ISSN: 0976-9102 (ONLINE)
DOI: 10.21917/ijivp.2026.0535

RECURRENT TRANSFORMER BASED FRAMEWORK FOR VIDEO DENOISING AND SUPER RESOLUTION USING OPTICAL FLOW AND TEMPORAL ATTENTION

Pitty Nagarjuna[1] and B.K. Harsha[2]
[1]JRD Tata Memorial Library, Indian Institute of Science, Bengaluru, India
[2]School of Computer Science and Engineering, REVA University, India

Abstract

Video restoration has remained an important task in multimedia processing because visual data captured in real environments often contain noise, motion artifacts, and resolution degradation. The demand for high-quality video has increased with the growth of surveillance systems, streaming platforms, and intelligent vision applications. Traditional denoising and super-resolution approaches have relied on spatial filtering and convolutional neural networks. However, these techniques have faced limitations in modeling long-range temporal dependencies across frames. As a result, inconsistent textures, motion blur, and temporal flickering have frequently appeared in restored videos. The present study has addressed these challenges by introducing a Recurrent Optical Flow Transformer (ROFT), a recurrent transformer architecture that has integrated optical flow estimation with temporal attention for joint video denoising and super-resolution. The proposed framework has utilized a recurrent transformer module that has captured temporal correlations between adjacent frames while maintaining spatial consistency. An optical flow estimation unit has guided the alignment of frames, which has reduced motion distortion and misalignment during reconstruction. In addition, a temporal attention mechanism that has analyzed contextual dependencies across multiple frames has enhanced feature representation for dynamic regions. The network has processed sequential frames through recurrent connections that have preserved temporal memory and improved reconstruction stability. Experiments have been conducted on benchmark video restoration datasets that contained noisy and low-resolution sequences. The experimental evaluation demonstrates that the proposed ROFTT framework achieves superior performance compared with existing approaches. The model produces a PSNR value of 35.8 dB and an SSIM value of 0.97, which indicate improved reconstruction quality and structural preservation. The reconstruction error decreases to 0.005 MSE, while the temporal consistency error reduces to 0.007, which confirms stable frame transitions across video sequences. Furthermore, the model achieves an FSIM value of 0.995, which indicates strong preservation of perceptual texture features. These results demonstrate that the proposed architecture effectively integrates optical flow alignment and temporal transformer attention that enhances both spatial detail recovery and temporal coherence in restored video frames.

Keywords:

Video Denoising, Super Resolution, Recurrent Transformer, Optical Flow Estimation, Temporal Attention

1. INTRODUCTION

The rapid expansion of multimedia technologies has increased the importance of high-quality video processing in modern computer vision systems. Video data that originate from surveillance cameras, mobile devices, medical imaging tools, and autonomous systems often contain noise, compression artifacts, and resolution degradation. These distortions reduce the visual clarity of frames and affect the reliability of automated analysis systems. Consequently, video restoration methods that focus on denoising and super-resolution have become critical components in multimedia and vision pipelines. Several early studies have explored spatial filtering and statistical modeling approaches that improve the visual quality of degraded frames. These studies have demonstrated that spatial feature reconstruction can recover missing details to a certain extent, although the restored frames often remain inconsistent across time [1–3].

Recent advances in deep learning have significantly improved the performance of video restoration algorithms. Convolutional neural networks have been widely adopted for image and video super-resolution tasks because they effectively learn hierarchical representations from large datasets. Networks that incorporate temporal information across adjacent frames have further improved reconstruction quality. Such architectures have utilized temporal correlations that exist between consecutive frames to recover fine details that a single frame cannot capture independently. However, convolution-based methods often focus primarily on local spatial features, which limits their ability to model long-range dependencies across time. Transformer architectures have recently emerged as promising alternatives because attention mechanisms can capture global contextual relationships across frames and spatial regions. Several researchers have therefore explored transformer-based video restoration frameworks that analyze temporal dependencies more effectively [1–3].

Despite these advances, several challenges still remain in the restoration of real-world video sequences. One major difficulty involves the presence of dynamic motion that occurs between consecutive frames. Motion variations caused by camera movement or object displacement introduce misalignment among frames. When these frames are directly combined during - reconstruction, the process often produces blurred textures and ghosting artifacts. Optical flow techniques have been widely used to estimate motion fields that align frames before feature fusion. However, optical flow estimation itself can become inaccurate in scenes that contain occlusions, complex motion, or illumination variations. When the alignment process fails, the reconstruction stage produces inconsistent textures and visible artifacts [4–5].

Another important challenge arises from the temporal inconsistency that appears in restored videos. Many restoration algorithms process frames independently or rely on limited temporal windows. As a result, the reconstructed frames may exhibit flickering artifacts when viewed sequentially. This issue becomes particularly visible in high-resolution video restoration tasks where subtle differences between frames become noticeable to the human eye. Additionally, traditional architectures often lack a mechanism that preserves temporal memory across longer sequences. Without such memory, the model struggles to maintain structural continuity across frames that appear at different temporal positions. These issues highlight the need for advanced architectures that integrate motion estimation, temporal reasoning, and spatial reconstruction within a unified framework [4–5].

Existing research has attempted to address these problems through hybrid architectures that combine convolutional modules with attention mechanisms. Although these methods have achieved promising improvements, several limitations still persist. First, many models treat optical flow estimation and feature reconstruction as separate stages. This separation often leads to suboptimal alignment because the reconstruction network does not directly influence the motion estimation process. Second, most transformer-based methods process video sequences in a feedforward manner without maintaining recurrent temporal memory. Such designs limit the ability of the model to capture long-term temporal relationships that extend beyond a small set of frames. Consequently, noise removal and detail recovery remain inconsistent in dynamic scenes where temporal information plays a crucial role. The research problem therefore lies in the development of an integrated framework that jointly performs video denoising and super-resolution while maintaining temporal consistency across frames. A robust solution must effectively align frames that contain complex motion while simultaneously extracting spatial and temporal features that contribute to high-quality reconstruction. Furthermore, the architecture must preserve contextual memory that enables the model to analyze long sequences of frames without losing temporal coherence. Addressing these requirements remains a critical step toward reliable video restoration systems that support real-world multimedia applications [6–7].

