H. Xia, S. Heo, and D. Wigdor, "You watch, you give, and you engage: a study of live streaming practices in china," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–13.

[2] L. Zhang, H. Zhang, H. Shen, and P. Li, "A super-resolution reconstruction algorithm for surveillance images," *Signal Processing*, vol. 90, no. 3, pp. 848–859, 2010.

[3] Z. Wan, B. Zhang, D. Chen, and J. Liao, "Bringing old films back to life," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 694–17 703.

[4] Q. Tang, Y. Zhao, M. Liu, and C. Yao, "A review of video super-resolution algorithms based on deep learning," *Acta Automatica Sinica*, 2024.

[5] L. Wang, Y. Guo, L. Liu, Z. Lin, X. Deng, and W. An, "Deep video super-resolution using hr optical flow estimation," *IEEE Transactions on Image Processing*, vol. 29, pp. 4323–4336, 2020.

[6] X. Wang, K. C. Chan, K. Yu, C. Dong, and C. Change Loy, "EDVR: Video restoration with enhanced deformable convolutional networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 1954–1963.

[7] K. C. Chan, X. Wang, K. Yu, C. Dong, and C. C. 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*, 2021, pp. 4947–4956.

[8] M. Liu, S. Jin, C. Yao, C. Lin, and Y. Zhao, "Temporal consistency learning of inter-frames for video super-resolution," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 4, pp. 1507–1520, 2022.

[9] M. Liu, C. Xu, C. Yao, C. Lin, and Y. Zhao, "JNMR: Joint non-linear motion regression for video frame interpolation," *IEEE Transactions on Image Processing*, vol. 32, pp. 5283–5295, 2023.

[10] S. Jin, M. Liu, C. Yao, C. Lin, and Y. Zhao, "Kernel dimension matters: To activate available kernels for real-time video super-resolution," in *Proceedings of the ACM International Conference on Multimedia*, 2023, pp. 8617–8625.

[11] Q. Tang, Y. Zhao, M. Liu, and C. Yao, "SeeClear: Semantic distillation enhances pixel condensation for video super-resolution," in *Annual Conference on Neural Information Processing Systems*, 2024.

[12] J. Cao, Y. Li, K. Zhang, and L. Van Gool, "Video super-resolution transformer," *arXiv preprint arXiv:2106.06847*, 2021.

[13] X. Zhou, L. Zhang, X. Zhao, K. Wang, L. Li, and S. Gu, "Video super-resolution transformer with masked inter&intra-frame attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 25 399–25 408.

[14] Q. Tang, Y. Zhao, M. Liu, J. Jin, and C. Yao, "Semantic lens: Instance-centric semantic alignment for video super-resolution," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024, pp. 5154–5161.

[15] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, no. 2, pp. 107–116, 1998.

[16] X. Chen, X. Wang, J. Zhou, Y. Qiao, and C. Dong, "Activating more pixels in image super-resolution transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 367–22 377.

[17] C. Liu, H. Yang, J. Fu, and X. Qian, "Learning trajectory-aware transformer for video super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5687–5696.

[18] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," *arXiv preprint arXiv:1611.01144*, 2016.

[19] S. Saha, "Image compression-from DCT to wavelets: a review," *XRDS: Crossroads, The ACM Magazine for Students*, vol. 6, no. 3, pp. 12–21, 2000.

[20] H. Yu, J. Huang, F. Zhao, J. Gu, C. C. Loy, D. Meng, C. Li *et al.*, "Deep fourier up-sampling," *Advances in Neural Information Processing Systems*, vol. 35, pp. 22 995–23 008, 2022.

[21] D. Zhang, F. Huang, S. Liu, X. Wang, and Z. Jin, "SwinFIR: Revisiting the swinir with fast fourier convolution and improved training for image super-resolution," *arXiv preprint arXiv:2208.11247*, 2022.

[22] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[23] S. Nah, S. Baik, S. Hong, G. Moon, S. Son, R. Timofte, and K. 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*, 2019.

[24] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, "Video enhancement with task-oriented flow," *International Journal of Computer Vision*, vol. 127, pp. 1106–1125, 2019.

[25] C. Liu and D. Sun, "On bayesian adaptive video super resolution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 2, pp. 346–360, 2013.

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

[27] Y. Jo, S. W. Oh, J. Kang, and S. J. 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*, 2018, pp. 3224–3232.

[28] M. Haris, G. Shakhnarovich, and N. Ukita, "Recurrent back-projection network for video super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3897–3906.

[29] Y. Huang, H. Dong, J. Pan, C. Zhu, B. Liang, Y. Guo, D. Liu, L. Fu, and F. Wang, "Boosting video super resolution with patch-based temporal redundancy optimization," in *Proceedings of the International Conference on Artificial Neural Networks*, 2023, pp. 362–375.

[30] W. Li, X. Tao, T. Guo, L. Qi, J. Lu, and J. Jia, "MuCAN: Multi-correspondence aggregation network for video super-resolution," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 335–351.

[31] A. A. Baniya, T.-K. Lee, P. W. Eklund, S. Aryal, and A. Robles-Kelly, "Online video super-resolution using information replenishing unidirectional recurrent model," *Neurocomputing*, vol. 546, p. 126355, 2023.

[32] C. Huang, J. Li, L. Chu, D. Liu, and Y. Lu, "Disentangle propagation and restoration for efficient video recovery," in *Proceedings of the ACM International Conference on Multimedia*, 2023, pp. 8336–8345.

[33] C. Zhang, X. Wang, R. Xiong, X. Fan, and D. Zhao, "Local-global dynamic filtering network for video super-resolution," *IEEE Transactions on Computational Imaging*, vol. 9, pp. 963–976, 2023.

[34] Z. Ai, X. Luo, Y. Qu, and Y. Xie, "SkipVSR: Adaptive patch routing for video super-resolution with inter-frame mask," in *Proceedings of the ACM International Conference on Multimedia*, 2024, pp. 5874–5882.

[35] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[37] W. Lai, J. Huang, N. Ahuja, and M. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 624–632.

[38] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, "Learning texture transformer network for image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5791–5800.

[39] J. Gu and C. Dong, "Interpreting super-resolution networks with local attribution maps," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9199–9208.