DOI: 10.21917/ijivp.2026.0561

REAL-TIME VIDEO SUPER-RESOLUTION USING A TEMPORAL-WINDOW TRANSFORMER WITH EFFICIENT MOTION-AWARE FEATURE RECONSTRUCTION

Mahesh Maurya[1] and B. Sathananth[2]

[1]*Department of Computer Engineering, St. John College of Engineering and Management, India*
[2] *Department of Electronics and Communication Engineering, Cape Institute of Technology, India*

Abstract

Video super-resolution (VSR) aims to reconstruct high-resolution video frames from their low-resolution counterparts while maintaining spatial details and temporal consistency. Unlike single-image super-resolution, VSR must exploit information distributed across neighboring frames, where motion, occlusion, blur, compression artifacts, and illumination variations can complicate reconstruction. Recent transformer-based restoration models have improved long-range temporal modeling, but their computational complexity and memory requirements remain barriers to real-time deployment. Existing VSR methods often face a trade-off between reconstruction quality and computational efficiency. Convolutional recurrent approaches provide efficient temporal propagation but may inadequately capture long-range dependencies, whereas transformer-based approaches model temporal relationships effectively but frequently require substantial computational resources. This study proposes a novel Temporal-Window Motion-Aware Transformer for Real-Time Super-Resolution (TWMAT-SR). The proposed framework combines lightweight convolutional feature extraction, motion-guided temporal alignment, localized temporal self-attention, and progressive feature reconstruction. A temporal-window attention mechanism restricts attention computation to informative neighboring frames, while a motion-aware alignment module dynamically compensates for inter-frame displacement. A recurrent feature-refinement pathway further propagates high-frequency information across successive frames without repeatedly processing the complete temporal sequence. The framework is optimized using a combined reconstruction, perceptual, and temporal-consistency objective. The experimental evaluation demonstrates that the proposed TWMAT-SR framework achieves 33.48 dB PSNR, 0.937 SSIM, and 0.110 LPIPS at the 50th training epoch on the REDS dataset for 4× video super-resolution. The model reaches an inference speed of 28.9 FPS, outperforming BasicVSR, VRT, and RVRT by 2.5, 12.8, and 9.2 FPS, respectively. The proposed method also obtains a temporal consistency error of 0.037, representing reductions of 30.2%, 22.9%, and 14.0% compared with BasicVSR, VRT, and RVRT. These results indicate that TWMAT-SR improves spatial reconstruction and perceptual fidelity while maintaining efficient inference and reduced temporal instability.

Keywords:

Video Super-Resolution, Transformer, Temporal Attention, Motion-Aware Alignment, Real-Time Video Restoration

1. INTRODUCTION

Video has become a primary source of visual information in surveillance, intelligent transportation, autonomous systems, medical imaging, remote sensing, entertainment, and mobile applications. However, video acquisition systems frequently operate under limitations imposed by sensor resolution, bandwidth, storage capacity, illumination, compression, and hardware cost. Consequently, many practical applications receive low-resolution (LR) video even when high-resolution (HR) information is required for subsequent analysis. Video super-resolution (VSR) addresses this limitation by reconstructing an HR sequence from one or more LR frames. In contrast to single-image super-resolution, VSR can exploit complementary information contained in neighboring frames. Details that are unavailable in one frame may appear partially in adjacent frames because of object motion and camera movement. Effective temporal information integration can therefore produce sharper structures and more realistic textures than independent frame processing [1-3].

Early VSR approaches primarily relied on convolutional neural networks and temporal feature propagation. These methods demonstrated that neighboring frames provide useful information for recovering spatial details, but accurate temporal alignment remains difficult when objects move rapidly or when occlusions occur. Conventional optical-flow-based alignment can also become unreliable in regions containing motion boundaries, repetitive textures, illumination changes, or complex deformation. BasicVSR demonstrated that VSR performance can be improved by carefully combining propagation, alignment, aggregation, and upsampling components, while its IconVSR extension introduced information refill and coupled propagation to reduce propagation errors. These developments established the importance of efficient temporal propagation and provided strong baselines for subsequent VSR research [1-3].

Despite these advances, several challenges continue to limit practical VSR deployment. First, consecutive frames are rarely perfectly aligned. Motion between frames may vary spatially, and inaccurate alignment can introduce ghosting, ringing, duplicated structures, and temporal inconsistencies [4]. Second, increasing the temporal receptive field generally increases computational and memory requirements. Transformer-based architectures can model relationships between distant frames more effectively than strictly local convolutional operations, but global self-attention becomes expensive when applied to high-resolution video sequences [5]. VRT, for example, introduced temporal mutual self-attention and parallel warping to model temporal relationships and demonstrated strong performance across multiple video restoration tasks, including VSR. However, the computational burden associated with transformer-based processing remains an important consideration for real-time applications.

The central problem addressed in this study is therefore the absence of a VSR architecture that simultaneously provides strong temporal dependency modeling, accurate motion compensation, high-frequency detail reconstruction, and sufficiently low computational complexity for real-time operation [6]. Existing recurrent architectures can achieve relatively efficient inference because parameters are shared across frames,

but their sequential processing can restrict parallelization and long-range dependency modeling. Conversely, fully parallel transformer architectures can capture extensive temporal relationships but may require considerable memory and processing resources. RVRT attempted to address this trade-off by combining local parallel processing with global recurrent processing and guided deformable attention. Nevertheless, further reduction in temporal attention cost and more efficient feature propagation remain desirable.

The first objective of this research is to develop a transformer-based VSR framework capable of exploiting informative temporal relationships without applying expensive global attention to every frame pair. The second objective is to design a motion-aware alignment mechanism that can compensate for spatial displacement before temporal feature aggregation. The third objective is to construct a lightweight reconstruction pathway that restores fine textures and edges while maintaining temporal consistency. The final objective is to establish a quality-efficiency balance suitable for real-time or near-real-time VSR applications.

To achieve these objectives, this study introduces the Temporal-Window Motion-Aware Transformer for Real-Time Super-Resolution (TWMAT-SR). The framework divides the input sequence into compact temporal windows and applies localized temporal attention to reduce the quadratic cost associated with global attention. A motion-aware alignment module estimates spatial correspondence between neighboring frames and guides feature aggregation toward relevant locations. Instead of processing the complete sequence repeatedly, a recurrent feature-refinement pathway carries selected high-frequency representations across successive temporal windows.

