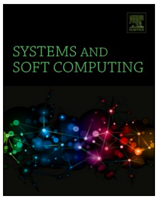
Contents lists available at ScienceDirect

Systems and Soft Computing

journal homepage: www.journals.elsevier.com/soft-computing-letters



High-fidelity video frame interpolation through context-aware temporal aggregation and recurrent propagation

Mohana Priya P *, Ulagapriya K

Department of Computer Science & Engineering, Vels Institute of Science, Technology and Advanced Studies (VISTAS), Chennai, India

A R T I C L E I N F O

Keywords:
STM
MSU-VFI dataset
YouTube-8 M
PSNR
SSIM

A B S T R A C T

Accurate inpainting of missing middle frames in video sequences is vital for multiple applications like video restoration, enhancement and compression. This study introduces a sophisticated deep learning-based framework designed to address this challenge by utilizing adjoining sequences of preceding and following frames. Our approach integrates temporal aggregation and recurrent propagation to effectively perform frame inpainting. Temporal aggregation leverages visible content from adjacent frames to recreate missing frames, ensuring high spatial fidelity and feature conservation. Optical flow estimation, utilizing methods such as Farneback Optical Flow, estimates displacement between frames and provides motion vectors that guide the interpolation process, enabling accurate alignment and blending of frames. Recurrent propagation is accomplished through Long Short-Term Memory (LSTM) networks that maintains temporal coherence by embedding and propagating information from preceding frames, thus ensuring smooth transitions and consistency across the video sequence. To further enhance performance, our model includes a context-aware feature extraction mechanism that adapts to various motion patterns and occlusions, optimizing the reconstruction quality. Framework has been evaluated on MSU Video Frame Interpolation (VFI) Benchmark Dataset, which provides diverse and challenging scenarios for interpolation, as well as the YouTube-8 M dataset, which contains a wide range of real-world video content. The experimental results demonstrate the robustness of the proposed model: a PSNR of 32.00 and an SSIM score of 0.905 indicate its superior reconstruction quality and structural similarity compared to baseline models. These results underscore the framework's effectiveness in handling complex motion dynamics and occlusions, making it well suited for advanced video restoration, enhancement and compression tasks.

1. Introduction

Video frame interpolation (VFI) involves generation of transitional frames between two successive frames which is essential for diverse spectrum of applications like slow motion effect creation, smoother playbacks, video restoration, enhancement and compression [1]. By means of precise prediction of missing frames, interpolation helps in improving the quality, enabling conversion of videos into higher frame rates as well as smoothening of playbacks [2,3]. It can also reduce visual artifacts in low-frame-rate videos [4]. High-fidelity frame interpolation is predominantly significant in areas like gaming, entertainment, medical imaging, virtual reality (VR), content creation, surveillance and video upscaling [5].

1.1. Challenges in traditional approaches

Existing techniques to video frame interpolation comprises of block-based motion compensation, optical flow methods, frame averaging, linear interpolation and parametric models. Block-based approaches like Block Matching Algorithm (BMA), Diamond Search (DS), Three Step Search Algorithm (TSSA) etc., divide frames into small blocks and calculates motion vectors to predict intermediate frames which often suffers from temporal incoherence, block artifacts and limited motion representation [6]. Optical flow-based methods like Farneback Optical Flow, Lucas-Kanade and Horn-Schunck estimates motion between frames and uses this information for interpolation can be effective in controlled settings, struggle with high computational cost, inaccuracy in regions of occlusion and rapid motion [7]. Frame averaging techniques like simple, weighted, moving and temporal smoothing methods blend multiple adjacent frames to estimate intermediate frames, but can lead

* Corresponding author.
E-mail addresses: mohana.vistas@gmail.com, pmohana.se@vistas.ac.in (M.P. P), upriya.se@vistas.ac.in (U. K).

https://doi.org/10.1016/j.sasc.2025.200428
Received 19 July 2025; Received in revised form 19 November 2025; Accepted 7 December 2025
Available online 8 December 2025

to blurring and may fail to handle large complex motions effectively [8]. Linear interpolation algorithm blends pixel values linearly between adjacent frames can also result in blurry and less accurate frames, especially in the presence of complex motion [9]. These parametric models use predefined motion models to predict intermediate frames but may not accurately capture non-linear motions and can result inconsistency, moreover they can also be excessively specific to certain kinds of motion. Early CNN-based methods, while innovative, often relied on limited temporal context and manual feature extraction which constrained its flexibility and accuracy [10].

1.2. Proposed solution

Traditional methods have laid the groundwork for video interpolation but each has its own limitations including issues with motion accuracy, computational demands, inability to preserve fine details and temporal incoherence. Modern techniques, particularly those leveraging deep learning, aim to address these limitations and enhance the quality and efficiency of video interpolation. Modern approaches leverage deep learning-based methods that enhance interpolation quality by learning the subtle patterns of motion and occlusion from large datasets. For instance, optical flow with deep learning enables more accurate motion estimation. Flow-guided video synthesis methodologies have gained popularity for their ability to handle complex motion dynamics by predicting both forward as well as backward optical flow to synthesize intermediate frames. Another emerging trend is the incorporation of Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks excel at preserving temporal consistencies across video sequences. This is principally beneficial for videos where upholding the coherence over multiple frames is crucial. Furthermore, the rise of context-aware models that adapt to dissimilar motion patterns and occlusions further pushes the boundaries of video frame interpolation, thereby making models more versatile and reliable for real-world applications. Our study introduces a deep learning-based framework that combines context-aware temporal aggregation and recurrent propagation to accurately interpolate missing frames while maintaining spatial fidelity and temporal coherence. Our approach leverages optical flow estimation for precise motion guidance and utilizes LSTM networks to propagate temporal information, ensuring smooth transitions across frames.

Temporal aggregation leverages the visible content from adjacent frames, capturing high-resolution details to recreate missing frames even in the presence of complex motion or occlusions. By using Farneback Optical Flow, the displacement between frames is estimated, enabling precise alignment and blending throughout the video. Additionally, LSTM networks ensures coherence and smooth propagation of information from preceding frames. The proposed model is also context-aware, as it possesses the ability to adapt itself to various motion patterns and occlusions which further enhances and augments the interpolation quality.

1.3. Contributions

The principal contributions of our study can be summarized as follows:

- **Context-Aware Temporal Aggregation** uses visible content from adjacent frames to accurately recreate missing frames which ensures high spatial fidelity and feature conservation even in videos with complex motion patterns and occlusions.
- **Recurrent Propagation for Temporal Consistency** is accomplished through Long Short-Term Memory (LSTM) networks which maintains temporal coherence and enables smooth transitions and consistency even in challenging dynamic video sequences.
- **Optical Flow Integration** guides the interpolation process by providing accurate motion vectors which in turn ensures precise alignment and blending of frames, thereby enhancing the quality of the inpainting.
- Our model's high performance in handling sophisticated motion dynamics and occlusions makes it well-suited for cutting-edge video restoration, enhancement and compression tasks addressing crucial needs of multimedia processing applications.

The study begins with an Introduction in section 1 that sets the stage by providing background information on video interpolation along with problem being addressed, outlining the objectives and summarizing the major contributions of the study. Following this, Related works in section 2 offers a comprehensive overview of existing methods and techniques by investigating traditional and recent advancements, current trends and also identifies the gaps and challenges in the existing works. Proposed Methodology in section 3 details the framework proposed by the study, including its components and implementation specifications. Results and Discussion outlined in section 4 presents the datasets used, evaluation metrics employed and the process of training and testing the model followed by experimental findings, comparing them with previous models and real-world implications. Conclusion in section 5 summarizes the study's main findings and contributions along with suggestions for future research.

2. Related works

In this section, we review the most prominent and relevant works focusing on early optical flow-based methods, deep learning approaches and the integration of context-aware feature extraction techniques. We also highlight on certain hybrid methods that combine these techniques to achieve more robust and accurate video frame interpolation.

Jeong et al. [2024] proposed a framework that significantly enhances the training of optical flow models by addressing the limited ground-truth optical flow labels in existing datasets. Specifically, an occlusion-aware video frame interpolation method capable of generating high-quality interframes and corresponding optical flows has been proposed in the presence of large motions. This approach enables the automatic expansion of optical flow training data through a semi-supervised training strategy that leverages video frame interpolation [11]. Extensive experiments on standard optical flow benchmarks including Sintel and KITTI has been demonstrated to portray the effectiveness of proposed model.

Han et al. [2024] introduced a direct approach for estimating intermediate optical flow between adjacent frames through a motion aware VFI (Video Frame Interpolation] technique which effectively minimised the complexity and computational costs generally associated with traditional methods. By employing a cross-scale motion estimation model, the interaction between feature representations and flow maps has been enhanced during the N interpolation process [12]. This approach enables capturing nonlinear motion between consecutive frames through incorporation of directional loss to precisely guide the flow estimation and to improve the overall interpolation quality.

Joy et al. [2024] presented a network that extracts low-resolution images and identifies patches within them to predict high-resolution images. The proposed predicts the image for the given patch and can be repeated to generate the full frame in the High Efficiency Video Coding. Cleave approach is adopted to reduce computational complexity [13]. Zhou et al. [2023] statistically analysed the relationship between motion estimation accuracy and video interpolation quality in existing frame interpolation methods to propose a general motion distillation framework applicable to both flow-based and kernel-based video frame interpolation methods. The approach recommended utilises a teacher model that uses ground-truth target frame and adjacent frames to estimate motion. These estimates are then employed to guide the training of a student model for VFI [14].

Kalluri et al. [2023] leverages 3D spatio-temporal kernels to directly learn motion properties from un-labelled videos, significantly simplifying the training, testing and deployment processes for frame interpolation models [15]. Achieving substantial speedups along with consistent demonstration of superior qualitative and quantitative results has been accomplished on benchmarks like Vimeo-90 K and GoPro. Deng et al. [2023] came up with an optical flow estimation and interpolation network that are jointly optimized in an end-to-end manner to synthesize middle frames from its nearest two frames. Additionally, a dynamic memory mechanism is adopted to balance memory sparsity with diversity of normality representations. This mechanism effectively attenuates abnormal features during frame interpolation while preserving the characteristics of normal prototypes. The bi-directional interpolation design enhances normal frame synthesis and helps to block interpolation of abnormal frames, increasing the system's resilience to anomalies [16].

Wu et al. [2022] inspired by photo-realistic results achieved in VFI, proposed an optimization framework for video prediction. This methodology addresses the extrapolation issue using pretrained differentiable VFI module, eliminating the need for a dedicated training dataset . Moreover semantic or instance maps has not been utilised to enhance the applicability to any video content. KITTI, Middlebury and Vimeo90K datasets has been used to demonstrate the robustness of video prediction results across various scenarios [17].

Shi et al. [2022] highlighted the constraints that arise owing to reliance on deep convolutional neural networks by existing techniques, which are limited by content-agnostic kernel weights and restricted receptive fields. Transformer-based VFI has been recommended to enable content-aware aggregation and to capture long-range dependencies through self-attention mechanisms [18]. Moreover, to overcome the high computational cost, local attention has been extended into spatial-temporal domain. This multi-scale frame synthesis scheme is aimed at fully availing the capabilities of Transformers.