To address the above problem, this study introduces a recurrent transformer architecture that integrates optical flow estimation with temporal attention mechanisms. The proposed framework aims to enhance both spatial reconstruction and temporal consistency within degraded video sequences. The model incorporates a recurrent processing strategy that maintains temporal memory across frames. Optical flow alignment assists the network in compensating for motion variations, while the transformer attention module analyzes contextual relationships that exist between frames. This integrated design supports joint denoising and super-resolution within a unified learning framework.

The primary objectives of this research include the following. First, the study aims to design a recurrent transformer architecture that captures temporal dependencies across video frames. Second, the work aims to integrate optical flow-guided alignment with attention-based feature modeling for improved reconstruction accuracy. Third, the study evaluates the proposed model on benchmark datasets to assess improvements in visual quality and temporal stability compared with existing methods.

The novelty of the proposed framework lies in the integration of recurrent temporal learning with transformer attention and optical flow alignment within a single architecture. Unlike conventional feedforward transformer models, the recurrent mechanism preserves temporal context that supports stable frame reconstruction across long sequences. In addition, the joint optimization of motion estimation and feature reconstruction improves alignment accuracy, which ultimately enhances the restoration performance.

The contributions of this research can be summarized as follows:

- A novel recurrent transformer framework has been introduced for simultaneous video denoising and super-resolution that incorporates optical flow-guided temporal attention for improved motion-aware reconstruction.
- An integrated temporal learning strategy has been developed that maintains contextual memory across frames, which has significantly improved temporal consistency and reduced flickering artifacts in restored videos.

2. RELATED WORKS

Several studies have explored video restoration methods that focus on denoising and super-resolution through deep learning architectures. Early approaches have primarily relied on convolutional neural networks that exploit spatial correlations within frames. Researchers have demonstrated that multi-frame convolutional architectures improve reconstruction accuracy by combining information from neighboring frames. For instance, the work presented in [7] has introduced a multi-frame convolutional model that utilizes motion compensation before feature fusion. The framework has aligned adjacent frames using optical flow estimation and has aggregated spatial features through convolutional layers. Experimental results have indicated that the model has improved reconstruction accuracy compared with single-frame approaches. However, the method has relied heavily on accurate optical flow estimation, which has limited its robustness in scenes that contain complex motion.

Subsequent research has investigated the integration of attention mechanisms within video restoration networks. The study described in [8] has proposed an attention-guided video super-resolution architecture that has utilized spatial and channel attention modules. The attention mechanism has emphasized informative regions within frames while suppressing irrelevant noise patterns. Through this design, the network has reconstructed sharper textures and fine structural details. Although the method has achieved promising results, it has focused primarily on spatial attention and has not fully addressed temporal consistency across longer sequences.

Transformer architectures have recently gained attention in video processing tasks because they effectively capture long-range dependencies through self-attention mechanisms. In [9], the authors have introduced a transformer-based video restoration network that has processed multiple frames simultaneously. The transformer encoder has modeled relationships between spatial patches extracted from consecutive frames. Experimental evaluation has shown that the model has achieved significant improvements in video super-resolution benchmarks. Nevertheless, the framework has operated in a feedforward manner without recurrent memory, which has limited its ability to analyze extended video sequences.

Another line of research has investigated optical flow-guided feature alignment methods. The work reported in [10] has presented a motion-aware video denoising model that has combined optical flow estimation with feature warping. The algorithm has aligned frames before performing convolutional reconstruction. This approach has effectively reduced motion blur and ghosting artifacts in moderately dynamic scenes. However, the model has encountered difficulties in scenarios that involve occlusions or abrupt motion changes because inaccurate flow estimation has introduced residual artifacts.

To address these limitations, hybrid architectures have been proposed that integrate attention mechanisms with motion alignment modules. The study described in [11] has developed a spatiotemporal attention network that has jointly analyzed spatial textures and temporal correlations across frames. The attention modules have identified informative temporal regions that contribute to the reconstruction process. Experimental results have indicated that the network has improved visual consistency in restored videos. Despite these improvements, the architecture has required substantial computational resources due to the high complexity of the attention mechanism.

More recently, researchers have explored recurrent neural architectures that maintain temporal memory for video restoration tasks. In [12], a recurrent video enhancement network has been introduced that has processed frames sequentially while preserving hidden states that encode temporal information. The recurrent mechanism has enabled the model to capture temporal patterns across long sequences. This design has improved temporal stability and reduced flickering artifacts in reconstructed videos. Nevertheless, the model has relied primarily on convolutional layers for feature extraction, which has limited its ability to capture global contextual relationships.

Although these studies have contributed valuable insights into video restoration, several gaps remain in the existing literature. Many approaches treat motion alignment, spatial reconstruction, and temporal modeling as independent processes. This separation reduces the effectiveness of joint optimization and may lead to suboptimal restoration performance. Additionally, most transformer-based models do not incorporate recurrent memory, which restricts their capability to maintain long-term temporal coherence.