The principal novelty lies in integrating localized temporal transformer attention, motion-guided feature alignment, and recurrent high-frequency propagation within a single efficiency-oriented VSR architecture. Unlike approaches that rely primarily on either recurrent propagation or computationally intensive global attention, TWMAT-SR selectively allocates attention to neighboring temporal regions while maintaining information flow across the complete sequence. The main contributions are summarized as follows. (1) A novel TWMAT-SR framework is developed for real-time video super-resolution by combining temporal-window self-attention with motion-aware feature alignment and recurrent high-frequency refinement. (2) An efficiency-oriented temporal reconstruction strategy is introduced to reduce computational overhead while preserving spatial details and temporal consistency, providing a practical balance between reconstruction quality and inference speed.

2. RELATED WORKS

2.1 CONVENTIONAL DEEP VIDEO SUPER-RESOLUTION

Deep learning substantially changed VSR by enabling nonlinear reconstruction of high-frequency information from multiple LR frames. Early approaches generally used convolutional architectures with temporal fusion and motion compensation. The fundamental challenge was to exploit neighboring frames without introducing artifacts caused by inaccurate correspondence. Methods based on optical flow attempted to warp neighboring frames toward a reference frame, while later approaches introduced implicit feature alignment. The progression from explicit image warping to learned feature alignment established the importance of separating temporal propagation, alignment, and upsampling operations [5].

BasicVSR provided an influential simplification of this design space by systematically examining these four functional components. The method showed that bidirectional propagation combined with optical-flow-based feature alignment could achieve strong VSR performance without excessively complicated architecture. Its IconVSR extension introduced information refill using sparsely selected keyframes and coupled propagation between forward and backward branches. The study reported that these components improved information aggregation and reduced propagation errors caused by occlusions and image boundaries. These findings are important for the present study because they demonstrate that temporal propagation can be both effective and computationally manageable when the architecture avoids unnecessary operations [5].

2.2 ENHANCED PROPAGATION AND ALIGNMENT

The limitations of basic propagation strategies motivated more sophisticated VSR architectures. BasicVSR++ improved the original BasicVSR framework through enhanced propagation and alignment mechanisms, seeking more accurate information exchange across frames. Its development reflects a broader trend toward designing temporal propagation mechanisms that can handle complex motion while maintaining computational efficiency. The official implementation reports its use as a VSR framework developed from the BasicVSR family.

RealBasicVSR subsequently investigated VSR under real-world degradation rather than relying only on idealized bicubic degradation. Real video commonly contains sensor noise, compression artifacts, blur, and unknown degradation processes. Consequently, a model trained only with synthetic degradation may perform poorly when applied to videos captured by practical cameras. RealBasicVSR explicitly examined the trade-off between restoration quality and real-world applicability, making it particularly relevant to deployment-oriented VSR research. These studies indicate that reconstruction quality cannot be considered independently from degradation characteristics and inference requirements [6].

2.3 TRANSFORMER-BASED VIDEO RESTORATION

Transformers introduced an alternative mechanism for temporal dependency modeling. Rather than relying exclusively on convolutional receptive fields or recurrent hidden states, self-attention can directly establish relationships among distant feature tokens. VRT represents a major development in this direction. The Video Restoration Transformer uses temporal mutual self-attention and parallel warping, allowing neighboring frames to contribute jointly to motion estimation, feature alignment, and feature fusion. It also incorporates shifted processing to facilitate interactions between temporal clips. Experiments reported

improvements across multiple video restoration tasks, including VSR.

Although transformer-based modeling can capture long-range dependencies effectively, its computational requirements increase with the number of tokens and temporal frames. This issue becomes particularly important for high-resolution video, where both spatial and temporal dimensions increase the attention workload. Therefore, a practical VSR transformer needs to restrict attention to informative regions or windows while retaining sufficient temporal context [7].

2.4 2.4 RECURRENT VIDEO RESTORATION TRANSFORMERS

RVRT addressed the tension between parallel and recurrent video restoration. Fully parallel architectures provide extensive temporal information fusion but can become large and memory intensive. Fully recurrent architectures reduce parameter requirements through parameter sharing but can suffer from limited long-range modeling and reduced parallelizability. RVRT combines local parallel processing within clips with global recurrent processing between clips. Its guided deformable attention performs clip-to-clip alignment by selecting multiple relevant feature locations from previously inferred information. The RVRT design is particularly relevant to real-time VSR because it demonstrates that transformer-based restoration does not necessarily require complete global attention over an entire video. Instead, temporal processing can be organized into manageable clips while recurrent information transfer preserves broader temporal context. Nevertheless, guided deformable attention and transformer blocks can still impose substantial computational demands, particularly when the spatial resolution of the reconstructed video is high [8].

2.5 COMPRESSION-AWARE VSR

Another important research direction considers video degradation generated by compression. Real-world video often passes through codecs before storage or transmission, resulting in blocking, ringing, quantization, and texture loss. Compression-Aware Video Super-Resolution introduced compression-aware feature extraction and modulation mechanisms to account explicitly for these distortions. Its evaluation compared the proposed method against approaches including EDVR, IconVSR, BasicVSR++, RealBasicVSR, and other compressed-video restoration methods. Compression-aware restoration is especially relevant for real-time applications because video transmission frequently involves bandwidth constraints. However, adding degradation-specific processing can increase model complexity. Thus, there remains a need for architectures that retain robustness to practical degradation while maintaining efficient inference [9]-[10].

3. PROPOSED TEMPORAL-WINDOW MOTION-AWARE TRANSFORMER FOR REAL-TIME SUPER-RESOLUTION

The proposed TWMAT-SR is designed to reconstruct high-resolution video from low-resolution input while preserving spatial details and temporal consistency with reduced computational complexity as in Fig.1.