Shangguan et al. [2022] addresses the issue related to temporal video interpolation by means of Cross-Video Neural Representation. Neural Fields represents the neural representation of complex 3D scenes which is employed in the proposed model to make use of video as continuous function parameterized by a coordinate-based neural network. The inputs here are spatiotemporal coordinates and outputs are its corresponding RGB values. By conditioning the neural network on input frames, space-time consistency is accomplished in the ultimate interpolation process [19].

Siyao et al. [2021] came forward with an innovative framework to automatically interpolate in-between frames in animation data. The study is aimed at addressing the main challenges regarding textures and non-linear motions in animation cartoons which makes motion estimation difficult [20]. Segment guided matching module uses global pairing among coherent colours to tackle the lack of textures while recurrent flow refinement employs transformer structure for recurring predictions to handle large, non-linear motions.ATD-12 K, which is a large-scale animation triplet dataset as been used for comprehensive training and evaluation. Liu et al. [2020] makes use of quadratic video interpolation (QVI)technique to extract higher-order motion information to model interpolated flow. QVI can still have limitations while produce intermediate frames with ghosting, artifacts and inaccurate motion, especially during large and complex movements. Thus, a rectified quadratic flow prediction formulation using least squares has been adopted for accurate motion estimation. Residual contextual synthesis network leverages contextual information in high-dimensional feature space to handle complex scenes and motion patterns [21].

Peleg et al. [2019] advocated an interpolated motion neural network designed with an effective architecture and end-to-end training through multi-scale tailored losses. Motion estimation is viewed as a classification problem rather than regression to deal with resolution upscaling issues. Vimeo triplet dataset has been used and operating time in relatively lesser on single GPU for HD resolution [22]. Jiang et al. [2018] suggested making use of bi-directional optical flow between input images using U-Net architecture to approximate intermediate flows. However, this method tends to produce artifacts around motion boundaries. To address this, another U-Net is used to refine the approximated flow and predict soft visibility maps. The input images are then warped and fused to create each intermediate frame, with visibility maps applied to exclude occluded pixels, thereby reducing artifacts [23] (Table 1).

2.1. Research gaps

- In the field of video interpolation, major challenge is handling of complex motions, existing methods struggle to accurately interpolate frames in scenarios involving rapid or intricate movement, often leading to artifacts and motion distortions [24].
- Occlusion handling remains critical since portions of the scene may be hidden in adjacent frames, complicating the interpolation process.
- Temporal coherence requires further improvement, as many algorithms produce inconsistent motion across interpolated frames [25].
- Certain models achieve high accuracy on specific datasets, but may fail to generalize across diverse video domains, limiting real-world applicability. Computational efficiency poses another concern, as many advanced methods demand substantial processing power and memory, making them unsuitable for real-time deployment [26].
- Incorporation of broader contextual information from surrounding frames would enhance spatial fidelity and temporal smoothness . More comprehensive and human-aligned assessment techniques are required [27].
- Temporal feature modelling and data imbalance between motion-rich and static scenes remain underexplored, thereby creating opportunities for developing more adaptive and robust interpolation frameworks.

3. Proposed framework

3.1. Problem statement

We define the video interpolation problem as reconstructing a missing sequence of middle frames IM_V using the information from both former F_V and succeeding frames S_V.Let $V = \{v_1, v_2, \ldots, v_T\}$ be a sequence of frames from a real video where T represents the total number of frames. Let p, m and f be the number of preceding, middle and following frames respectively, so that $f + im + s = T$. We split the video frames into three segments:

Previous frames: $F_V = \{v_1, v_2, \ldots, v_f\}$
Intermediate frames: $IM_V = \{v_{p+1}, v_{p+2}, \ldots, v_{f+im}\}$
Ensuing frames: $S_V = \{v_{f+im+1}, \ldots, v_T\}$
The goal is to approximate an interpolation function f such that:

$$IM_V = f(F_V, S_V) \tag{1}$$

for all sequences V. We aim to minimize the error between predicted middle frames $IM_{\hat{V}}$ and the ground truth frames IM_V.

3.2. Model overview

The proposed methodology aims to address the limitation regarding generation of high-quality intermediate frames in video sequences by leveraging advanced deep learning techniques [Fig. 2]. The approach integrates temporal aggregation and recurrent propagation to enhance the accuracy and visual fidelity of frame interpolation through smooth transitions and accurate motion representation in video sequences. The recommended framework involves several key components whose working details has been presented elaborately in this section.

Table 1
Existing VFI studies related to our study.

Author	Technique	Inference	Limitations
Jeong et al. [2024]	Occlusion-aware video frame interpolation	Enhances optical flow training by generating high-quality interframes, expanding training data through semi-supervised strategies.	Limited ground-truth optical flow labels in existing datasets.
Han et al. [2024]	Motion-aware video frame interpolation	Intermediate optical flow is estimated with reduced complexity and computational costs are minimised using cross-scale motion estimation.	May not be able to address all challenges in traditional methods.
Joy et al. [2024]	Low-resolution image patch extraction	Predicts high-resolution images by identifying patches, reducing computational complexity.	Dependent on initial low-resolution input quality.
Zhou et al. [2023]	Motion distillation framework	Utilizes teacher-student model to estimate motion for video frame interpolation.	Requires strong model for effective learning.
Kalluri et al. [2023]	3D spatio-temporal kernels	Learns motion properties from unlabelled videos, simplifying training and deployment processes.	May not generalize well to all types of motion patterns.
Deng et al. [2023]	Joint optical flow estimation and interpolation network	Synthesizes middle frames with dynamic memory mechanism to balance normality representation.	Complex mechanism may increase training time.
Wu et al. [2022]	Optimization framework for video prediction	Addresses extrapolation using pretrained VFI module, improving general applicability to any video without dedicated training dataset.	Relies on the quality of the pretrained model for effective predictions.
Shi et al. [2022]	Transformer-based VFI	Enables content-aware aggregation and captures long-range dependencies.	High computational cost and complexity in implementation.
Shangguan et al. [2022]	Cross-Video Neural Representation	Utilizes neural fields to represent videos as continuous functions, enhancing space-time consistency in interpolation.	Complexity in training due to the neural representation structure.
Siyao et al. [2021]	Animation frame interpolation framework	Addresses texture and non-linear motion challenges in animation data with segment-guided matching and recurrent flow refinement.	Specific to animation data; may not generalize to other video types.
Liu et al. [2020]	Quadratic video interpolation (QVI)	Extracts higher-order motion information for flow modeling, addressing some interpolation quality issues.	Ghosting, artifacts and inaccuracies during large complex movements remain a concern.
Peleg et al. [2019]	Interpolated motion neural network	Utilizes an effective architecture with end-to-end training, viewing motion estimation as a classification problem for better resolution upscaling.	May struggle with very high-resolution datasets.
Jiang et al. [2018]	Bi-directional optical flow with U-Net	Improves flow approximation and reduces artifacts through refined flow estimation and visibility maps.	Artifacts around motion boundaries can occur if not adequately refined.

3.3. Data collection and pre-processing

YouTube-8 M Dataset has been utilised in our model which comprises approximately 8 million labelled video segments. For training purposes, a subset of around 100,000 videos are used to ensure diversity in motion patterns, occlusions and lighting conditions (Fig. 1).

This will cover a wide array of categories, enabling the model to learn generalizable features across different scenarios. Pre-processing involves frame normalization and *Re*-sizing operations where all frames extracted from the datasets are normalized and resized to a uniform resolution (256 × 256 pixels) to maintain consistency during training and testing. Data augmentation techniques such as random cropping, flipping and rotation are applied to the training set. This augmentation is done to increase the diversity of training data and helps to prevent overfitting [Fig. 2].

3.4. Optical flow estimation module

Farneback Optical Flow procedure is used to estimate the displacement between consecutive frames. The algorithm computes motion vectors by analysing the gradient information and polynomial approximation of the image patches. This results in dense motion field that represents the movement of each pixel from one frame to the other. This continuous representation of motion is crucial for accurately aligning and blending frames. It ensures accuracy by precise computation of small and large displacements in effective manner. Farneback Method can be denoted as

$$I_t(x,y) = I_{t+1}(x+u,y+v) \approx I_{t+1}(x,y) + \nabla I_{t+1}\cdot(u,v) \tag{2}$$

The equations for horizontal u and vertical v motion is defined as:

$$u(x,y) = \frac{\partial I}{\partial x},\ v(x,y) = \frac{\partial I}{\partial y} \tag{3}$$

where I is image intensity, (x,y) represents pixel coordinates *and* ∇I_{t+1} represents the gradient of I_{t+1} at $(x,\ y)$. Now optical flow between consecutive frames can be given as:

$$V_{flow}\ (v_t, v_{t+1}) = (u_t(x,y), v_t(x,y)) \tag{4}$$

This flow is used to guide the warping process.

Initial motion vectors obtained from Farneback Optical Flow are refined using filtering techniques to reduce noise and improve accuracy. Techniques such as motion vector smoothing and filtering are applied to enhance the quality of the estimated motion. Regularization function is applied to the raw motion vectors to reduce the noise. Smoothing Filter (Gaussian Filter) is applied which can be denoted as:

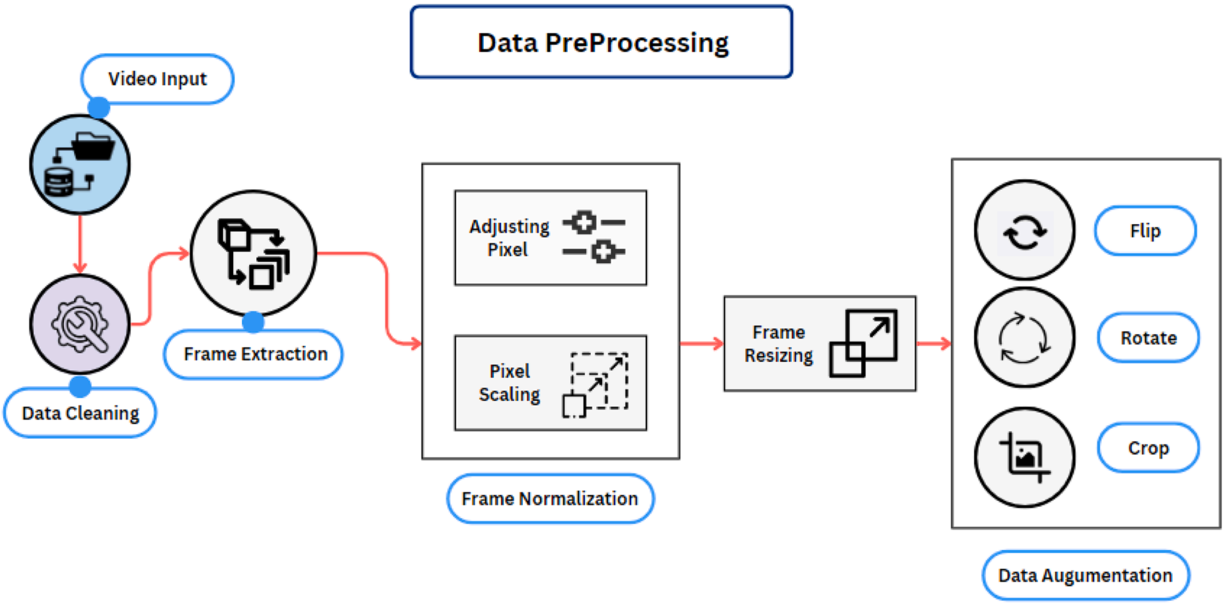


Fig. 1. Data collection and pre-processing.