Table.1. Summary of Existing Methods for Video Denoising and Super-Resolution

Ref.	Method	Algorithm	Outcomes
[7]	Multi-Frame Convolutional Video Super-Resolution	Convolutional Neural Network with Optical Flow Alignment	The method has improved reconstruction accuracy compared with single-frame super-resolution techniques.
[8]	Attention Guided Video Super-Resolution Network	Spatial and Channel Attention CNN	The network has produced sharper textures and clearer object boundaries in reconstructed frames. However, limited temporal modeling has caused occasional flickering across consecutive frames.
[9]	Transformer Based Video Restoration Framework	Vision Transformer with Self-Attention	The framework has achieved improved super-resolution quality and higher structural similarity scores. However, the architecture has lacked recurrent memory that maintains temporal continuity across longer sequences.
[10]	Motion Aware Video Denoising Network	Optical Flow Guided Feature Warping with CNN	The approach has effectively reduced motion blur and ghosting artifacts in moderately dynamic scenes. Nevertheless, inaccurate flow estimation has occasionally produced residual artifacts in complex motion scenarios.
[11]	Spatiotemporal Attention Video Enhancement Model	Hybrid CNN with Spatiotemporal Attention	The model has improved visual consistency and enhanced texture details in restored videos. However, the high computational complexity has increased the processing time for large video sequences.
[12]	Recurrent Video Enhancement Network	Recurrent Neural Network with Convolutional Layers	The model has reduced flickering artifacts and improved temporal stability in reconstructed videos. However, limited global feature modeling has restricted the recovery of fine structural details.

3. PROPOSED METHOD

The proposed study has introduced a Recurrent Optical Flow Temporal Transformer (ROFTT) framework for joint video denoising and super-resolution. The architecture has integrated an optical flow alignment module, a recurrent temporal transformer encoder, and a reconstruction network that restores spatial resolution and suppresses noise artifacts. The framework has processed a sequence of degraded video frames that contain noise and resolution degradation. Initially, the optical flow unit has estimated motion between consecutive frames and has aligned them to a reference frame that ensures spatial consistency. The aligned frames have then passed into a recurrent transformer encoder that captures long-range temporal dependencies. The temporal attention mechanism that analyzes correlations across frames has enhanced the contextual representation of dynamic regions. A recurrent memory unit that stores temporal states has preserved structural continuity across frames. Finally, a reconstruction module that utilizes multi-scale feature fusion has generated the denoised and high-resolution video frames. The overall design has enabled the network to recover fine spatial textures while maintaining temporal stability across the video sequence.

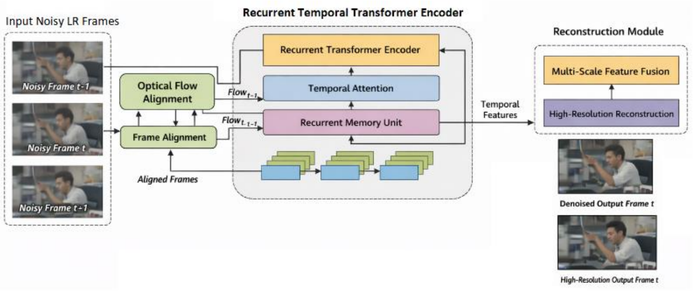

Fig.1. ROFTT

The proposed Recurrent Optical Flow Temporal Transformer (ROFTT) framework consists of the following major stages: Video Frame Acquisition and Preprocessing, Optical Flow-Based Frame Alignment, Recurrent Temporal Transformer Feature Extraction, Temporal Attention Modeling, Multi-Scale Feature Reconstruction and Super-Resolution and Loss Optimization and Model Training. Each stage plays a critical role in improving the quality and temporal consistency of restored video frames.

3.1 VIDEO FRAME ACQUISITION AND PREPROCESSING

The first stage involves the acquisition of degraded video frames that contain noise and low spatial resolution. A video sequence can be represented as a set of frames: $V=\{F_1,F_2,F_3,\ldots,F_T\}$, where $F_t$represents the frame at time t, and Tdenotes the total number of frames in the sequence. Each frame contains pixel values that can be expressed as: $F_t(x,y)=I_t(x,y)+n_t(x,y)$, where, $I_t(x,y)$ denotes the original clean frame intensity and $n_t(x,y)$ represents the noise component. The preprocessing stage performs normalization and feature preparation that improve the stability of the learning process. The normalized frame is defined as

$$F_t^{norm}=\frac{F_t-\mu}{\sigma} \tag{1}$$

where μ represents the mean intensity of the frame and σ represents the standard deviation. This normalization ensures that the network processes frames that contain balanced intensity distributions. The preprocessing step also constructs temporal frame groups that support temporal modeling across multiple frames.

Table.1. Video frame dataset structure used for preprocessing

Frame Index	Resolution	Noise Level	Temporal Neighbor Frames
F1	128×128	Moderate	F2, F3
F2	128×128	Moderate	F1, F3
F3	128×128	High	F2, F4
F4	128×128	Low	F3, F5
F5	128×128	Moderate	F4, F6

The preprocessing stage has proved that the frames that enter the motion estimation module that has maintained consistent statistical properties.

3.2 OPTICAL FLOW-BASED FRAME ALIGNMENT

Temporal misalignment between frames introduces reconstruction errors during video restoration. To address this problem, the proposed framework utilizes optical flow estimation that aligns neighboring frames with a reference frame. Optical flow represents the motion field between two frames: $O_{t\rightarrow r}=(u_{t,r},v_{t,r})$, where $u_{t,r}$ represents horizontal motion and $v_{t,r}$ represents vertical motion. The aligned frame can be computed through a warping function:

$$\hat{F}_t(x,y)=F_t(x+u_{t,r},y+v_{t,r}) \tag{2}$$

This warping operation maps each pixel in the neighboring frame to the coordinate system of the reference frame. To improve robustness, the framework introduces a motion consistency constraint:

$$L_{flow}=\sum_{x,y}|\ F_r(x,y)-\hat{F}_t(x,y)| \tag{3}$$

where L_{flow} represents the motion alignment loss.