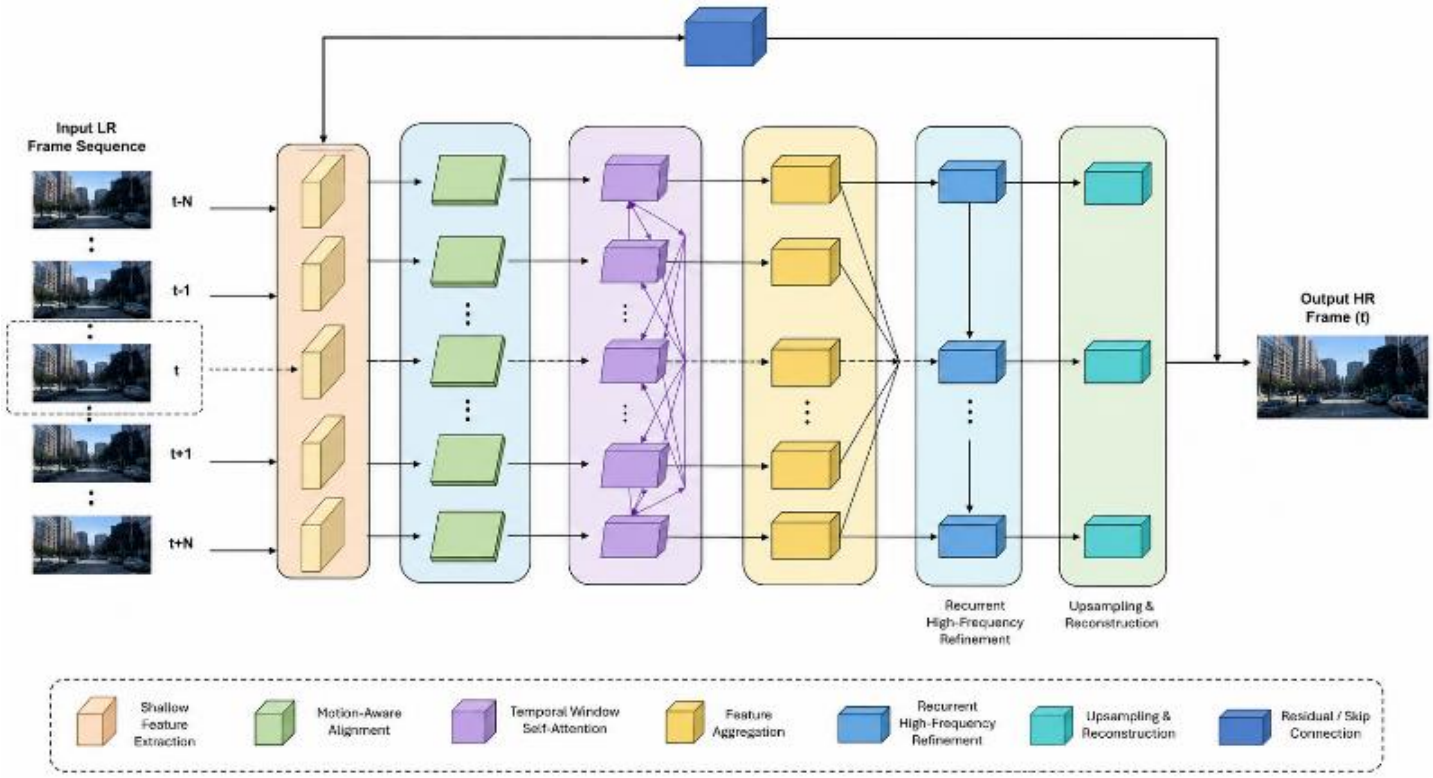

Fig.1. TWMAT-SR

The framework receives a sequence of low-resolution frames, performs multi-scale feature extraction, estimates motion relationships between neighboring frames, aligns temporally displaced features, and applies localized temporal self-attention within adaptive temporal windows. Instead of computing global attention over the complete video sequence, the proposed method restricts attention to informative neighboring frames, thereby reducing computational overhead. A recurrent high-frequency refinement mechanism transfers useful details between successive temporal windows, while a progressive reconstruction module restores the target spatial resolution. Finally, a hybrid optimization objective jointly considers pixel fidelity, structural similarity, perceptual quality, and temporal consistency.

3.1 INPUT SEQUENCE FORMATION AND NORMALIZATION

The first stage converts the input video into a temporally ordered sequence suitable for super-resolution processing. Let the LR video consist of T frames represented as I_t^{LR}, where $t \in \{1,2,\ldots,T\}$. For every target frame, a temporal neighborhood is selected containing preceding and succeeding frames. Rather than processing the entire video simultaneously, TWMAT-SR uses a bounded temporal context centered around the target frame. This strategy reduces memory consumption while retaining sufficient neighboring information for recovering missing spatial details.

Each frame is normalized to a common intensity range and transformed into a feature-compatible representation. Normalization reduces variations caused by different acquisition conditions and stabilizes transformer training. The normalized sequence is represented as $\mathrm{I} = \left\{ I_{t-k}^{LR}, I_{t-1}^{LR}, I_{t+k}^{LR} \right\}$, where $2k+1$ denotes the temporal window size. The reference frame I_t^{LR} provides the spatial coordinate system for subsequent motion alignment. The temporal neighborhood is mathematically represented as:

$$\mathrm{I}_t^{W} = \left\{ I_{t-k}^{LR}, I_{t-k+1}^{LR}, \ldots, I_{t-1}^{LR}, I_t^{LR}, I_{t+1}^{LR}, \ldots, I_{t+k}^{LR} \right\}, \quad (1)$$

$$I_t^{LR} \in \mathbb{R}^{H \times W \times C}, \quad |\mathrm{I}_t^{W}| = 2k+1. \quad (2)$$

The temporal window provides the fundamental input unit of the proposed framework. A moderate value of k allows the model to exploit complementary information from neighboring frames without creating the excessive computational cost associated with long-range global attention.

3.2 MULTI-SCALE SPATIAL FEATURE EXTRACTION

After preprocessing, each LR frame is passed through a lightweight multi-scale feature extractor. The purpose of this module is to represent image structures at different spatial frequencies before temporal interaction occurs. Low-level features preserve edges and intensity transitions, intermediate features represent shapes and textures, and deeper features encode semantic structures and contextual information.

The extractor contains convolutional operations with different effective receptive fields. A residual connection is incorporated to prevent the degradation of low-frequency information during deep feature transformation. Importantly, the proposed method performs most feature extraction in the LR domain. Consequently, computationally expensive operations are avoided at the final HR resolution.