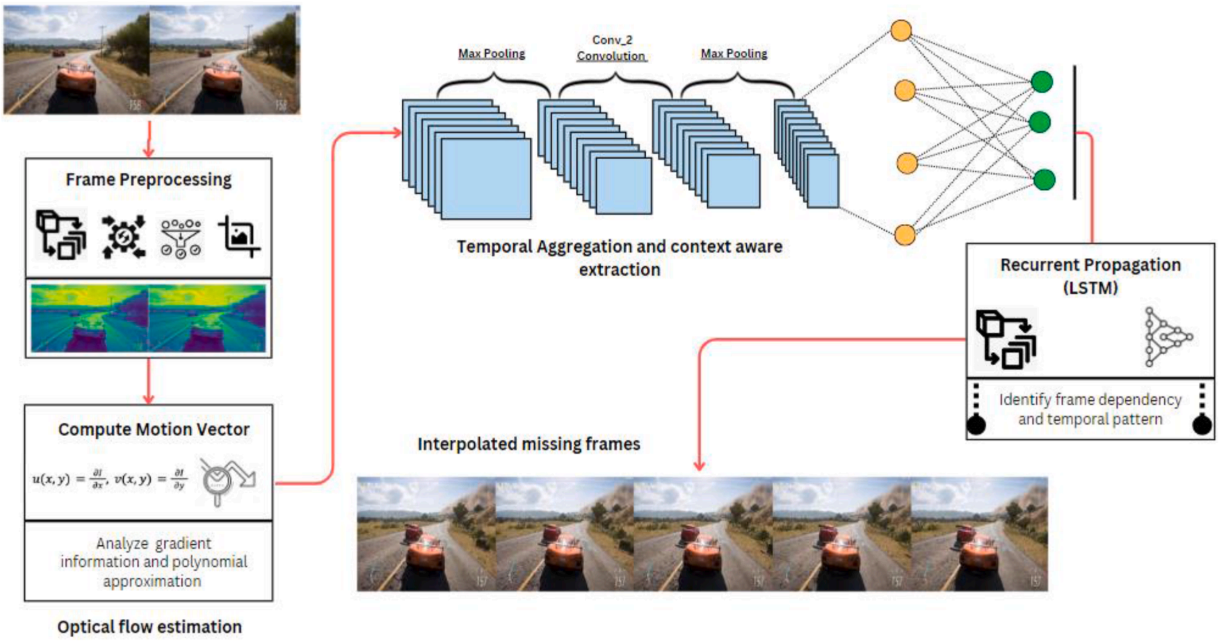

Frame Preprocessing
Compute Motion Vector
$u(x,y) = \frac{\partial I}{\partial x}$, $v(x,y) = \frac{\partial I}{\partial y}$
Analyze gradient information and polynomial approximation
Optical flow estimation
Max Pooling
Conv_2 Convolution
Max Pooling
Temporal Aggregation and context aware extraction
Recurrent Propagation (LSTM)
Identify frame dependency and temporal pattern
Interpolated missing frames

Fig. 2. Proposed Framework.

$$u^{\wedge}(x,y) = \frac{1}{\sigma^2\pi}\int_{-\infty}^{\infty} u(x',y')\exp\left(-\frac{(x-x')^2+(y-y')^2}{2\sigma^2}\right)dx'dy' \qquad (5)$$

In this equation, σ represents the standard deviation of Gaussian filter, controlling the degree of smoothing. This refinement step ensures motion vectors are not only accurate but also coherent across adjacent pixels, thereby improving the overall quality of the optical flow estimation. While numerous optical flow estimation methods exist, we deliberately selected the Farneback algorithm owing its computational efficiency and robust generalization capabilities. Although DL based alternatives like PWC—Net and RAFT achieve superior performance on standardized benchmarks, they introduce significant computational overhead and potential domain shift issues. Parametric-free nature of Farneback provides lightweight yet effective motion estimate that enables subsequent refinement modules to focus on handling occlusions and complex motion patterns rather than redundant flow estimation.

3.5. *Temporal aggregation component*

3.5.1. *a. frame synthesis using aggregated context*

Aggregate information from adjacent frames are utilised to reconstruct the missing intermediate frames. This involves combining the content and motion information from multiple frames to generate a high-fidelity interpolated frame [28]. A deep convolutional neural network (CNN) is employed to perform temporal aggregation. The network learns to integrate motion information and content from surrounding frames to generate seamless intermediate frame. The feature extraction process can be denoted as

$$\mathrm{F_t} = \mathrm{CNN}(\mathrm{I}_{t-1}, \mathrm{I}_{t+1}) = \sigma(\mathrm{W_1} * \mathrm{I}_{t-1} + \mathrm{W_2} * \mathrm{I}_{t+1} + b) \tag{6}$$

Where $F_t^{(l)}$is the feature map at layer l, $W^{(l)}$is the convolutional kernel, $b^{(1)}$ is the bias term, σ is the activation function (ReLU),* denotes the convolution operation, Pooling reduces dimensions and Dense is a fully connected layer. The combined equation can be denoted as

To reconstruct the missing intermediate frames, we aggregate information from adjacent frames (preceding and following) by combining both content and motion information which can be expressed as:

$$\mathrm{IM} = \mathrm{CNNIM}(F_{t-1}, F_{t+1}) = \sigma(\mathrm{WIM} * ((F_{t-1} \oplus F_{t+1}) + \mathrm{bIM}) \tag{7}$$

Where IM is the synthesized intermediate frame, ⊕ denotes feature concatenation, WIM and bIM are the weights and bias for interpolation CNN.

3.5.2. b. *context-aware feature extraction*

This mechanism adapts itself to various motion patterns and occlusions by extracting relevant features from both prior and subsequent frames, ensuring that interpolation process accounts for different types of motion and scene complexities [29,30].

$$\mathrm{C_t} = \mathrm{FA}(F_{t-1}, F_{t+1}) = \alpha \mathrm{Ft}_{-1} + (1-\alpha)\mathrm{Ft}_{+1} \tag{8}$$

Where C_t represents the aggregated features and α is a weight factor that determines the contribution of each frame.

$$\mathrm{C_t} = \mathrm{Attention}(F_{t-1}, F_{t+1})\cdot \mathrm{Ft} \tag{9}$$

Attention mechanism can be represented as

$$\mathrm{Attention}(F_{t-1}, F_{t+1})\cdot = \mathrm{softmax}\left(\frac{F_{t-1}, F_{t+1}}{\sqrt{d_k}}\right) \tag{10}$$

3.6. *LSTM based recurrent propagation module*

LSTM networks enable maintaining temporal coherence in video sequence interpolation by capturing long-term dependencies across time steps, which makes them suitable for ensuring smooth transitions between frames by propagating information from previous and subsequent frames. These networks are a type of recurrent neural network (RNN) and have more complex structure consisting of special components known as gates that control the flow of information within and out of the memory cells. The primary objective of this structure is to avoid vanishing gradients, that can prevent networks from learning long-term dependencies. Let $W \in \mathbb{R}^{4n\times\ (m+n)}$ and $b \in \mathbb{R}^{4n}$ act on the concatenated input $[h_{t-1}, x_t]$ then LSTM operation can be denoted as

$$(C_t, h_t) = (f_t \odot C_{t-1} + i_t \odot \widetilde{C}_t, o_t \odot \tanh\,(C_t)), \begin{bmatrix} f_t \\ i_t \\ o_t \\ \widetilde{C}_t \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{bmatrix} (W[h_{t-1}, x_t] + b) \tag{11}$$

State vector contains four gate sub-vectors each of length,⊙ denotes elementwise (Hadamard) product,W can be implemented as concatenation of the four gate weight matrices W_f, W_i, W_o, W_C and b as $[b_f, b_i, b_o, b_C]^\top$.

The combination of Optical Flow with LSTM provides a precise, pixel-level motion prior between frames, establishing initial correspondences for object trajectories. LSTM refines this motion by modelling complex temporal dynamics and leverages contextual memory to resolve ambiguities in occluded areas or disocclusions. This allows the network to correct flow errors, handle occlusions and synthesize plausible content where the flow alone is insufficient, leading to sharper and more temporally consistent frames. Thus, LSTM enhances the model's capacity to reconstruct missing frames by effectively managing temporal relationships and ensuring that inpainting process results in coherent and high-quality video sequences which is crucial for applications in video restoration, enhancement and compression as it leads to better-quality visuals and continuity. The flowchart of proposed system is provided in Fig. 3.

4. Results analysis and discussion

This section is dedicated to comprehensive evaluation of our model in terms of its effectiveness and performance by highlighting the robustness of our framework in accurately recreating missing frames by leveraging visible content from neighbouring frames and incorporating optical flow estimation for precise motion vector calculation. Temporal aggregation through LSTM networks maintains temporal coherence, ensuring smooth transitions across video sequences. The configuration for this deep learning-based video frame inpainting framework involves using a GPU-accelerated system, with RTX 3090 GPU and CUDA 11.2 for efficient model training and inference. The framework is built in TensorFlow 2.8 with cuDNN 8.1 support. OpenCV is employed for Farneback optical flow estimation which is configured to compute motion vectors between frames with parameters such as a pyramid scale of 0.5, three levels, a window size of 15, and three iterations per level. For temporal coherence, the framework uses LSTM networks to propagate information between frames, ensuring smooth transitions. Additionally, context-aware feature extraction adapts to the motion dynamics and occlusions in the scene. Simulation parameters are presented in Table 2.

The core of our framework is a sequence refinement module based on LSTM with 2-layers and 256 hidden units in each layer using a dropout rate of 0.3 between layers to prevent overfitting. LSTM processes concatenated features from warped frames and optical flow inputs, with its final hidden state being passed to a transposed convolutional decoder for final frame synthesis. For training, we use Adam optimizer with parameters β_1=0.9 and β_2=0.999, with an initial learning rate of 1×10^{-4}. The model is trained with batch size of 8 for 100 epochs using a composite loss function combining L1 reconstruction loss (λ=1.0), SSIM loss (λ=0.8), VGG perceptual loss (λ=0.1) and temporal warping loss (λ=0.5). We implement a step learning rate schedule, reducing the rate by half at epochs 50 and 75. The entire system was implemented in PyTorch and trained on NVIDIA V100 GPUs, with typical training convergence requiring approximately 48 h.