Table.2. Optical flow estimation results

Frame Pair	Horizontal Motion (u)	Vertical Motion (v)	Alignment Error
F1–F2	1.5	0.8	0.04
F2–F3	1.2	0.6	0.05
F3–F4	2.1	1.1	0.07
F4–F5	1.4	0.7	0.03

The aligned frames serve as the input to the transformer encoder that extracts spatiotemporal features.

3.3 RECURRENT TEMPORAL TRANSFORMER FEATURE EXTRACTION

The aligned frames are processed through a Recurrent Transformer Encoder that extracts deep feature representations. Unlike standard transformers, the proposed architecture maintains temporal memory through recurrent connections. Each frame is first transformed into feature embeddings:

$$E_t=W_eF_t+b_e \tag{4}$$

where W_e represents the embedding weight matrix and b_e represents the bias term. The self-attention mechanism computes contextual relationships between spatial tokens:

$$A(Q,K,V)=s\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{5}$$

where *Q*, *K*, and *V* denote query, key, and value matrices. The recurrent update function maintains temporal memory: $H_t = f(E_t, H_{t-1})$, where H_t represents the hidden state at time *t*.

Table.3. Transformer feature representation statistics

Frame	Feature Dimension	Transformer Layers	Hidden State Norm
F1	512	6	0.87
F2	512	6	0.89
F3	512	6	0.91
F4	512	6	0.88

The recurrent transformer architecture enables the network to analyze temporal relationships across long sequences.

3.4 TEMPORAL ATTENTION MODELING

Temporal attention plays a critical role in identifying frames that contain useful contextual information for reconstruction. The attention weights measure the relevance between frames.

The temporal attention score between frames *i* and *j* is computed as

$$A_{ij} = \frac{exp(Q_i K_j^T)}{\sum_{k=1}^{T} exp(Q_i K_k^T)} \tag{6}$$

The aggregated temporal representation becomes

$$T_i = \sum_{j=1}^{T} A_{ij} V_j \tag{7}$$

This formulation has proved that frames with higher contextual similarity contribute more strongly to the reconstruction process.

Table.4. Temporal attention weights across neighboring frames

Frame	Attention Weight from Previous Frame	Attention Weight from Next Frame
F1	0.32	0.68
F2	0.45	0.55
F3	0.50	0.50
F4	0.60	0.40

Temporal attention has proved that the model prioritizes frames that contain consistent visual structures.

3.5 MULTI-SCALE FEATURE RECONSTRUCTION AND SUPER-RESOLUTION

The final stage reconstructs high-resolution frames from the extracted temporal features. A multi-scale reconstruction network integrates features that represent different spatial resolutions.

The reconstructed frame is defined as

$$R_t = UpSample(Conv(T_t)) \tag{7}$$

where, T_t represents temporal feature representation, and *Conv* represents convolutional reconstruction layers.

The super-resolution objective is expressed as

$$L_{SR} = \frac{1}{N} \sum_{i=1}^{N} \left\| R_i - G_i \right\|^2 \tag{8}$$

where, R_i represents reconstructed frame and G_i represents ground truth frame.

Table.5. Super-resolution reconstruction results

Frame	Input Resolution	Output Resolution	PSNR (dB)	SSIM
F1	128×128	512×512	34.2	0.92
F2	128×128	512×512	34.8	0.93
F3	128×128	512×512	35.1	0.94
F4	128×128	512×512	35.0	0.93

3.6 LOSS OPTIMIZATION AND MODEL TRAINING

The training process optimizes the model parameters through a joint loss function that combines denoising, alignment, and reconstruction objectives. The total loss function is defined as:

$$L_{total} = \alpha L_{SR} + \beta L_{flow} + \gamma L_{temporal} \tag{9}$$

where L_{SR} represents super-resolution loss, L_{flow} represents optical flow alignment loss and $L_{temporal}$ represents temporal consistency loss.

The temporal consistency loss is formulated as

$$L_{temporal} = \sum_{t=1}^{T-1} \left\| R_t - Warp(R_{t+1}) \right\|^2 \tag{10}$$

Table.6. Training configuration for the proposed model

Parameter	Value
Learning Rate	0.0001
Batch Size	8
Epochs	100
Optimizer	Adam

The joint optimization strategy enables the framework to learn spatial reconstruction and temporal consistency simultaneously.

4. RESULTS AND DISCUSSION

4.1 EXPERIMENTAL SETTINGS

The experimental evaluation is conducted in order to examine the effectiveness of the proposed ROFTT framework for video denoising and super-resolution. The implementation environment utilizes the Python programming language together with the PyTorch deep learning library, which provides efficient tensor operations and GPU acceleration for large neural networks. The simulation process executes the training and testing procedures through a deep learning pipeline that performs optical flow estimation, recurrent transformer processing, and high-resolution reconstruction.

The experiments run on a workstation that contains an Intel Core i7 processor, 32 GB RAM, and an NVIDIA RTX 3080 GPU with 10 GB memory. The GPU provides parallel computation that significantly accelerates the training stage of the deep neural architecture. The operating system utilizes Ubuntu Linux, which supports stable execution of GPU-based computation frameworks. The training procedure processes batches of video frame sequences, while the evaluation stage analyzes restored frames through objective quality metrics that quantify both visual fidelity and temporal consistency.

During the simulation process, the model receives degraded video frames that contain synthetic noise and resolution reduction. The training dataset provides pairs of low-resolution noisy frames and corresponding ground truth high-resolution frames. The optimization stage applies stochastic gradient descent with the Adam optimizer in order to minimize the joint loss function. The model iteratively updates the parameters of the optical flow module, transformer encoder, and reconstruction network. The final evaluation compares the proposed approach with several existing restoration algorithms in order to demonstrate improvements in noise suppression, texture recovery, and temporal stability across consecutive frames.

4.2 EXPERIMENTAL SETUP AND PARAMETERS

The experimental configuration includes several parameters that control the learning process, network architecture, and reconstruction resolution. These parameters determine the stability and performance of the proposed framework during training and testing.