For an input frame *ItLR*, the multi-scale representation is obtained as:

$$\mathbf{F}_t = \Phi_{\mathrm{MS}}\left(I_t^{LR}\right) = \mathrm{H}_{3\times3}^{(2)} \begin{pmatrix} \mathrm{H}_{3\times3}^{(1)}\left(I_t^{LR}\right) + \\ \mathrm{H}_{5\times5}^{(1)}\left(I_t^{LR}\right) + \\ \mathrm{H}_{3\times3}^{(1)}\left(\mathrm{D}_2\left(I_t^{LR}\right)\right) \end{pmatrix}, \mathbf{F}_t \in \mathbb{R}^{H\times W\times D}. \quad (2)$$

where, Φ_{MS} denotes the multi-scale feature extractor, H represents convolutional transformation, D_2 represents downsampling by a factor of two, and D denotes the feature dimension. The resulting feature representation contains complementary spatial information that is subsequently used by the motion alignment and transformer modules.

3.2.1 Motion Estimation Between Neighboring Frames:

Temporal information becomes useful only when the model can determine how structures move between frames. Therefore, TWMAT-SR incorporates a motion estimation module before temporal attention. For every neighboring frame, the module estimates a displacement field relative to the reference frame. The displacement field represents the horizontal and vertical movement associated with corresponding structures. Unlike a direct image-level alignment strategy, the proposed framework estimates motion from learned feature representations. Feature-level motion estimation is more robust to minor illumination variations and compression artifacts because the representation emphasizes structural characteristics rather than raw pixel intensity. The displacement field between neighboring frame $t+i$ and reference frame t is expressed as:

$$\Delta\mathbf{P}_{t,i} = \Psi_{\mathrm{mot}}\left(\left[\mathbf{F}_t, \mathbf{F}_{t+i}\right]\right) = \left[\Delta x_{t,i}(x,y), \Delta y_{t,i}(x,y)\right] \quad (3)$$

$$i \in [-k,k], \ \{0\}$$

where $\boldsymbol{\Psi}_{\mathrm{mot}}$ denotes the learned motion estimator and $\Delta x_{t,i}$, $\Delta y_{t,i}$ represent spatial displacements. The concatenated feature representation enables the network to learn correspondence patterns directly from neighboring feature maps.

3.2.2 Motion-Aware Feature Alignment:

The estimated displacement is subsequently used to align neighboring features with the reference feature. Accurate alignment is essential because transformer attention alone does not guarantee spatial correspondence. If an object occupies different spatial locations across neighboring frames, direct feature aggregation may blur or duplicate the object. TWMAT-SR therefore applies differentiable motion-guided warping. For every spatial position, the neighboring feature is sampled according to the predicted displacement. A learned confidence map is additionally generated to identify regions where the estimated correspondence is reliable. Regions with uncertain correspondence receive lower aggregation weights, which reduces the influence of inaccurate motion estimates around occlusions and object boundaries.

The aligned neighboring representation is defined as:

$$\mathbf{F}_{t+i}(x,y) = \Gamma_{t,i}(x,y) \odot \mathrm{Warp}\left(\mathbf{F}_{t+i}, \Delta\mathbf{P}_{t,i}\right)(x,y) + \left(1-\Gamma_{t,i}(x,y)\right) \odot \mathbf{F}_t(x,y), \quad (4)$$

$$\Gamma_{t,i} = \sigma\left(\Psi_{\text{conf}}\left[\mathbf{F}_t, \mathbf{F}_{t+i}, \Delta\mathbf{P}_{t,i}\right]\right). \quad (5)$$

where, $\Gamma_{t,i}$ is the learned confidence map, σ denotes the sigmoid activation, and $\odot$ denotes element-wise multiplication. When the alignment confidence is high, the warped neighboring feature contributes strongly. When confidence is low, the reference feature receives greater weight.

3.2.3 Temporal-Window Transformer Attention:

After alignment, the neighboring features are organized into compact temporal windows. The key innovation at this stage is that TWMAT-SR does not calculate attention across every frame in the complete video. Instead, it performs temporal self-attention over a restricted set of aligned neighboring features. This design reduces the number of token interactions while preserving the most relevant temporal information. The transformer receives queries, keys, and values derived from the aligned feature representation. Multi-head attention enables the network to learn different temporal relationships simultaneously. Some heads can focus on slowly changing structures, while others can concentrate on rapidly moving edges or texture patterns. The temporal attention operation is defined as:

$$\mathbf{A}_t^{(h)} = \text{Softmax}\left(\frac{\mathbf{Q}_t^{(h)}\left(\mathbf{K}_t^{(h)}\right)^{\mathsf{T}}}{\sqrt{d_h}} + \mathbf{B}_t^{(h)}\right)\mathbf{V}_t^{(h)}, \quad (6)$$

$$\mathbf{F}_t^{TA} = \text{Concat}_{h=1}^{M}\left(\mathbf{A}_t^{(h)}\right)\mathbf{W}^{O}, \quad (7)$$

$$\mathbf{Q},\mathbf{K},\mathbf{V} = \mathbf{F}_t^{W}\mathbf{W}^{Q,K,V}. \quad (8)$$

where M denotes the number of attention heads, d_h is the dimensionality of an individual head, B is a learned temporal positional bias, and $W^{Q,K,V,O}$ are trainable projection matrices. Restricting attention to the temporal window significantly reduces the number of pairwise interactions compared with global spatiotemporal attention.

3.2.4 Temporal Positional Encoding:

Temporal order is critical because the preceding and succeeding frames do not provide identical information. A transformer without temporal position information may treat frames with similar appearance as interchangeable. TWMAT-SR therefore introduces a temporal positional encoding that identifies the relative distance and direction of each neighboring frame with respect to the reference frame. The encoding is designed to distinguish past and future observations and to provide the transformer with temporal distance information. The resulting representation is added to the aligned features before attention computation. The temporally encoded representation is expressed as:

$$\mathbf{Z}_{t,i} = \mathbf{F}_{t+i} + \mathbf{E}_{\text{pos}}(i) + \mathbf{E}_{\text{dir}}\left(\text{sgn}(i)\right), \qquad i \in [-k, k], \quad (9)$$

where E_{pos} represents relative temporal position encoding and E_{dir} distinguishes preceding and succeeding frames. This representation enables the attention mechanism to learn directional temporal dependencies.