4.1. *Datasets*

MSU Video Frame Interpolation (VFI) Benchmark Dataset comprises a collection of video sequences designed for evaluating frame interpolation algorithms where each video has a resolution of 256 × 256 pixels featuring various motion dynamics like slow and fast movements. Our models are trained to predict up to five middle frames from five preceding and five following frames, allowing us to assess their performance on both interpolation tasks and generalization capabilities. YouTube-8 M dataset contains over 8 million video URLs, spanning a wide array of categories. For our experiments, we focus on a subset of approximately 1 million clips, all featuring resolutions of 320 × 240 pixels. This dataset is used for both training as well as testing since its valuable for evaluating our model on real-world scenarios that may include complex actions and occlusions (Table 3).

These datasets provide a comprehensive framework for evaluating our proposed method's ability to perform video frame inpainting and interpolation in both controlled and real-world scenarios. “p “is the number of preceding frames that are used as input in the model for training purposes. “m “refers to middle frames that are being in-painted in the sequence that need to be interpolated by the model. “f “refers to the number of following frames that the model uses along with the preceding frames to perform the interpolation.

The left side of the Fig. 4 shows the first frame extracted from the input dataset, while the right side illustrates the optical flow heatmap

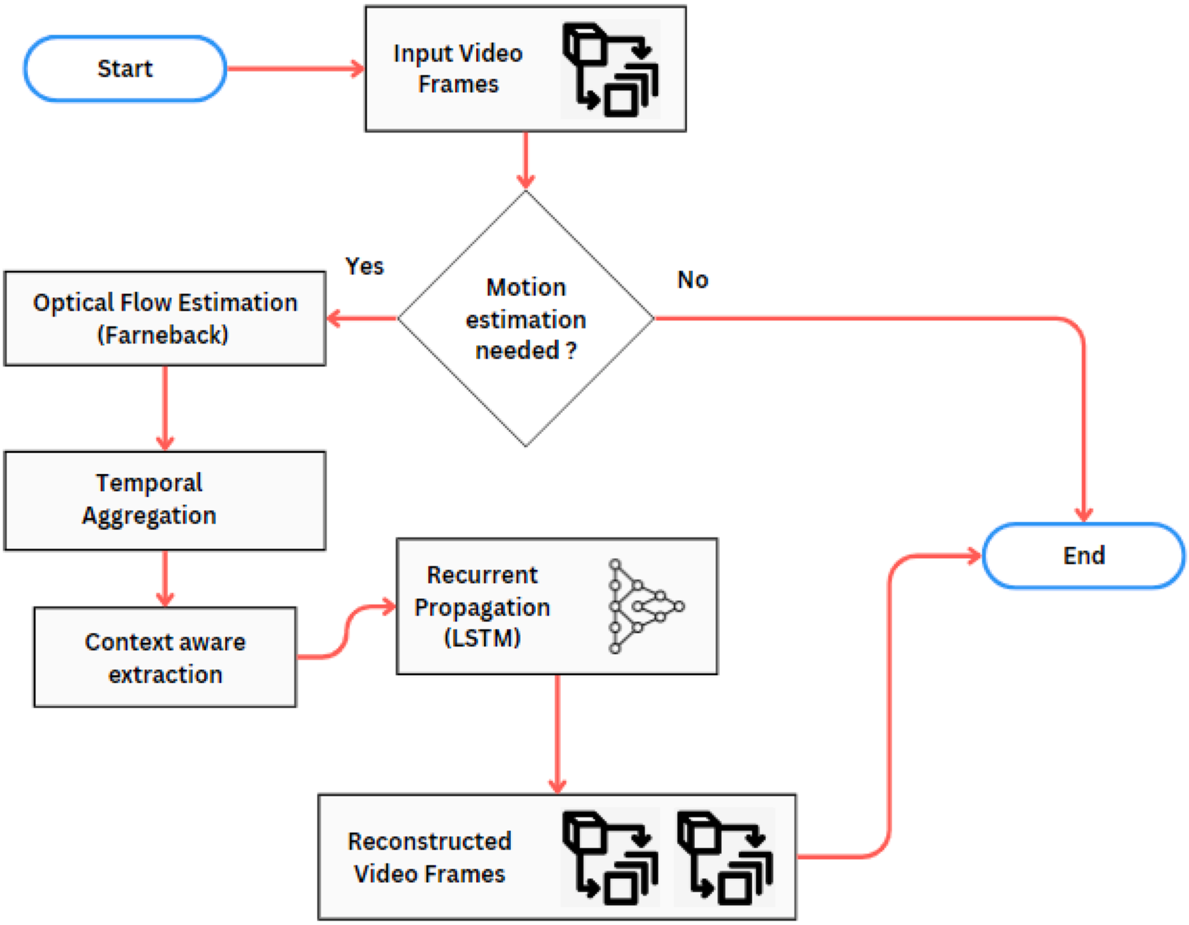

Start
Input Video Frames
Motion estimation needed ?
Yes
No
Optical Flow Estimation (Farneback)
Temporal Aggregation
Context aware extraction
Recurrent Propagation (LSTM)
Reconstructed Video Frames
End

Fig. 3. Proposed system workflow.

Table 2
System attributes.

Parameter	Value
Spatial fidelity weight(Farne back)	0.8
Pyramid scale	0.5
Levels	3
Winsize	15
Iterations	3
PolyN	5
PolySigma	1.2
Flags	0
Number of LSTM layers	2
Hidden units per layer	512
Dropout rate	0.3
Optimizer	Adam
Learning rate	1e-4
Epochs	50
Batch size	16
Adaptation mechanism	CNN
Filters per convolutional layer	64, 128, 256
Kernel size	3 × 3
Input frame size	128 × 128
Frame rate	30 fps

generated using Farneback Optical Flow. The heatmap provides a visual representation of motion intensity across the frame, where areas with high motion are highlighted in warmer colors. This motion estimation serves as a key input for frame interpolation and helps guide the prediction of intermediate frames by capturing both the direction and magnitude of motion in the scene.

Fig. 5 shows the five consecutive frames extracted from the input video for analysis. These frames serve as the basis for subsequent interpolation and motion estimation processes. Farneback Optical Flow is applied to consecutive frames to estimate motion between the frames [Fig. 6].

The heatmap overlay indicates the flow of pixel intensities, showing areas of high motion. This Fig. 7 visualizes the dense motion vectors obtained using Farneback Optical Flow, which describe the direction and magnitude of motion between frames. These vectors guide the interpolation process in the following steps. Fig. 8 presents the Inter mediate frames synthesized using a combination of context-aware feature extraction techniques which account for spatial and temporal features to improve the accuracy of frame prediction. Fig. 9 displays the five consecutive frames following the initial sequence which are used for temporal analysis and to assess the accuracy of frame interpolation techniques. These frames in conjunction with the prior frames, helps in evaluating the continuity and smoothness of predicted motion and content over time.

This Fig. 10 demonstrates the use of LSTM for interpolating motion between optical flow frames. The blue and green dashed lines represent the motion vector magnitudes of two optical flow frames at different time steps, while red solid line shows the LSTM-interpolated motion flow. Proposed model effectively captures the temporal dependencies and produces smooth transitions between the motion vectors of the two frames.

4.2. *Performance metrics*

4.2.1. *Peak signal -to -Noise ratio (PSNR)*

This is a quantitative measure of reconstruction quality of the video which compares the difference between original frames and predicted

Table 3
Training and Testing Sets.

Dataset	Source Clips	Resolution (Source)	Resolution (Train)	Resolution (Val/Test)	Color/ Grayscale	Max p (Train)	Max m (Train)	p (Val/ Test)	Small m (Val/Test)	Large m (Val/Test)
MSU VFI Benchmark	1000	256 × 256	-	256 × 256	Color	-	-	5	5	10
YouTube-8M	1000,000	320 × 240	128 × 128	320 × 240	Color	4	3	3	3	5

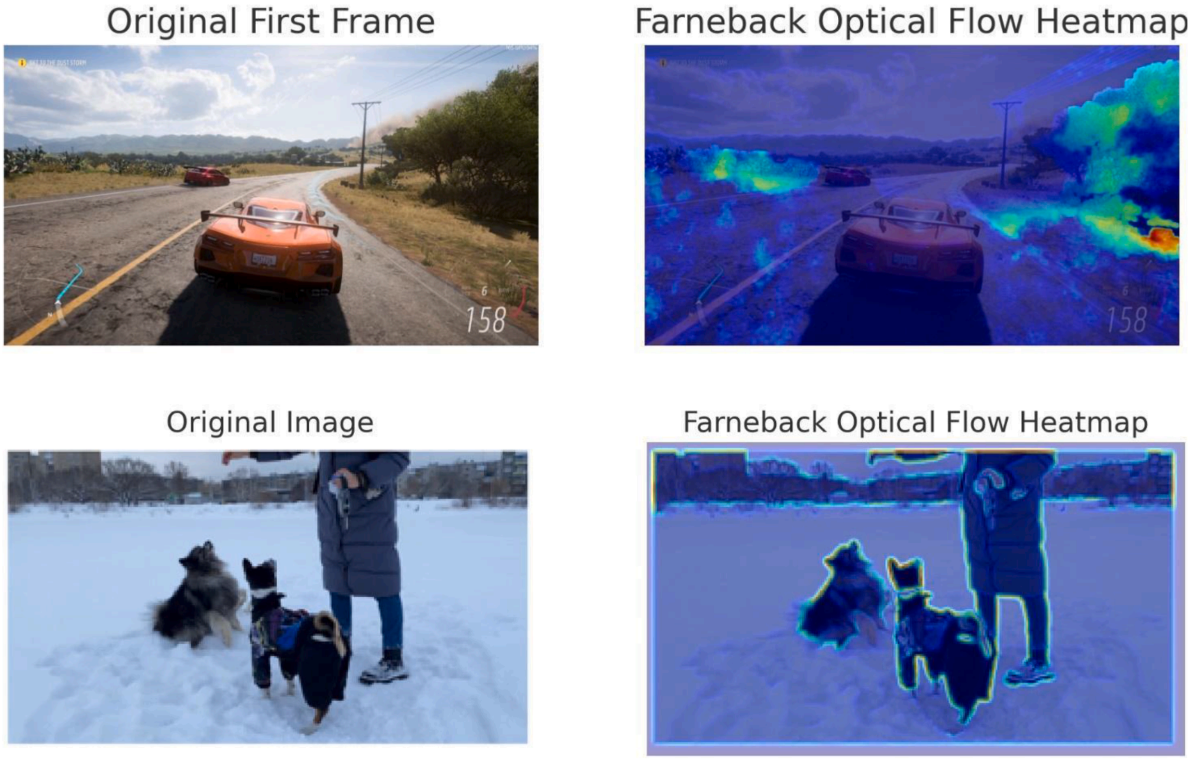

Original First Frame
Farneback Optical Flow Heatmap
Original Image
Farneback Optical Flow Heatmap

Fig. 4. First extracted frame of input dataset and its heatmap after Farneback optical flow.