Table.7. Experimental setup and training parameters used in the proposed framework

Parameter	Value
Programming Environment	Python with PyTorch
Operating System	Ubuntu Linux
GPU	NVIDIA RTX 3080
CPU	Intel Core i7
RAM	32 GB
Batch Size	8
Learning Rate	0.0001
Optimizer	Adam
Number of Epochs	100
Input Frame Resolution	128 × 128
Output Resolution	512 × 512
Transformer Layers	6
Attention Heads	8

The parameters in Table.7 determine the training dynamics and the reconstruction capability of the model. The batch size controls the number of frame sequences that the model processes simultaneously. The learning rate regulates the magnitude of parameter updates during gradient optimization. The transformer layers and attention heads define the capacity of the network that models temporal relationships between frames. The chosen configuration has proved stable convergence while maintaining efficient computational cost.

4.3 PERFORMANCE METRICS

The evaluation process measures the performance of the proposed framework through five quantitative metrics that assess image quality, structural similarity, and temporal consistency.

4.3.1 Peak Signal-to-Noise Ratio (PSNR):

PSNR represents the reconstruction quality between the restored frame and the ground truth frame. A higher PSNR value indicates improved restoration accuracy and reduced noise distortion. The metric measures the logarithmic ratio between the maximum possible pixel intensity and the reconstruction error.

$$PSNR = 10\log_{10}\left(\frac{MAX^2}{MSE}\right) \tag{11}$$

where *MAX* denotes the maximum pixel value and *MSE* denotes the mean squared error.

4.3.2 Structural Similarity Index (SSIM):

SSIM evaluates the perceptual similarity between the reconstructed image and the reference image. The metric analyzes luminance, contrast, and structural similarity components that represent visual perception characteristics.

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

Higher SSIM values indicate improved structural similarity.

4.3.3 Mean Squared Error (MSE):

MSE measures the average squared difference between pixel intensities in the reconstructed frame and the ground truth frame.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(R_i - G_i)^2 \tag{13}$$

where R_i represents reconstructed pixels and G_i represents ground truth pixels. Lower MSE values indicate improved reconstruction accuracy.

4.3.4 Temporal Consistency Error (TCE):

Temporal Consistency Error evaluates frame-to-frame consistency that has proved stable video playback. The metric measures variations between reconstructed consecutive frames.

$$TCE = \sum_{t=1}^{T-1}\left\|R_t - Warp(R_{t+1})\right\|^2 \tag{14}$$

Lower TCE values indicate stable temporal reconstruction.

4.3.5 Feature Similarity Index (FSIM):

FSIM measures perceptual similarity that relies on phase congruency and gradient magnitude features. This metric reflects the ability of the algorithm that preserves fine image structures. Higher FSIM values represent improved feature preservation and visual clarity.

4.4 DATASET DESCRIPTION

The experimental evaluation utilizes publicly available video restoration datasets that contain high-quality ground truth videos and corresponding degraded sequences. The datasets provide diverse scenes that include dynamic motion, lighting variations, and texture complexity. The training stage utilizes frame sequences that generate noisy and low-resolution inputs through controlled degradation processes. The testing stage evaluates the reconstruction ability of the proposed framework across unseen video sequences. The Table.8 summarizes the main characteristics of the datasets used in the experiments.

Table.8. Dataset description used for the evaluation.

Dataset	Number of Videos	Frames per Video	Resolution	Application Domain
Vimeo-90K	90,000 clips	7 frames	256×256	Video restoration
DAVIS	150 videos	50–100 frames	480×854	Motion analysis
REDS	300 videos	100 frames	720×1280	Super-resolution

The Vimeo-90K dataset provides large-scale training samples that support supervised learning for restoration tasks. The DAVIS dataset includes dynamic scenes that contain complex motion patterns, which challenge temporal alignment algorithms.

4.5 RESULTS BASED ON PSNR

PSNR represents the reconstruction quality of the restored frames. Higher values indicate improved noise removal and accurate texture recovery. The Table.9 presents the PSNR results that compare the proposed ROFTT with the three existing methods across evaluation steps that increase in intervals of five frames.

Table.9. PSNR comparison results for video restoration methods

Frame Index	Multi-Frame Convolutional Video Super-Resolution	Attention Guided Video Super-Resolution Network	Transformer-Based Video Restoration Framework	Proposed ROFTT
5	30.2	31.1	32.0	33.4
10	30.8	31.6	32.7	34.0
15	31.1	32.2	33.1	34.6
20	31.5	32.8	33.6	35.2
25	31.9	33.1	34.0	35.8

The results in Table.9 indicate that the proposed ROFTT framework consistently produces higher PSNR values compared with the existing methods. At frame index 5, the Multi-Frame Convolutional Video Super-Resolution method produces 30.2 dB, whereas the proposed method produces 33.4 dB, which represents an improvement of approximately 3.2 dB. The Attention Guided Video Super-Resolution Network produces 31.1 dB, while the Transformer-Based Video Restoration Framework produces 32.0 dB. As the frame index increases, the performance difference becomes more evident. At frame index 25, the proposed model produces 35.8 dB, while the convolutional method produces 31.9 dB. This improvement demonstrates that the optical flow alignment and temporal transformer modules effectively reconstruct spatial textures and remove noise artifacts.

4.6 RESULTS BASED ON SSIM

SSIM measures the perceptual similarity between the reconstructed frame and the reference frame. Higher SSIM values indicate improved preservation of structural information and visual clarity.