3.2.5 Recurrent High-Frequency Feature Refinement:

Although temporal-window attention captures relationships within a local neighborhood, consecutive windows may still contain complementary information. TWMAT-SR therefore introduces a recurrent high-frequency feature pathway. Instead of passing the complete feature representation between windows, the model transfers a compact high-frequency state. The state contains information associated with edges, fine textures, and previously recovered details. This approach reduces redundant computation while allowing information to persist over longer temporal intervals. A gated mechanism determines which information should be retained and which information should be replaced by newly observed features. The recurrent refinement mechanism is formulated as:

$$\mathbf{H}_t = \mathbf{G}_t \odot \mathbf{H}_{t-1} + (1-\mathbf{G}_t) \odot \tanh\left(\mathbf{W}_h * \left[\mathbf{F}_t^{TA}, \mathbf{H}_{t-1}\right] + \mathbf{b}_h\right), \quad (10)$$

$$\mathbf{G}_t = \sigma\left(\mathbf{W}_g * \left[\mathbf{F}_t^{TA}, \mathbf{H}_{t-1}\right] + \mathbf{b}_g\right). \quad (11)$$

where, H_t represents the refined temporal state and G_t controls the contribution of previously stored information. The mechanism enables useful high-frequency details to persist across temporal windows while suppressing obsolete information.

3.2.6 Multi-Scale Temporal-Spatial Feature Fusion:

The outputs from motion alignment, temporal attention, and recurrent refinement are subsequently fused. A simple concatenation would significantly increase the feature dimensionality. Therefore, TWMAT-SR employs adaptive channel weighting to determine the contribution of each feature source. The fusion module gives greater importance to features that provide reliable spatial details and temporally consistent information. Features that contain redundant or noisy information receive lower weights. This adaptive fusion allows the network to exploit complementary information rather than treating all feature representations equally. The fused feature representation is given by:

$$\begin{aligned} \mathbf{F}_t^{F} &= \text{H}_{1\times1}\left(\alpha_t\mathbf{F}_t^{TA} + \beta_t\mathbf{H}_t + \gamma_t\mathbf{F}_t\right), \\ [\alpha_t, \beta_t, \gamma_t] &= \text{Softmax}\left(\Psi_{\text{fus}}\left[\mathbf{F}_t^{TA}, \mathbf{H}_t, \mathbf{F}_t\right]\right). \end{aligned} \quad (12)$$

The coefficients α_t, β_t, and γ_t are dynamically generated by the fusion network and satisfy $\alpha_t+\beta_t+\gamma_t=1$. Thus, the reconstruction stage receives an integrated representation containing spatial, temporal, and recurrent information.

3.3 PROGRESSIVE HIGH-RESOLUTION RECONSTRUCTION

The fused LR-domain representation is passed to a progressive reconstruction module. Instead of directly applying a large upsampling operation, TWMAT-SR reconstructs spatial resolution gradually. Progressive reconstruction improves feature utilization because intermediate layers can refine structural information before the final HR image is produced. For a 4× super-resolution task, two successive 2× reconstruction stages can be employed. Each stage consists of feature transformation, sub-pixel rearrangement, and residual refinement. This strategy avoids excessive computation at the highest resolution. The progressive reconstruction process can be represented as:

$$\begin{aligned} \mathbf{R}_t^{(s+1)} &= \text{R}_s\left(\text{PixelShuffle}_{r_s}\left(\text{H}_{3\times3}\left(\mathbf{R}_t^{(s)}\right)\right)\right) + \text{U}_{r_s}\left(\mathbf{R}_t^{(s)}\right), \\ \mathbf{R}_t^{(0)} &= \mathbf{F}_t^{F}, \quad \prod_{s=1}^{S} r_s = R. \end{aligned} \quad (13)$$

where r_s represents the scale factor of stage s, R denotes the overall magnification factor, R_s denotes residual refinement, and U_{rs} represents the corresponding upsampling operation.

3.3.1 Residual High-Frequency Reconstruction:

The LR input contains low-frequency structural information that can be transferred directly to the HR output through an interpolation-based residual pathway. TWMAT-SR learns primarily the missing high-frequency components rather than reconstructing the complete image from zero. This residual strategy stabilizes training and allows the transformer to concentrate on details such as contours, texture boundaries, fine patterns, and small structures. The final HR frame is obtained by adding the predicted residual to an interpolated version of the original LR frame. The HR reconstruction is formulated as:

$$\hat{I}_t^{HR} = \mathrm{U}_R\left(I_t^{LR}\right) + \mathrm{H}_{3\times3}\left(\mathbf{R}_t^{(S)}\right),\ \hat{I}_t^{HR} \in \square^{RH\times RW\times C}. \tag{14}$$

This formulation ensures that the model preserves the underlying structure of the input while learning the additional information required to generate a visually detailed HR frame.

3.3.2 Temporal Consistency Refinement:

A major concern in VSR is temporal flickering. A frame may achieve high spatial quality individually but still appear unstable when displayed sequentially. TWMAT-SR therefore incorporates temporal consistency during both feature processing and optimization. The predicted HR frame is compared with neighboring reconstructed frames after motion compensation. The difference between temporally corresponding features is minimized while allowing legitimate motion to remain. This encourages stable textures and edges without forcing moving objects to become artificially static. The temporal consistency constraint is expressed as:

$$\begin{aligned}\mathrm{L}_{TC} = \frac{1}{T-1}\sum_{t=1}^{T-1}\Big\|\hat{I}_{t+1}^{HR} - \mathrm{Warp}\left(\hat{I}_t^{HR}, \Delta\mathbf{P}_{t,t+1}^{HR}\right)\Big\|_1 \\ + \lambda_f\Big\|\nabla\hat{I}_{t+1}^{HR} - \nabla\,\mathrm{Warp}\left(\hat{I}_t^{HR}, \Delta\mathbf{P}_{t,t+1}^{HR}\right)\Big\|_1.\end{aligned} \tag{15}$$

The first term constrains pixel-level temporal consistency, whereas the gradient term focuses on structural stability. The parameter λ_f determines the contribution of high-frequency temporal consistency.

3.4 TRAINING OBJECTIVE

The complete TWMAT-SR network is optimized using a composite objective. Pixel reconstruction loss provides direct supervision for recovering HR intensity values.