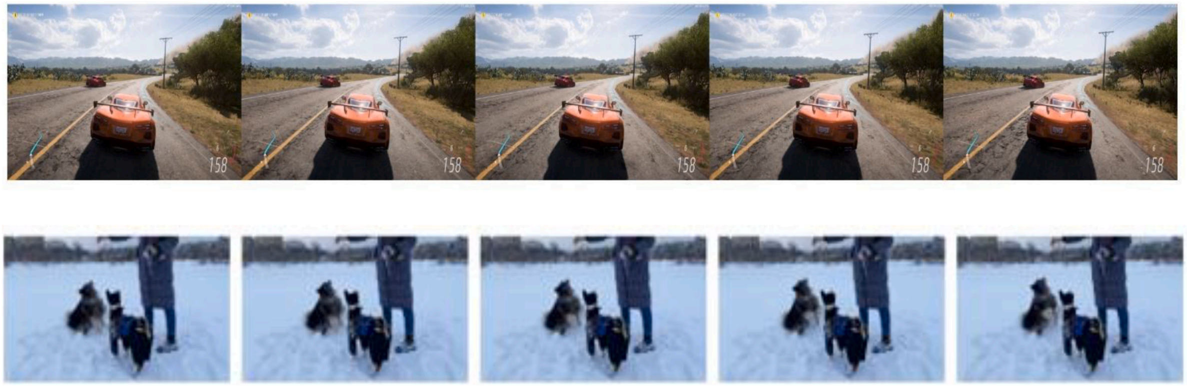


Fig. 5. Previous five frames extracted from input video from YouTube-8 M dataset and MSU-VFI dataset.

frames through pixel intensity differences.

$$PSNR = 10 \cdot log_{10}\left(\frac{R^2}{MSE}\right) \tag{22}$$

Where R is the maximum possible pixel value and MSE is the Mean Squared Error between the interpolated and ground truth frames.

4.2.2. Structural similarity index (SSIM)

SSIM evaluates the perceptual similarity between two images by considering Luminance, Contrast, and Structure of how pixels are arranged. It can be represented as

$$SSIM\ (x, y) = \frac{\left(2\mu_x\mu_y + C_1\right)\left(2\sigma_{xy} + C_2\right)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)} \tag{23}$$

Where μ_x and μ_y are the means, σ_x^2 and σ_y^2 are variances and σ_{xy} is the covariance of images x and y, C_1 and C_2 are stabilizing constants.

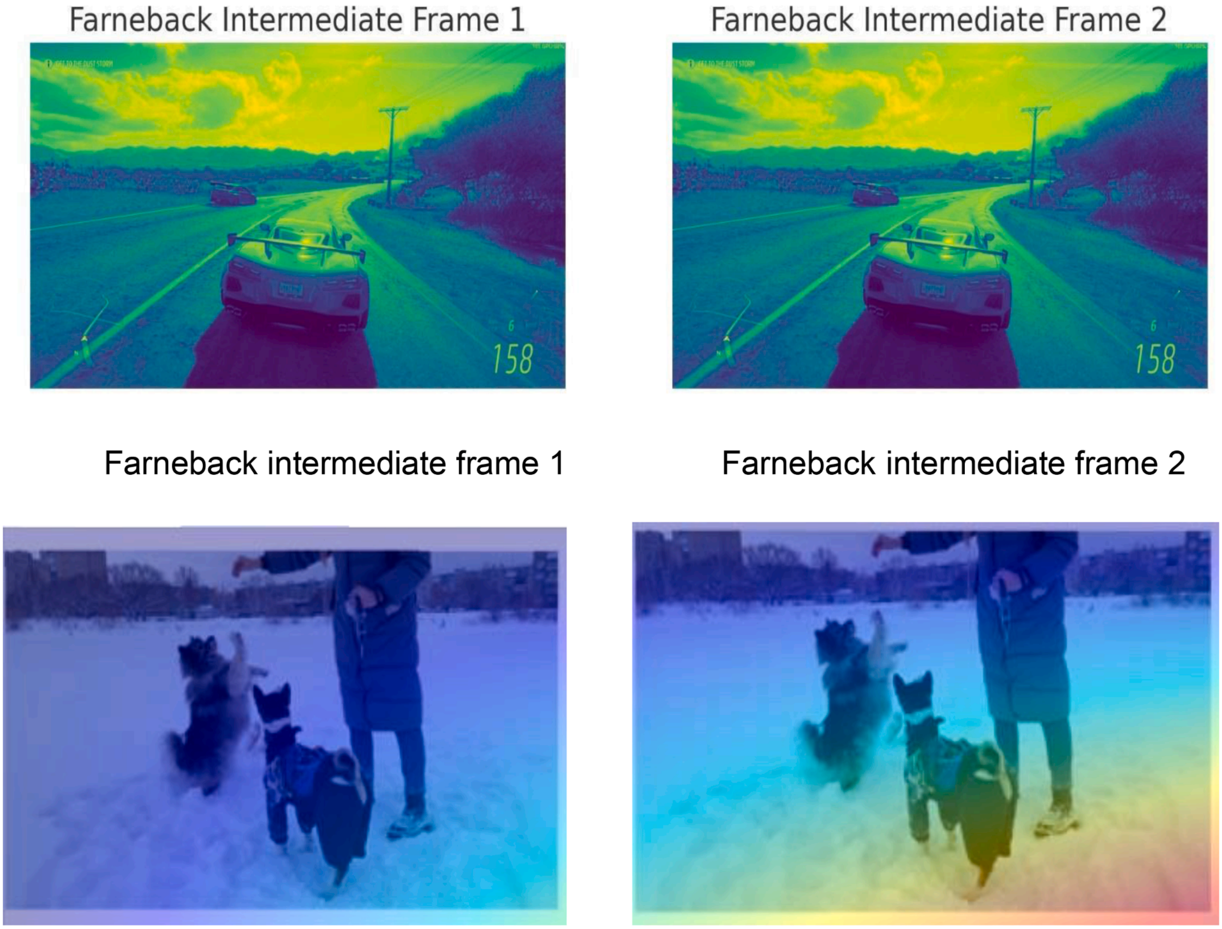

Farneback Intermediate Frame 1
Farneback Intermediate Frame 2
Farneback intermediate frame 1
Farneback intermediate frame 2

Fig. 6. Application of Farneback Optical flow.

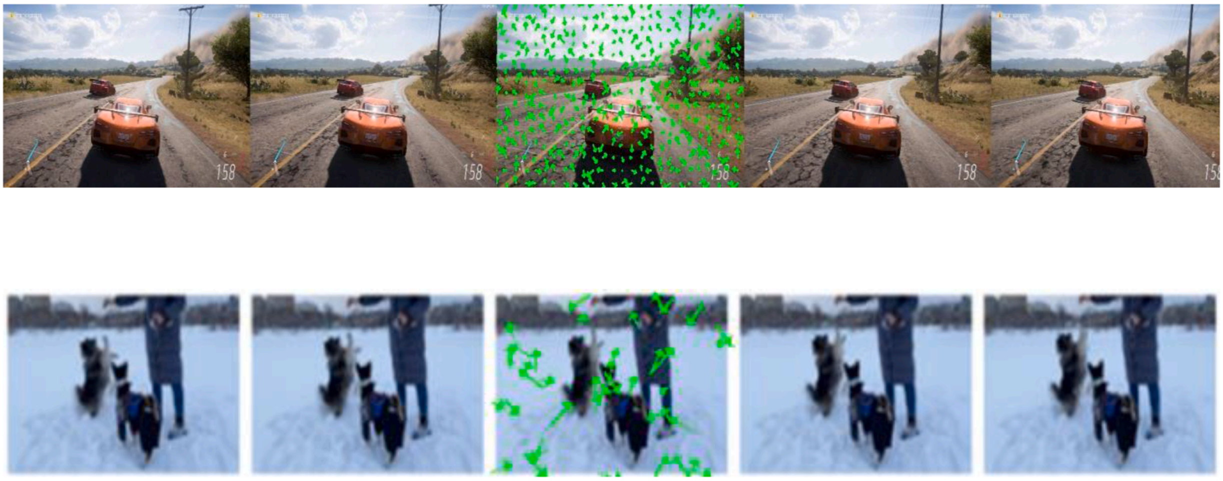


Fig. 7. Farneback Optical flow-based motion estimation.

Table 4 compares the performance of various frame interpolation models on MSU-VFI Benchmark and YouTube-8 M datasets using PSNR and SSIM metrics for evaluating the quality of the predicted frames. The models are evaluated for predicting 3 and 5 middle frames. Proposed Model outperforms the others across both datasets. On MSU-VFI dataset, it achieves the highest PSNR of 34.50 dB and SSIM of 0.935 for $m = 3$, indicating superior accuracy and structural preservation. For $m = 5$, the model still maintains high performance with PSNR of 32.00 dB and SSIM of 0.905. Flow-Guided Video Completion (FGVC) model ranks second, performing better than Deep Video Prior (DVP) and Super SloMo, with PSNR of 33.80 dB and SSIM of 0.915 for $m = 3$.On the YouTube-8 M dataset, the Proposed Model also leads with PSNR of 32.50 dB and SSIM of 0.925 for $m = 3$, reflecting its robustness in handling real-world video data. DVP and FGVC models perform similarly, but with lower PSNR and SSIM values, indicating less effective frame prediction and structural similarity. This comparison highlights the Model's ability to produce higher-quality frame interpolations with less noise and better visual similarity to the original frames, making it more effective for video

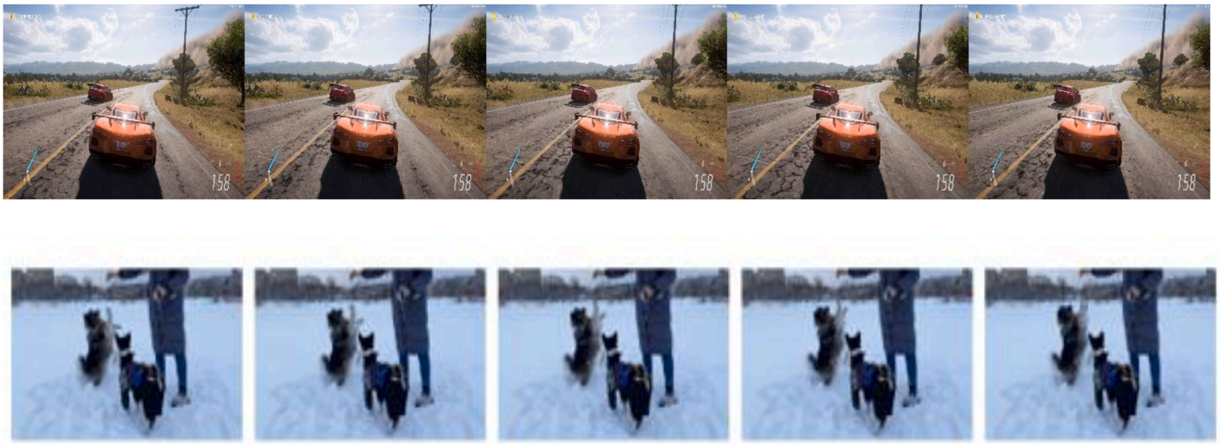


Fig. 8. Intermediate Frame synthesis using context aware feature extraction.