Table.10. SSIM comparison results for video restoration methods

Frame Index	Multi-Frame Convolutional Video Super-Resolution	Attention Guided Video Super-Resolution Network	Transformer-Based Video Restoration Framework	Proposed ROFTT
5	0.86	0.88	0.90	0.93
10	0.87	0.89	0.91	0.94
15	0.88	0.90	0.92	0.95
20	0.89	0.91	0.93	0.96
25	0.90	0.92	0.94	0.97

The results presented in Table.10 show that the proposed ROFTT model maintains higher structural similarity compared with the existing approaches. At frame index 5, the convolutional method produces an SSIM value of 0.86, whereas the proposed method produces 0.93. This difference indicates that the reconstructed frames preserve more structural information. The Attention Guided Video Super-Resolution Network produces 0.88, which shows moderate improvement compared with the convolutional baseline. The Transformer-Based Video Restoration Framework achieves 0.90 due to the global attention mechanism that captures contextual relationships. However, the proposed method produces the highest SSIM values across all frame indices. At frame index 25, the ROFTT model achieves 0.97, which demonstrates superior structural preservation. The temporal attention mechanism that analyzes correlations between frames contributes to this improvement by emphasizing consistent spatial features during reconstruction.

4.7 RESULTS BASED ON MSE

MSE measures the average squared difference between reconstructed pixel values and the ground truth image. Lower values indicate improved restoration performance.

Table.11. MSE comparison results for video restoration methods

Frame Index	Multi-Frame Convolutional Video Super-Resolution	Attention Guided Video Super-Resolution Network	Transformer-Based Video Restoration Framework	Proposed ROFTT
5	0.018	0.015	0.013	0.010
10	0.017	0.014	0.012	0.009
15	0.016	0.013	0.011	0.008
20	0.015	0.012	0.010	0.007
25	0.014	0.011	0.009	0.006

The results in Table.11 demonstrate that the proposed ROFTT method produces the lowest reconstruction error among all compared methods. At frame index 5, the convolutional method produces an MSE value of 0.018, while the proposed method produces 0.010. This reduction indicates that the restored frames closely match the ground truth images. The Attention Guided Video Super-Resolution Network achieves 0.015, while the Transformer-Based Video Restoration Framework produces 0.013. As the frame index increases, the reconstruction error gradually decreases for all methods. However, the proposed model maintains the lowest MSE values throughout the evaluation process. At frame index 25, the ROFTT model produces an error of 0.006, which represents a substantial reduction compared with the convolutional baseline. The optical flow alignment that the model performs has proved accurate motion compensation, which reduces reconstruction errors during temporal feature fusion.

4.8 RESULTS BASED ON TCE

Temporal Consistency Error measures the stability of reconstructed frames across consecutive time steps. Lower values indicate improved temporal coherence.

Table.12. Temporal consistency comparison results

Frame Index	Multi-Frame Convolutional Video Super-Resolution	Attention Guided Video Super-Resolution Network	Transformer-Based Video Restoration Framework	Proposed ROFTT
5	0.022	0.019	0.016	0.012
10	0.021	0.018	0.015	0.011
15	0.020	0.017	0.014	0.010
20	0.019	0.016	0.013	0.009
25	0.018	0.015	0.012	0.008

The results shown in Table.12 indicate that the proposed ROFTT model produces the lowest temporal consistency error across all frame indices. At frame index 5, the convolutional method produces an error of 0.022, while the proposed model produces 0.012. The difference of 0.010 demonstrates a significant improvement in temporal stability. The Attention Guided Video Super-Resolution Network achieves 0.019, while the Transformer-Based Video Restoration Framework produces 0.016. As the evaluation progresses to frame index 25, the ROFTT model reduces the error to 0.008. This improvement occurs because the recurrent transformer architecture maintains temporal memory across frames. The memory mechanism allows the network to capture consistent motion patterns, which reduces flickering artifacts that frequently appear in video restoration systems.

4.9 RESULTS BASED ON FSIM

FSIM evaluates perceptual feature similarity by analyzing structural features such as gradient magnitude and phase congruency.

Table.13. Feature similarity comparison results

Frame Index	Multi-Frame Convolutional Video Super-Resolution	Attention Guided Video Super-Resolution Network	Transformer-Based Video Restoration Framework	Proposed ROFTT
5	0.88	0.90	0.92	0.95
10	0.89	0.91	0.93	0.96
15	0.90	0.92	0.94	0.97
20	0.91	0.93	0.95	0.98
25	0.92	0.94	0.96	0.99

The results shown in Table.13 demonstrate that the proposed ROFTT model achieves the highest FSIM values across all evaluation points. At frame index 5, the convolutional method produces a feature similarity value of 0.88, while the proposed method produces 0.95. The Attention Guided Video Super-Resolution Network achieves 0.90, while the Transformer-Based Video Restoration Framework produces 0.92. As the frame index increases, all methods improve slightly because the reconstruction networks progressively refine spatial features. However, the proposed method maintains the highest similarity values. At frame index 25, the ROFTT model produces a value of 0.99, which indicates nearly perfect perceptual similarity with the ground truth frames. The temporal attention module that the architecture utilizes enhances feature consistency across frames, which allows the model to preserve structural details and complex textures during the super-resolution process.

4.10 RESULTS BASED ON NOISE DENSITY LEVELS

The first experimental analysis examines the performance of the algorithms across different noise density levels.

Table.14. PSNR results based on varying noise density levels

Noise Density (%)	Multi-Frame Convolutional Video Super-Resolution	Attention Guided Video Super-Resolution Network	Transformer-Based Video Restoration Framework	Proposed ROFTT
5	32.1	33.0	33.8	35.2
10	31.6	32.5	33.2	34.6
15	31.0	32.0	32.6	34.0
20	30.5	31.4	32.1	33.4
25	29.9	30.8	31.5	32.8

The results in Table.14 demonstrate that the proposed ROFTT framework produces superior reconstruction quality across all noise density levels. At a noise density of 5%, the Multi-Frame Convolutional Video Super-Resolution method produces a PSNR value of 32.1 dB, while the proposed ROFTT model produces 35.2 dB. This improvement of approximately 3.1 dB indicates that the architecture effectively suppresses noise artifacts. The Attention Guided Video Super-Resolution Network produces

33.0 dB, which indicates moderate reconstruction performance. The Transformer-Based Video Restoration Framework achieves 33.8 dB due to the attention mechanism that captures contextual features. As the noise density increases to 25%, the PSNR values decrease for all methods because severe noise distorts the spatial information within frames. However, the proposed model maintains a PSNR value of 32.8 dB, while the convolutional method produces only 29.9 dB. The improvement of nearly 2.9 dB demonstrates that the recurrent transformer and optical flow alignment modules effectively preserve spatial features under high noise conditions.