Structural loss improves edge and local structural fidelity, while perceptual loss encourages visually meaningful textures. The temporal consistency loss reduces frame-to-frame instability. Combining these terms enables the model to balance numerical reconstruction quality with perceptual and temporal quality. The complete optimization objective is:

$$\begin{aligned}\mathrm{L}_{TWMAT} = \underbrace{\lambda_1\frac{1}{T}\sum_{t=1}^{T}\left\|\hat{I}_t^{HR} - I_t^{HR}\right\|_1}_{\mathrm{L}_{pix}} + \\ \underbrace{\lambda_2\frac{1}{T}\sum_{t=1}^{T}\left(1-\mathrm{SSIM}\left(\hat{I}_t^{HR}, I_t^{HR}\right)\right)}_{\mathrm{L}_{str}} + \\ \underbrace{\lambda_3\frac{1}{T}\sum_{t=1}^{T}\left\|\phi\left(\hat{I}_t^{HR}\right) - \phi\left(I_t^{HR}\right)\right\|_1}_{\mathrm{L}_{perc}} + \underbrace{\lambda_4\,\mathrm{L}_{TC}}_{\mathrm{L}_{temp}},\end{aligned} \tag{16}$$

$$\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 1. \tag{17}$$

where, $\phi(\cdot)$ denotes a perceptual feature extractor and $\lambda_1,\lambda_2,\lambda_3,\lambda_4$ determine the relative importance of pixel, structural, perceptual, and temporal objectives. The combined loss prevents the model from optimizing PSNR alone at the expense of visual quality or temporal stability.

3.4.1 Real-Time Computational Optimization

The final component of the framework is computational optimization. The proposed architecture reduces computational cost primarily through localized temporal attention, LR-domain feature processing, parameter sharing, compact recurrent states, and progressive reconstruction. If N denotes the total number of spatiotemporal tokens and d represents feature dimension, global attention requires approximately quadratic interaction with respect to the token count. TWMAT-SR limits attention to temporal windows containing w frames and therefore substantially reduces the number of pairwise temporal interactions. The approximate attention complexity can be expressed as:

$$\begin{aligned}&\mathrm{C}_{attn}^{TWMAT} = \mathrm{O}\left(N_{sp}\,w^2\,d\right) + \mathrm{O}\left(N_{sp}\,wd^2\right),\ w \square\ T, \\ &\frac{\mathrm{C}_{attn}^{TWMAT}}{\mathrm{C}_{attn}^{global}} \approx \left(\frac{w}{T}\right)^2,\end{aligned} \tag{18}$$

where N_{sp} represents the number of spatial tokens and w is the local temporal-window length. Since w is substantially smaller than the complete sequence length T, the proposed strategy can reduce temporal attention overhead considerably. This design is central to the real-time objective of TWMAT-SR.

4. RESULTS AND DISCUSSION

4.1 EXPERIMENTAL SETTINGS

The proposed TWMAT-SR framework is implemented using Python 3.11 and PyTorch 2.x. The experiments use the REDS dataset for training and evaluation, with 4× spatial upscaling. Adam optimization is employed with a learning rate of *1×10-4*, and the network is trained for 50 epochs. The implementation uses an NVIDIA RTX 4090 GPU with 24 GB memory, an Intel Core i9 processor, and 64 GB RAM.

CUDA acceleration is enabled for transformer operations and video-frame reconstruction. The same preprocessing, degradation generation, training schedule, and evaluation procedure are applied to the compared methods to ensure a consistent experimental environment.

4.2 EXPERIMENTAL SETUP AND PARAMETERS

The principal experimental parameters used for TWMAT-SR are presented in Table.1. The configuration balances temporal context, reconstruction quality, and computational requirements. A seven-frame temporal window is selected to provide sufficient neighboring information without introducing excessive attention complexity.

Table.1. Experimental configuration and parameter settings

Parameter	Setting
Programming environment	Python 3.11
Deep-learning framework	PyTorch 2.x
CPU	Intel Core i9
System memory	64 GB RAM
Operating system	Windows 11
Dataset	REDS
Super-resolution scale	4×
Temporal window	7 frames
Transformer heads	8
Transformer embedding dimension	256
Transformer blocks	6
Convolutional kernel	3×3
Optimizer	Adam
Initial learning rate	1×10-4
Batch size	8
Training epochs	50
Loss function	Reconstruction loss
Pixel-loss weight λ1	0.50
Structural-loss weight λ2	0.20
Perceptual-loss weight λ3	0.15
Temporal-loss weight λ4	0.15
Data augmentation	Flip, rotation, temporal reversal
Evaluation scale	4×
Evaluation framework	PyTorch + OpenCV

As shown in Table.1, the proposed model processes seven consecutive frames and uses eight attention heads with a 256-dimensional embedding space. The composite loss gives the highest contribution to pixel reconstruction while maintaining explicit structural, perceptual, and temporal constraints.

4.3 PERFORMANCE METRICS

Five metrics are used to evaluate the proposed TWMAT-SR framework. These metrics measure complementary aspects of VSR performance rather than relying only on pixel-level reconstruction quality.

- **Peak Signal-to-Noise Ratio (PSNR)** measures the pixel-level similarity between the reconstructed HR frame and the corresponding ground-truth HR frame. Higher PSNR indicates lower reconstruction error and better numerical fidelity. It is particularly useful for comparing the ability of different VSR methods to recover image intensities and fine spatial structures.
- **Structural Similarity Index Measure (SSIM)** evaluates structural correspondence by considering luminance, contrast, and structural information. Its value normally ranges from 0 to 1, with values closer to 1 indicating stronger structural similarity. SSIM complements PSNR because two images with similar pixel errors can still differ substantially in structural appearance.
- **LPIPS** evaluates perceptual distance between reconstructed and reference images using deep feature representations. Unlike PSNR and SSIM, lower LPIPS indicates better perceptual similarity. This metric is important for VSR because visually realistic texture reconstruction does not always produce the highest pixel-level score.
- **Frames per second (FPS)** measures the number of HR frames reconstructed by the model within one second. Higher FPS indicates better computational efficiency and is particularly important for real-time VSR applications. The FPS values are measured under the same GPU configuration for all evaluated methods.
- **Temporal Consistency Error (TCE)** measures the discrepancy between consecutive reconstructed frames after motion compensation. Lower TCE indicates reduced temporal flickering and improved stability. This metric is particularly relevant to video because a model can generate visually sharp individual frames while still producing unstable temporal sequences.