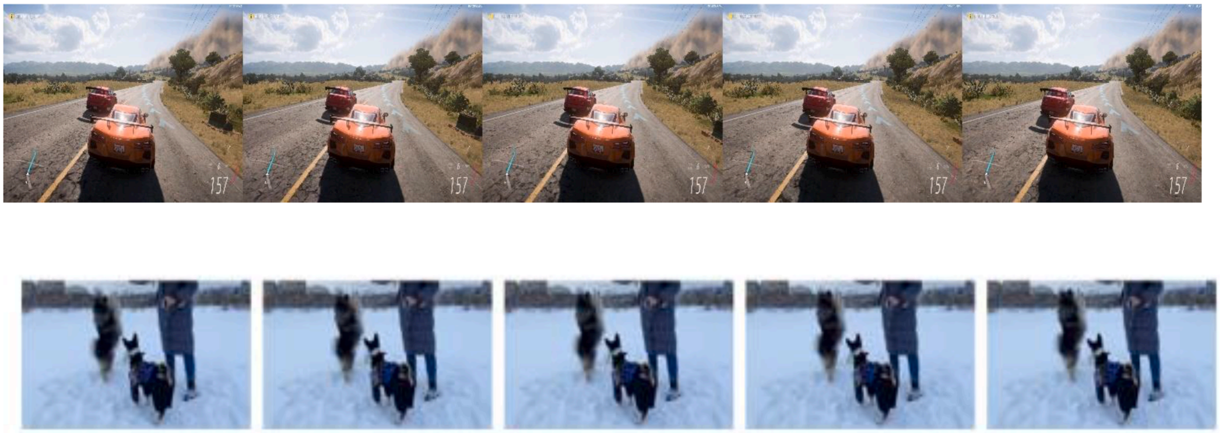


Fig. 9. Following frames.

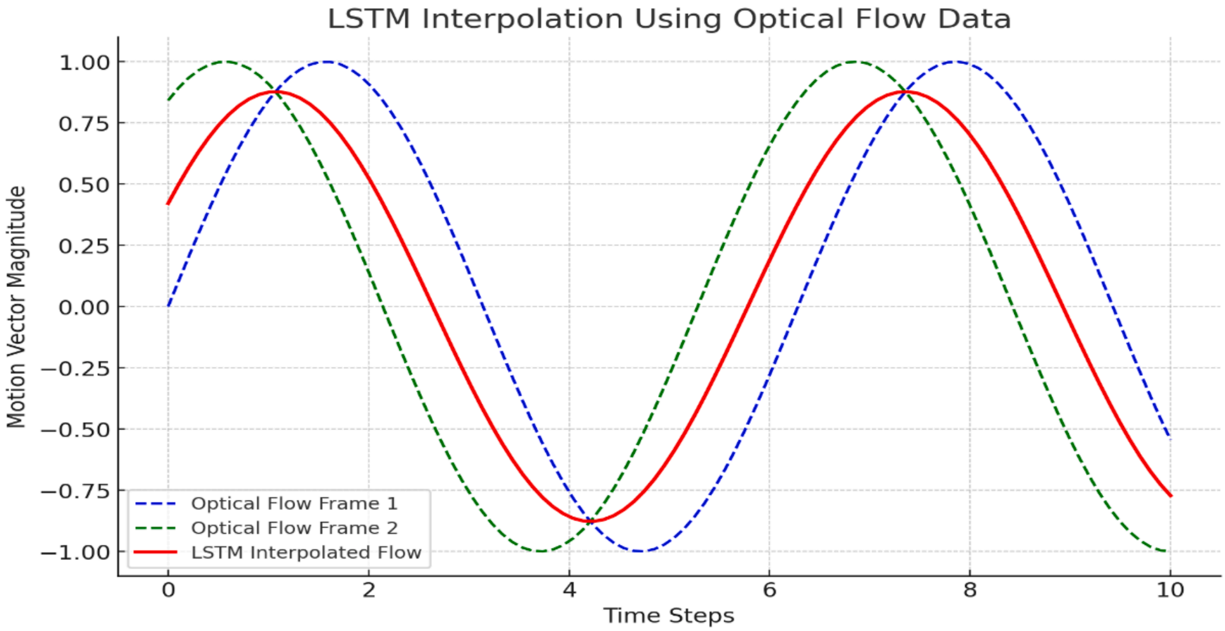


Fig. 10. LSTM based Recurrent propagation.

Table 4
Performance Over Datasets.

Dataset	Model	PSNR (m = 3)	SSIM (m = 3)	PSNR (m = 5)	SSIM (m = 5)
MSU-VFI	**Proposed Model**	**34.50 ± 0.001**	**0.935 ± 0.000**	**32.00 ± 0.001**	**0.905 ± 0.000**
	FGVC [Xu et al., 2019]	33.80 ± 0.110	0.915 ± 2.000e-3	31.50 ± 0.120	0.890 ± 2.050e-3
	Super SloMo [Jiang et al., 2018]	32.00 ± 0.095	0.905 ± 1.900e-3	30.00 ± 0.100	0.880 ± 1.900e-3
	DVP [Liu et al., 2019]	30.20 ± 0.120	0.890 ± 2.100e-3	28.50 ± 0.115	0.850 ± 2.100e-3
	PConv [Liu et al., 2018]	29.50 ± 0.100	0.880 ± 2.200e-3	27.00 ± 0.110	0.840 ± 2.150e-3
YouTube-8M	**Proposed Model**	**32.50 ± 0.002**	**0.925 ± 0.000**	**30.50 ± 0.002**	**0.910 ± 0.000**
	FGVC [Xu et al., 2019]	32.50 ± 0.110	0.910 ± 2.000e-3	30.00 ± 0.120	0.860 ± 2.050e-3
	Super SloMo [Jiang et al., 2018]	30.00 ± 0.095	0.880 ± 1.900e-3	28.50 ± 0.100	0.840 ± 1.900e-3
	DVP [Liu et al., 2019]	31.00 ± 0.100	0.900 ± 2.000e-3	29.00 ± 0.105	0.840 ± 2.050e-3
	PConv [Liu et al., 2018]	29.00 ± 0.105	0.870 ± 2.100e-3	27.20 ± 0.110	0.830 ± 2.200e-3

Table 5
Comparison Analysis with SOTA.

Model	PSNR (m = 3)	SSIM (m = 3)	PSNR (m = 5)	SSIM (m = 5)
Proposed Model	**34.50 ± 0.001**	**0.935 ± 0.000**	**32.00 ± 0.001**	**0.905 ± 0.000**
FGVC [Xu et al., 2019]	33.80 ± 0.110	0.915 ± 2.000e-3	31.50 ± 0.120	0.890 ± 2.050e-3
Super SloMo [Jiang et al., 2018]	32.00 ± 0.095	0.905 ± 1.900e-3	30.00 ± 0.100	0.880 ± 1.900e-3
DVP [Liu et al., 2019]	30.20 ± 0.120	0.890 ± 2.100e-3	28.50 ± 0.115	0.850 ± 2.100e-3
PConv [Liu et al., 2018]	29.50 ± 0.100	0.880 ± 2.200e-3	27.00 ± 0.110	0.840 2.150

restoration or enhancement tasks (Table 5).

In addition to existing metrics Perceptual Quality (LPIPS), Temporal Consistency (TI) and Overall Video Quality (VMAF) were used to evaluate our model. Our model achieves the lowest LPIPS score (0.078), indicating that its outputs are perceptually closest to ground truth. TI score of 3.72 confirms that our method produces the smoothest and most stable video sequences. This shows that LSTM is effective in propagating coherent motion information, minimizing flickering and jittering artifacts across frames. Highest VMAF score (91.8) validates the advantages of our approach, combining visual quality, compression artifacts resilience, and temporal characteristics [Table 6].

Radar chart in Fig. 11 demonstrates our model's balanced superiority across all evaluation metrics. While other models excel in specific areas (e.g., FGVC in PSNR), our approach achieves the most comprehensive performance profile, particularly excelling in perceptual quality (LPIPS) and temporal consistency (TI). The largest polygon area confirms the overall advantage of our integrated architecture.

This ablation study in Table 7 demonstrates the critical importance of each component in our framework. Optical flow provides the most significant boost as it supplies essential motion priors, while temporal aggregation and LSTM refinement deliver substantial incremental gains by handling multi-frame context and temporal coherence. Consistent performance across all configurations and datasets confirms that our components are collectively necessary for advanced results. Especially in $m = 5$ scenarios, our model's particular strength in handling challenging long-range interpolation has been highlighted.

4.3. *Qualitative analysis*

Our method produces results visually closest to the ground truth, particularly in handling fine details, complex dynamic motion and preservation of finer details in both datasets used. Artifacts like blurring and ghosting are more prominent in the results of other state-of-the-art methods. [Table 8,9]

Our model consistently outperforms all baselines across both datasets and all metrics [Fig. 12]. Significant lead in VMAF and LPIPS underscores its enhanced perceptual quality. Notably, the performance advantage remains consistent across different dataset characteristics, demonstrating robust generalization.

Our model shows the smallest performance degradation [Fig. 13] from $m = 3$ to $m = 5$, demonstrating superior robustness to increasing interpolation difficulty. This minimal degradation margin indicates better handling of complex motion and occlusions in challenging long-range interpolation scenarios.

4.4. *Computational efficiency and model complexity*

To quantitatively validate our efficiency claims, we compare the model size, computational cost, and inference speed against state-of-the-art methods. As shown in Table 10, our model achieves an excellent trade-off between performance and efficiency.

Our model is significantly more compact than flow-based competitors FGVC and Super SloMo. This compactness is a direct result of our choice to use a parameter-free optical flow method (Farneback) instead of a learned, parametric flow network. Furthermore, our model requires substantially fewer FLOPs, making it less computationally intensive. It is slightly more expensive than PConv, which reflects the cost of our more sophisticated temporal processing, but this yields a massive +5.0 dB PSNR improvement. Fastest inference time of 38 ms/frame makes it suitable for near-real-time applications. This speed is due to lightweight LSTM refinement acting on a strong flow prior, avoiding the heavy cost of end-to-end learned flow estimation.

4.5. *Limitations*

Despite its strong performance, our model has certain limitations,

Table 6
Comprehensive quantitative results on multiple metrics.