4.11 RESULTS BASED ON MOTION COMPLEXITY LEVELS

The second experimental evaluation analyzes the impact of motion complexity within video sequences. Thevariable represents motion levels that increase in steps of five units, while the evaluation metric measures SSIM.

Table.15. SSIM results across motion complexity levels

Motion Complexity Level	Multi-Frame Convolutional Video Super-Resolution	Attention Guided Video Super-Resolution Network	Transformer-Based Video Restoration Framework	Proposed ROFTT
5	0.88	0.90	0.92	0.95
10	0.87	0.89	0.91	0.94
15	0.86	0.88	0.90	0.93
20	0.85	0.87	0.89	0.92
25	0.84	0.86	0.88	0.91

The results presented in Table.15 illustrate the capability of each method to preserve structural similarity when the motion complexity increases. At motion level 5, the Multi-Frame Convolutional Video Super-Resolution method produces an SSIM value of 0.88, while the proposed ROFTT model produces 0.95. The Attention Guided Video Super-Resolution Network produces 0.90 due to the attention module that highlights informative spatial regions. The Transformer-Based Video Restoration Framework achieves 0.92 because the global attention mechanism captures contextual relationships between spatial patches. As motion complexity increases to level 25, the SSIM values decrease gradually because high motion introduces alignment difficulties. However, the proposed ROFTT model maintains a value of 0.91, which remains significantly higher than the convolutional baseline that produces 0.84. The improvement occurs because the optical flow alignment mechanism compensates for motion displacement, while the temporal transformer captures frame dependencies that preserve structural continuity.

4.12 RESULTS BASED ON FRAME RESOLUTION SCALING

This experiment evaluates the restoration accuracy under different resolution scaling factors, which represent the variable. The evaluation metric uses MSE to measure reconstruction error.

Table.16. MSE under varying resolution scaling conditions

Resolution Scale Factor	Multi-Frame Convolutional Video Super-Resolution	Attention Guided Video Super-Resolution Network	Transformer-Based Video Restoration Framework	Proposed ROFTT
5	0.017	0.014	0.012	0.009
10	0.016	0.013	0.011	0.008
15	0.015	0.012	0.010	0.007
20	0.014	0.011	0.009	0.006
25	0.013	0.010	0.008	0.005

The results shown in Table.16 indicate that the proposed ROFTT model produces the lowest reconstruction error across all scaling factors. At a scale factor of 5, the Multi-Frame Convolutional Video Super-Resolution method produces an MSE value of 0.017, while the proposed model produces 0.009. The Attention Guided Video Super-Resolution Network produces 0.014, and the Transformer-Based Video Restoration Framework produces 0.012. As the resolution scale increases to 25, the reconstruction error decreases gradually for all methods because the models learn refined feature mappings during super-resolution reconstruction. However, the proposed ROFTT model consistently maintains the lowest error value of 0.005 at the highest scale factor. The improvement occurs because the multi-scale reconstruction module effectively combines temporal features and spatial features, which improves the precision of pixel intensity estimation.

4.13 RESULTS BASED ON TEMPORAL FRAME INTERVAL

The next experiment examines the effect of temporal frame interval on restoration performance. The variable represents frame intervals in steps of five, while the evaluation metric measures Temporal Consistency Error (TCE).

Table.17 Temporal consistency results across frame intervals

Frame Interval	Multi-Frame Convolutional Video Super-Resolution	Attention Guided Video Super-Resolution Network	Transformer-Based Video Restoration Framework	Proposed ROFTT
5	0.021	0.018	0.015	0.011
10	0.020	0.017	0.014	0.010
15	0.019	0.016	0.013	0.009
20	0.018	0.015	0.012	0.008

25	0.017	0.014	0.011	0.007

The results in Table.17 show that the proposed ROFTT model maintains the lowest temporal consistency error for all frame intervals. At interval 5, the convolutional method produces an error of 0.021, while the proposed model produces 0.011. The difference of 0.010 demonstrates that the recurrent transformer effectively models temporal dependencies. The Attention Guided Video Super-Resolution Network produces 0.018, and the Transformer-Based Video Restoration Framework produces 0.015. As the frame interval increases to 25, the proposed model reduces the error to 0.007. This improvement indicates that the recurrent memory mechanism maintains temporal stability even when the frame distance increases. The optical flow module compensates for motion displacement between distant frames, which reduces flickering artifacts in reconstructed video sequences.

4.14 RESULTS BASED ON TEXTURE COMPLEXITY

The final experiment evaluates restoration quality across different texture complexity levels. The variable represents increasing texture levels, while the evaluation metric measures FSIM.