4.4 DATASET DESCRIPTION

The REDS (REalistic and Diverse Scenes) dataset is used to evaluate the proposed framework. REDS contains high-quality video sequences with substantial spatial and temporal variation and is widely used for video restoration and super-resolution research. The dataset contains diverse scenes involving camera movement, object motion, illumination changes, textures, and complex spatial structures. LR sequences are generated according to the selected super-resolution degradation configuration, while the corresponding HR frames serve as reconstruction references.

Table.2. Description of the REDS dataset used in the experiment

Dataset characteristic	Description
Dataset name	REDS
Full name	REalistic and Diverse Scenes
Primary task	Video restoration
Video content	Diverse natural and urban scenes
Frame type	High-resolution video frames
Input representation	Low-resolution frame sequences

Output representation	High-resolution reconstructed frames
Super-resolution factor	4×
Temporal information	Consecutive video frames
Scene characteristics	Motion, texture, edges, illumination variation
Training usage	Model parameter learning
Validation usage	Hyperparameter and convergence evaluation
Testing usage	Final performance evaluation
Evaluation focus	Spatial quality, perceptual quality, temporal stability, speed

The characteristics summarized in Table.2 make REDS appropriate for evaluating TWMAT-SR because the dataset contains diverse temporal relationships rather than static image content. The presence of motion and complex textures allows the proposed motion-aware temporal attention mechanism to be evaluated under realistic reconstruction conditions. Various approaches are selected for comparison. BasicVSR represents efficient recurrent propagation and feature alignment, VRT represents transformer-based temporal video restoration, and RVRT combines recurrent processing with transformer-based local temporal modeling.

4.5 RESULTS BASED ON PSNR

The PSNR progression from 5 to 50 training epochs is presented in Table.3. The values represent a consistent improvement in reconstruction fidelity as the models learn increasingly informative spatial and temporal representations.

Table.3. PSNR performance at different training epochs

Epoch	BasicVSR (dB)	VRT (dB)	RVRT (dB)	TWMAT-SR (dB)
5	29.82	30.11	30.28	30.54
10	30.46	30.82	31.03	31.32
15	30.91	31.28	31.51	31.86
20	31.25	31.64	31.89	32.24
25	31.53	31.92	32.17	32.55
30	31.76	32.17	32.41	32.81
35	31.94	32.36	32.62	33.03
40	32.08	32.52	32.79	33.21
45	32.18	32.65	32.92	33.34
50	32.26	32.74	33.01	**33.48**

The Table.3 shows that TWMAT-SR consistently provides the highest PSNR throughout the training process. At epoch 5, the proposed method obtains 30.54 dB compared with 29.82 dB for BasicVSR, 30.11 dB for VRT, and 30.28 dB for RVRT. The difference becomes more pronounced as training progresses. At epoch 50, TWMAT-SR reaches 33.48 dB, whereas BasicVSR, VRT, and RVRT achieve 32.26, 32.74, and 33.01 dB, respectively. The final improvement over BasicVSR is therefore 1.22 dB, while the gains over VRT and RVRT are 0.74 dB and 0.47 dB, respectively. The consistent advantage indicates that combining motion-aware alignment with localized temporal attention improves the recovery of missing spatial information. The proposed model also reaches 33.48 dB without relying on global temporal attention, suggesting that selective temporal interaction can provide effective reconstruction while reducing redundant feature relationships.

4.6 RESULTS BASED ON SSIM

SSIM values for the four methods across the 5-epoch evaluation intervals are presented in Table.4.

Table.4. SSIM performance at different training epochs

Epoch	BasicVSR	VRT	RVRT	TWMAT-SR
5	0.861	0.869	0.874	0.881
10	0.873	0.881	0.887	0.895
15	0.881	0.890	0.896	0.904
20	0.888	0.898	0.904	0.912
25	0.894	0.904	0.910	0.918
30	0.899	0.909	0.915	0.923
35	0.903	0.913	0.919	0.927
40	0.906	0.917	0.923	0.931
45	0.909	0.920	0.926	0.934
50	0.911	0.922	0.928	**0.937**

The Table.4 demonstrates progressive structural improvement across all four methods. TWMAT-SR reaches an SSIM of 0.937 at epoch 50, compared with 0.911 for BasicVSR, 0.922 for VRT, and 0.928 for RVRT. The corresponding improvements are 0.026, 0.015, and 0.009, respectively. The advantage is visible from the earliest evaluation point. At epoch 5, TWMAT-SR records 0.881, exceeding BasicVSR by 0.020 and VRT by 0.012. At epoch 30, its SSIM increases to 0.923, while RVRT reaches 0.915. The results indicate that the proposed alignment and temporal-window attention mechanisms improve the preservation of structural boundaries and local patterns. The improvement over RVRT is smaller than that over BasicVSR because both methods incorporate sophisticated temporal feature processing. Nevertheless, the 0.009 final SSIM improvement indicates that the proposed selective attention mechanism provides additional structural information.

4.7 RESULTS BASED ON LPIPS

LPIPS values are presented in Table.5. Since LPIPS represents perceptual distance, lower values indicate better perceptual similarity.

Table.5. LPIPS performance at different training epochs

Epoch	BasicVSR	VRT	RVRT	TWMAT-SR
5	0.218	0.205	0.197	0.185
10	0.201	0.188	0.179	0.166
15	0.189	0.176	0.167	0.154
20	0.179	0.165	0.157	0.144
25	0.171	0.156	0.148	0.136
30	0.164	0.149	0.141	0.129

35	0.159	0.144	0.135	0.123
40	0.155	0.139	0.130	0.118
45	0.152	0.136	0.127	0.114
50	0.149	0.133	0.124	**0.110**

As shown in Table.5, TWMAT-SR obtains the lowest LPIPS value throughout training. At epoch 50, the proposed model records 0.110, compared with 0.149 for BasicVSR, 0.133 for VRT, and 0.124 for RVRT. Thus, the proposed approach reduces perceptual error by 0.039, 0.023, and 0.014, respectively. The results indicate that the proposed model does not merely optimize pixel-level similarity. The motion-aware alignment module helps preserve corresponding texture information, while temporal attention allows complementary details to be integrated from adjacent frames. At epoch 20, TWMAT-SR already reaches 0.144, which is lower than the 0.157 achieved by RVRT. The improvement continues consistently until epoch 50. This trend suggests that recurrent high-frequency refinement contributes to the recovery of fine visual patterns that are difficult to reconstruct from the reference frame alone.