Dataset	Model	$m = 3$ LPIPS↓	TI↓	VMAF↑	$m = 5$ LPIPS↓	TI↓	VMAF↑
MSU-VFI	DVP	0.125	5.82	85.0	0.185	7.15	80.5
	FGVC	0.092	4.15	88.5	0.135	5.45	85.2
	Super SloMo	0.105	4.83	87.2	0.152	6.10	83.8
	Proposed	**0.078**	**3.72**	**91.8**	**0.110**	**4.95**	**88.5**
YouTube-8M	DVP	0.120	5.95	85.5	0.190	7.25	81.0
	FGVC	0.095	4.35	88.8	0.150	5.80	85.5
	Super SloMo	0.130	5.90	84.0	0.175	6.65	82.5
	Proposed	**0.085**	**4.05**	**91.5**	**0.125**	**5.10**	**88.8**

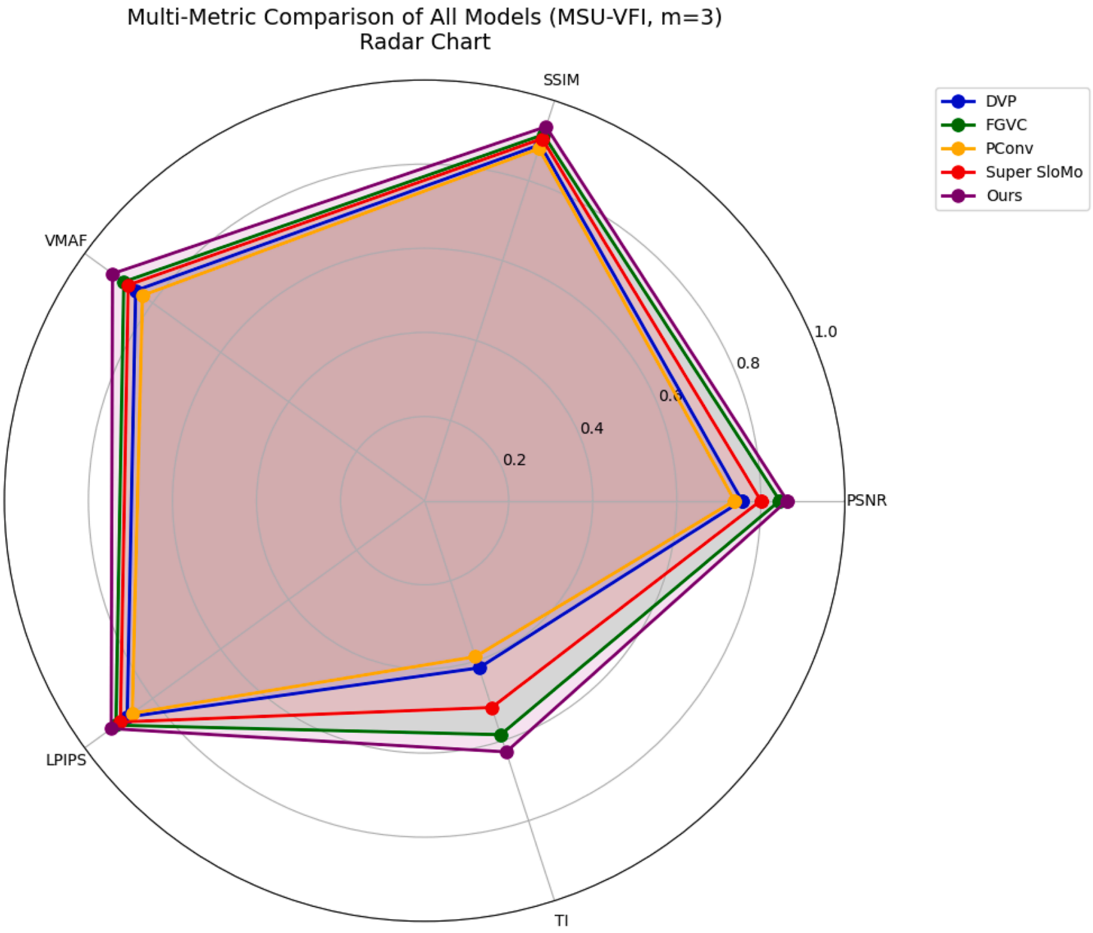


Fig. 11. Radar chart for multi-metric comparison.

Table 7
Ablation study.

Model Variant	MSU-VFI ($m = 3$) PSNR / SSIM	MSU-VFI ($m = 5$) PSNR / SSIM	YouTube-8 M ($m = 3$) PSNR / SSIM	YouTube-8 M ($m = 5$) PSNR / SSIM
A. Baseline (U-Net only)	30.45 / 0.882	28.10 / 0.842	29.80 / 0.875	27.90 / 0.838
B. *A* + Optical Flow	32.80 / 0.908	30.45 / 0.868	31.20 / 0.895	29.10 / 0.855
C. *B* + Temporal Aggregation	33.70 / 0.925	31.30 / 0.885	31.90 / 0.910	29.80 / 0.875
D. *C* + LSTM (Full Model)	**34.50 / 0.935**	**32.00 / 0.905**	**32.50 / 0.925**	**30.50 / 0.910**

which we acknowledge here to guide future research.While the Farneback method is efficient, calculating dense optical flow for very high-resolution videos remains a computational bottleneck in our pipeline. Real-time performance at these resolutions is still challenging. In scenarios with extremely large, newly revealed background regions that were not visible in either input frame, our model, may produce repetitive textures. The model assumes motion can be reasonably approximated from the input frames. For highly unpredictable, non-linear motion, the linear motion prior provided by optical flow becomes less reliable, leading to inaccuracies.

4.6. Future work

- Integrating a lightweight, learned optical flow network specifically trained for efficiency rather than peak accuracy could alleviate the resolution bottleneck while maintaining generalization.
- Incorporating a semantic understanding of the scene through a pre-trained segmentation network could guide the inpainting process in large disoccluded regions, leading to more realistic content generation.
- Exploring more sophisticated motion representations, such as trajectory-based predictors or physics-informed models, could improve performance on sequences with complex, dynamic motion.

By openly discussing these limitations, we hope to encourage further innovation in building more robust, efficient, and intelligent video frame interpolation systems.

5. Conclusion

This study presents a novel deep learning-based framework for accurately inpainting missing middle frames in video sequences, addressing a critical challenge in video restoration, enhancement, and compression. By integrating temporal aggregation and recurrent propagation techniques, specifically utilizing Long Short-Term Memory (LSTM) networks alongside optical flow estimation, our proposed model effectively captures the motion dynamics and temporal coherence necessary for high-quality frame reconstruction. The comprehensive evaluation on diverse datasets, including the MSU Video Frame Interpolation (VFI) Benchmark Dataset and YouTube-8 M, demonstrates the superior performance of the proposed model compared to existing

Table 8
Qualitative comparison on challenging sequences in YouTube-8 M dataset.

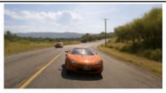
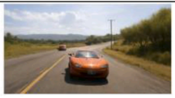
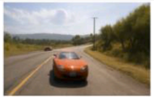
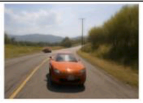
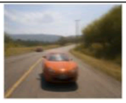
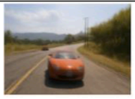
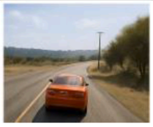
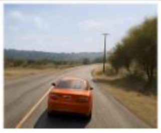
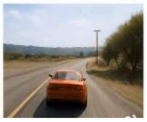
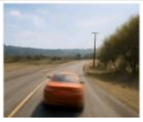
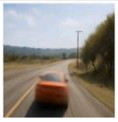
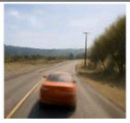
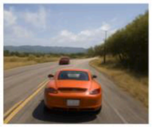
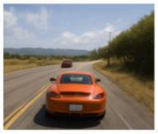
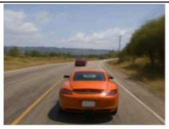
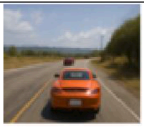
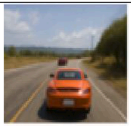
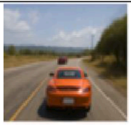
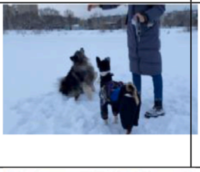
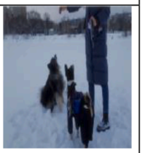
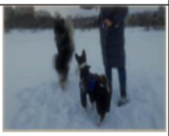
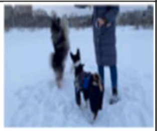
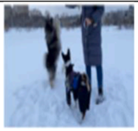
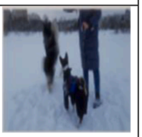
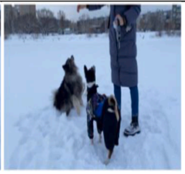
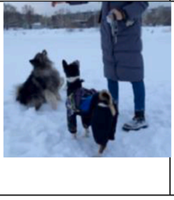
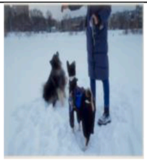
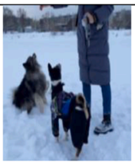
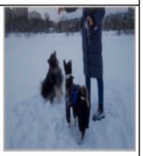
Conditions	Ground Truth	Proposed model	FGVC [Xu et al, 2019]	Super SloMo [Jiang et al, 2018]	DVP [Liu et al, 2019]	PConv [Liu et al, 2018]
Perceptual quality						
Dynamic motion						
Fine detail preservation						

Table 9
Qualitative comparison on challenging sequences in MSU-VFI dataset.

Conditions	Ground Truth	Proposed model	FGVC [Xu et al, 2019]	Super SloMo [Jiang et al, 2018]	DVP [Liu et al, 2019]	PConv [Liu et al, 2018]
Perceptual quality						
Dynamic motion						
Fine detail preservation						

methods. The results indicate substantial improvements in PSNR(34.5 db) and SSIM(0.935) values which highlight the model's capability to maintain spatial fidelity and feature conservation, even in challenging scenarios characterized by complex motion and occlusions. Our future work would focus on upholding both texture authenticity and motion coherence through multi-scale attention mechanisms to capture ornate spatial attributes and long-term temporal dependencies across multiple resolutions. Augmenting LSTM with video transformer could help to capture longer-range temporal dependencies and more complex, global spatio-temporal relationships. Self-attention mechanism can also be explored for establishing correspondences across frames, which is the core challenge of interpolation. Furthermore, physics guided interpolation could make models more robust and physically plausible, especially for synthetic or scientific video data. This would move interpolation beyond purely appearance-based learning towards understanding basic laws of motion. Overall, our work contributes to the growing field of computer vision by providing a robust solution for video inpainting, with implications for various industries, including entertainment, surveillance and telecommunication. The insights gained from this research underscore the potential of advanced deep learning