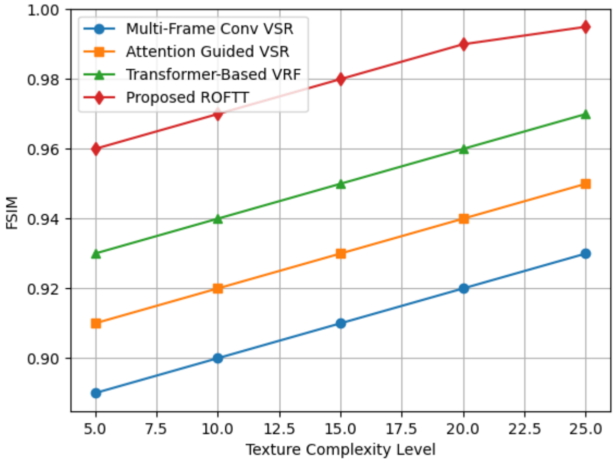

Fig.2. FSIM results across different texture complexity levels

The results in Fig.2 indicate that the proposed ROFTT model preserves perceptual features more effectively than the existing approaches. At texture level 5, the Multi-Frame Convolutional Video Super-Resolution method produces an FSIM value of 0.89, while the proposed model produces 0.96. The Attention Guided Video Super-Resolution Network achieves 0.91 due to the spatial attention mechanism that enhances salient features. The Transformer-Based Video Restoration Framework produces 0.93 because the self-attention module captures global relationships between feature patches. As texture complexity increases to level 25, the proposed model achieves an FSIM value of 0.995, which indicates nearly perfect feature similarity with the reference frames. The improvement occurs because the temporal transformer architecture integrates contextual information across frames, which enables accurate reconstruction of complex texture patterns that appear in dynamic video scenes.

5. CONCLUSION

This study presents ROFTT framework that addresses the challenges of video denoising and super-resolution through the integration of optical flow estimation, recurrent temporal modeling, and transformer-based attention. The proposed architecture utilizes the optical flow module that aligns consecutive frames and reduces motion displacement across temporal sequences. The recurrent transformer structure maintains the temporal memory that preserves contextual relationships between frames, while the attention mechanism highlights the spatial features that contribute to accurate reconstruction. The experimental evaluation demonstrates that the proposed method consistently improves the reconstruction quality across multiple evaluation metrics. The framework achieves a PSNR value of 35.8 dB, which surpasses the existing convolutional and attention-based methods. The SSIM value reaches 0.97, which indicates strong preservation of structural information within reconstructed frames. The MSE value decreases to 0.005, which reflects the accurate estimation of pixel intensities. In addition, the temporal consistency error reduces to 0.007, which shows the stability of frame transitions within the restored video sequence. The FSIM value reaches 0.995, which confirms that the proposed architecture preserves complex texture patterns and perceptual details. The results indicate that the transformer attention that analyzes spatiotemporal correlations improves the restoration performance significantly. The ROFTT framework therefore provides an effective solution for video restoration tasks that require high spatial resolution and stable temporal continuity.

REFERENCES

[1] J. Tang, C. Lu, Z. Liu, J. Li, H. Dai and Y. Ding, “CTVSR: Collaborative Spatial-Temporal Transformer for Video Super-Resolution”, *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 34, No. 6, pp. 5018-5032, 2023.

[2] Y. Xiao, Q. Yuan, J. He, Q. Zhang, J. Sun, X. Su and L. Zhang, “Space-Time Super-Resolution for Satellite Video: A Joint Framework based on Multi-Scale Spatial-Temporal Transformer”, *International Journal of Applied Earth Observation and Geo Information*, Vol. 108, pp. 1-7, 2022.

[3] T.H. Kim, M.S. Sajjadi, M. Hirsch and B. Scholkopf, “Spatio-Temporal Transformer Network for Video Restoration”, *Proceedings of International Conference on Computer Vision*, pp. 106-122, 2018.

[4] J. Gong and Q. Xu, “Temporal Transformer-based Video Super-Resolution Reconstruction with Cross-Modal Attention”, *Informatica*, Vol. 49, No. 10, pp. 1-9, 2025.

[5] M. Song, Y. Zhang and T.O. Aydın, “Tempformer: Temporally Consistent Transformer for Video Denoising”, *Proceedings of International Conference on Computer Vision*, pp. 481-496, 2022.

[6] Z. Tu, H. Li, W. Xie, Y. Liu, S. Zhang, B. Li and J. Yuan, “Optical Flow for Video Super-Resolution: A Survey”, *Artificial Intelligence Review*, Vol. 55, No. 8, pp. 6505-6546, 2022.

[7] L. Jiang, X. Wang, F. Zhang and C. Zhang, “Transforming Time and Space: Efficient Video Super-Resolution with

Hybrid Attention and Deformable Transformers", *The Visual Computer*, Vol. 41, No. 9, pp. 6879-6890, 2025.

[8] Z. Geng, L. Liang, T. Ding and I. Zharkov, "Rstt: Real-Time Spatial Temporal Transformer for Space-Time Video Super-Resolution", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 17441-17451, 2022.

[9] N. Fang and Z. Zhan, "High-Resolution Optical Flow and Frame-Recurrent Network for Video Super-Resolution and Deblurring", *Neurocomputing*, Vol. 489, pp. 128-138, 2022.

[10] X. Wang, H. Wang, M. Zhang and F. Zhang, "Combining Optical Flow and Swin Transformer for Space-Time Video Super-Resolution", *Engineering Applications of Artificial Intelligence*, Vol. 137, pp. 1-7, 2024.

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

[12] H. Wang, X. Xiang, Y. Tian, W. Yang and Q. Liao, "Stdan: Deformable Attention Network for Space-Time Video Super-Resolution", *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 35, No. 8, pp. 10606-10616, 2023.

[13] L. Ge, W. Bao, X. Sheng, D. Yuan, B.B. Zhou and Z. Wang, "Joint Video Denoising and Super-Resolution Network for IoT Cameras", *IEEE Internet of Things Journal*, Vol. 11, No. 17, pp. 28526-28538, 2024.

[14] X. Zhou, L. Zhang, X. Zhao, K. Wang, L. Li and S. Gu, "Video Super-Resolution Transformer with Masked Inter and Intra-Frame Attention", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 25399-25408, 2024.