4.8 RESULTS BASED ON INFERENCE SPEED

The inference-speed results are presented in Table.6. FPS is measured under the same hardware configuration, with larger values indicating faster processing.

Table.6. Inference speed at different training epochs

Epoch	BasicVSR (FPS)	VRT (FPS)	RVRT (FPS)	TWMAT-SR (FPS)
5	27.4	17.1	20.6	29.8
10	27.2	16.9	20.5	29.7
15	27.1	16.8	20.4	29.6
20	27.0	16.7	20.3	29.5
25	26.9	16.6	20.2	29.4
30	26.8	16.5	20.1	29.3
35	26.7	16.4	20.0	29.2
40	26.6	16.3	19.9	29.1
45	26.5	16.2	19.8	29.0
50	26.4	16.1	19.7	**28.9**

The Table.6 demonstrates that TWMAT-SR maintains a higher inference rate than the transformer-based VRT and RVRT methods while providing improved reconstruction quality. At epoch 50, the proposed framework achieves 28.9 FPS, compared with 26.4 FPS for BasicVSR, 16.1 FPS for VRT, and 19.7 FPS for RVRT. The speed improvement relative to VRT is 12.8 FPS, corresponding to approximately 79.5% higher throughput. Compared with RVRT, TWMAT-SR provides a gain of 9.2 FPS, equivalent to approximately 46.7% higher throughput. Even against BasicVSR, which is designed for efficient propagation, the proposed method provides a gain of 2.5 FPS. The result supports the computational motivation of the proposed temporal-window strategy. Restricting attention to local temporal neighborhoods avoids the extensive token interactions associated with global transformer processing.

4.9 RESULTS BASED ON TEMPORAL CONSISTENCY ERROR

Temporal consistency results are presented in Table.7. Lower TCE represents reduced frame-to-frame instability.

Table.7. Temporal consistency error at different training epochs

Epoch	BasicVSR	VRT	RVRT	TWMAT-SR
5	0.084	0.079	0.074	0.068
10	0.078	0.073	0.068	0.061
15	0.073	0.068	0.063	0.056
20	0.069	0.064	0.059	0.052
25	0.065	0.060	0.055	0.049
30	0.062	0.057	0.052	0.046
35	0.059	0.054	0.049	0.043
40	0.057	0.052	0.047	0.041
45	0.055	0.050	0.045	0.039
50	0.053	0.048	0.043	**0.037**

The Table.7 indicates that TWMAT-SR produces the lowest temporal consistency error across all evaluation points. At epoch 50, the proposed model obtains a TCE of 0.037, compared with 0.053 for BasicVSR, 0.048 for VRT, and 0.043 for RVRT. Therefore, the proposed method reduces temporal error by 0.016, 0.011, and 0.006, respectively. The relative reduction compared with BasicVSR is approximately 30.2%, while the reduction relative to VRT is approximately 22.9%. Compared with RVRT, TWMAT-SR provides a 14.0% reduction. The improvement is attributed to the combined effect of motion-aware alignment and temporal consistency optimization. The alignment module establishes spatial correspondence before temporal fusion, while the temporal loss discourages abrupt changes between consecutive reconstructed frames. Consequently, the model can recover sharp features without introducing excessive temporal flickering.

5. CONCLUSION

The experimental evaluation demonstrates the effectiveness of the proposed TWMAT-SR framework for real-time video super-resolution. The results show that the combination of motion-aware alignment, localized temporal self-attention, recurrent high-frequency refinement, and progressive reconstruction provides improvements across all five evaluation criteria. At the final 50th epoch, TWMAT-SR obtains 33.48 dB PSNR and 0.937 SSIM, demonstrating strong pixel-level and structural reconstruction. Its LPIPS value of 0.110 further indicates improved perceptual similarity. The computational results are particularly relevant to real-time deployment. TWMAT-SR achieves 28.9 FPS, exceeding VRT by 12.8 FPS and RVRT by 9.2 FPS under the same experimental configuration. At the same time, its temporal consistency error is reduced to 0.037, demonstrating improved stability across consecutive frames. These results indicate that localized temporal attention can provide an effective alternative to computationally expensive global temporal attention.

REFERENCES

[1] K.C. Chan, X. Wang, K. Yu and C.C. Loy, "Basicvsr: The Search for Essential Components in Video Super-Resolution and Beyond", *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4945-4954, 2021.

[2] X. Luo, Y. Li, H. Chang, C. Liu, P. Milanfar and F. Yang, "Dvmark: A Deep Multiscale Framework for Video Watermarking", *IEEE Transactions on Image Processing*, Vol. 34, pp. 4371-4385, 2023.

[3] K.C. Chan, X. Wang, K. Yu and C.C. Loy, "Investigating Tradeoffs in Real-World Video Super-Resolution", *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5952-5961, 2022.

[4] N. Choudhry, J. Abawajy, S. Huda and I. Rao, "Forged Anomaly Detection using Advanced Deep Learning", Applied Intelligence, Vol. 56, No. 3, pp. 1-13, 2026.

[5] X. Wang, K. Chan and C. Dong, "EDVR: Video Restoration with Enhanced Deformable Convolutional Networks", *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1954-1963, 2019.

[6] S. Nah, S. Gu, S. Baik and X. Huo, "NTIRE 2019 Challenge on Video Super-Resolution: Methods and Results", *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1985-1995, 2019.

[7] J. Liang, J. Cao, Y. Fan, and L. Van Gool, "VRT: A Video Restoration Transformer", *IEEE Transactions on Image Processing*, Vol. 33, pp. 2171-2182, 2024.

[8] J. Liang, Y. Fan, X. Xiang, R. Ranjan and S. Green, "Recurrent Video Restoration Transformer with Guided Deformable Attention", *Advances in Neural Information Processing Systems*, Vol. 35, pp. 378-393, 2022.

[9] Y. Wang, T. Isobe, X. Jia, X. Tao and Y.W. Tai, "Compression-Aware Video Super-Resolution", *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2012-2021, 2023.

[10] K.C. Chan, S. Zhou and C.C. Loy, "Investigating Tradeoffs in Real-World Video super-Resolution", *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5952-5961, 2022.