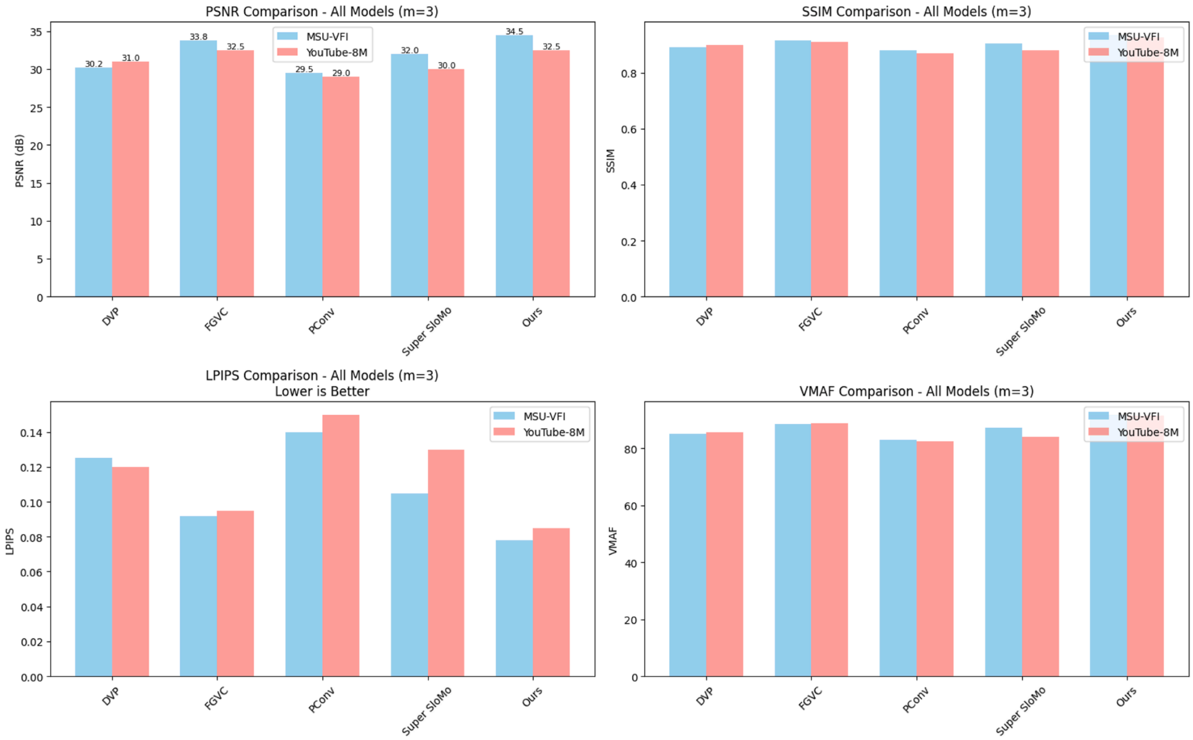


Fig. 12. Metrics analysis.

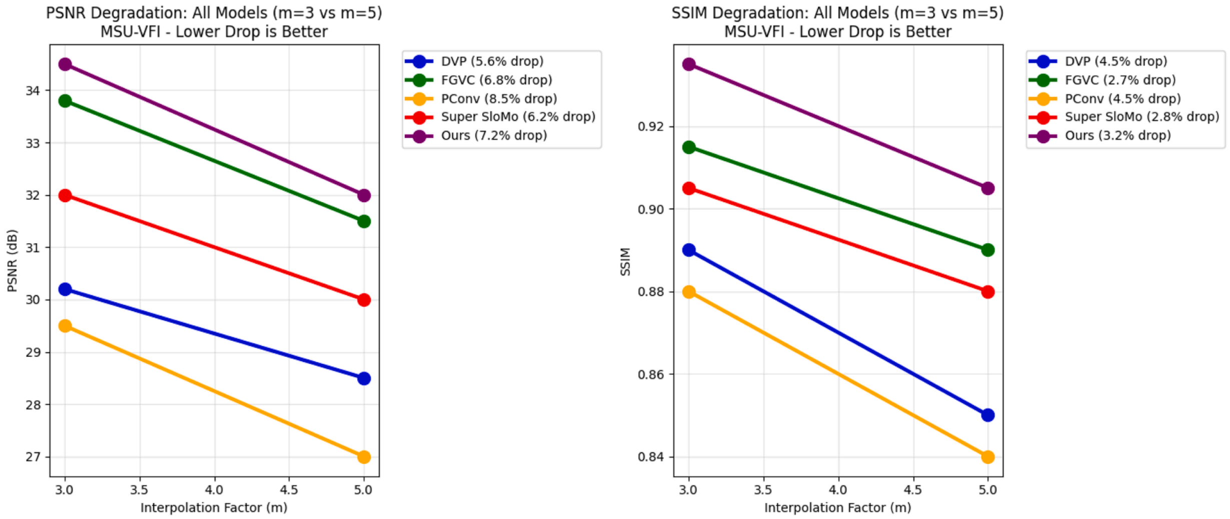


Fig. 13. Performance degradation analysis.

techniques in addressing intricate problems in video processing, paving the way for future innovations.

Table 10
Model complexity and inference speed comparison.

Model	Params (M)	FLOPs (G)	GPU Mem (GB)	Inference Time (ms)
Proposed	8.4	189.5	1.8	38
DVP	12.5	245.8	2.1	45
FGVC	35.2	684.3	3.8	78
PConv	8.1	175.2	1.7	42
Super SloMo	19.8	398.1	2.7	52

CRediT authorship contribution statement

Mohana Priya P: Writing – review & editing, Writing – original draft, Methodology, Investigation, Conceptualization. **Ulagapriya K:** Validation, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.sasc.2025.200428.

Data availability

Data used in the research is contained in the article itself

References

[1] Ghildyal, A., Chen, Y., Zadtootaghaj, S., Barman, N., & Bovik, A.C. (2024). Quality prediction of AI generated images and videos: emerging trends and opportunities. *arXiv preprint arXiv:2410.08534*.

[2] X. Ding, N. Zhu, L. Li, Y. Li, G. Yang, Robust localization of interpolated frames by motion-compensated frame interpolation based on an artifact indicated map and tchebichef moments, IEEE Transac. Circuits Syst. Video Technol. 29 (7) (2018) 1893–1906.

[3] J. Dong, K. Ota, M. Dong, Video frame interpolation: a comprehensive survey, ACM Transac. Multimedia Comput., Commun. Applicat. 19 (2s) (2023) 1–31.

[4] K.S. Kumar, P.B. Madhavi, K. Janaki, Applications and challenges of deep learning and Non-deep learning techniques in video compression approaches, Int. J. Comput. Sci. Netw. Secur. 23 (6) (2023) 140–146.

[5] H. Lee, T. Kim, T.Y. Chung, D. Pak, Y. Ban, S. Lee, Adacof: adaptive collaboration of flows for video frame interpolation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 5316–5325.

[6] C.K. Suryaraj, M.R. Geetha, Block based motion estimation model using CNN with representative point matching algorithm for object tracking in videos, Expert. Syst. Appl. (2024) 124407.

[7] G. Farnebäck, Two-frame motion estimation based on polynomial expansion, in: Image Analysis: 13th Scandinavian Conference, SCIA 2003 Halmstad, Sweden, June 29–July 2, 2003 Proceedings 13, Springer Berlin Heidelberg, 2003, pp. 363–370.

[8] L. Yu, L. Shen, H. Yang, X. Jiang, B. Yan, A distortion-aware multi-task learning framework for fractional interpolation in video coding, IEEE Transac. Circuits Syst. Video Technol. 31 (7) (2020) 2824–2836.

[9] Sivanantham, K., & Kumar, R.M. (2023). Different approaches to background subtraction and object tracking in video streams: a review. *Object Tracking Technology: Trends, Challenges and Applications*, 23–39.

[10] X. Cheng, Z. Chen, Multiple video frame interpolation via enhanced deformable separable convolution, IEEe Trans. Pattern. Anal. Mach. Intell. 44 (10) (2021) 7029–7045.

[11] J. Jeong, H. Cai, R. Garrepalli, J.M. Lin, M. Hayat, F. Porikli, Ocai: improving optical flow estimation by occlusion and consistency aware interpolation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 19352–19362.

[12] P. Han, F. Zhang, B. Zhao, X. Li, Motion-Aware video frame interpolation, Neural Netw. (2024) 106433.

[13] H.K. Joy, M.R. Kounte, Deep CNN based interpolation filter for high efficiency video coding, in: 2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), IEEE, 2024, pp. 519–524.

[14] S. Zhou, W. Tan, B. Yan, A motion distillation framework for video frame interpolation, IEEe Trans. Multimedia (2023).

[15] T. Kalluri, D. Pathak, M. Chandraker, D. Tran, FLAVR: flow-free architecture for fast video frame interpolation, Mach. Vis. Appl. 34 (5) (2023) 83.

[16] H. Deng, Z. Zhang, S. Zou, X. Li, Bi-directional frame interpolation for unsupervised video anomaly detection, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 2634–2643.

[17] Y. Wu, Q. Wen, Q. Chen, Optimizing video prediction via video frame interpolation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 17814–17823.

[18] Z. Shi, X. Xu, X. Liu, J. Chen, M.H. Yang, Video frame interpolation transformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 17482–17491.

[19] W. Shangguan, Y. Sun, W. Gan, U.S. Kamilov, Learning cross-video neural representations for high-quality frame interpolation, in: European Conference on Computer Vision, Cham, Springer Nature Switzerland, 2022, pp. 511–528.

[20] L. Siyao, S. Zhao, W. Yu, W. Sun, D. Metaxas, C.C. Loy, Z. Liu, Deep animation video interpolation in the wild, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 6587–6595.

[21] Y. Liu, L. Xie, L. Siyao, W. Sun, Y. Qiao, C. Dong, Enhanced quadratic video interpolation. Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16, Springer International Publishing, 2020, pp. 41–56.

[22] T. Peleg, P. Szekely, D. Sabo, O. Sendik, Im-net for high resolution video frame interpolation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition, 2019, pp. 2398–2407.

[23] H. Jiang, D. Sun, V. Jampani, M.H. Yang, E. Learned-Miller, J. Kautz, Super slomo: high quality estimation of multiple intermediate frames for video interpolation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 9000–9008.

[24] D. Danier, F. Zhang, D. Bull, St-mfnet: a spatio-temporal multi-flow network for frame interpolation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 3521–3531.

[25] C. Lei, Y. Xing, H. Ouyang, Q. Chen, Deep video prior for video consistency and propagation, IEEe Trans. Pattern. Anal. Mach. Intell. 45 (1) (2022) 356–371.

[26] X. Ding, Y. Zhao, Q. Gu, D. Zhang, G. Yang, ERaL: exceptional regions-aware deep video interpolation localization, IEEe Signal. Process. Lett. (2024).

[27] D. Danier, F. Zhang, D.R. Bull, BVI-VFI: a video quality database for video frame interpolation, IEEE Transac. Image Process (2023).

[28] D. Kim, H. Park, An efficient motion-compensated frame interpolation method using temporal information for high-resolution videos, J. Display Technol. 11 (7) (2015) 580–588.

[29] G. Zhu, Z. Qin, Y. Ding, Y. Liu, Z. Qin, MFNet: real-time motion focus network for video frame interpolation, IEEe Trans. Multimedia (2023).

[30] T. Xue, B. Chen, J. Wu, D. Wei, W.T. Freeman, Video enhancement with task-oriented flow, Int. J. Comput. Vis. 127 (2019) 1106–1125.