

Video Prediction: Generating Future Frames from Past Video Clips

¹Mrs.Pooja K N, ²Nanditha R

Assitant Professor Department of Master of Computer Applications, BIET, Davangere-577004.

poojakn531@gmail.com, Davangere, Karnataka, India¹

P.G. Student, Department of Master of Computer Applications, BIET, Davangere-577004.

nandithar1308@gmail.com, Davangere, Karnataka, India²

ABSTRACT-Video prediction involves generating future frames of a video sequence based on previously observed frames. It is one of the fundamental problem in computer vision which has numerous applications in autonomous driving, robotics, surveillance, and medical imaging. Recent advancements in deep learning, particularly spatiotemporal neural networks and generative models, have significantly improved the quality of future frame synthesis. This paper presents a deep learning based framework for video

prediction using convolutional recurrent networks and generative adversarial models. The Proposed System captures temporal dynamics and spatial features in order to generate realistic future frames. Experimental evaluation on benchmark video datasets demonstrates improved prediction accuracy and visual quality compared to conventional approaches. This framework provides a scalable solution for real time video forecasting tasks.

I.INTRODUCTION

Video prediction is a challenging task that involves modeling spatial and temporal dependencies to forecast future frames from past observations. Unlike static image generation, video prediction must capture motion dynamics, object interactions, and scene consistency over time. Accurate future frame prediction is essential for applications such as autonomous navigation, traffic forecasting, and human activity recognition.

Recent developments in deep learning have enabled significant progress in spatiotemporal modeling. Convolutional

Neural Networks effectively capture spatial features, while recurrent architectures such as LSTM model temporal dependencies. In 2023 and 2024, researchers demonstrated that transformer based video models and diffusion based generative frameworks outperform traditional recurrent models in long term video forecasting tasks .

Generative Adversarial Network and variational models have also been applied to produce realistic future frames while minimizing blurriness in predictions. These advancements motivate the design of a robust deep learning based video prediction

system capable of generating high quality future frames from past video clips.

II. RELATED WORK

In Recent literature from 2023–2026 highlights significant improvements in video prediction architectures. In a spatiotemporal transformer model was introduced for long term video prediction, demonstrating improved motion modeling compared to convolutional recurrent networks. Similarly, proposed a diffusion based generative framework for video forecasting, achieving superior visual realism and temporal coherence.

Recurrent convolutional models remain relevant for short term prediction tasks. A 2023 study in utilized ConvLSTM architectures combined with adversarial loss to improve frame sharpness. Hybrid architectures integrating CNNs with attention mechanisms have also shown strong performance in capturing motion patterns .

Video prediction models increasingly incorporate self-supervised learning techniques to leverage unlabeled video data. Additionally, efficient video forecasting models suitable for edge devices were explored in , emphasizing low latency deployment for real time systems.

These recent contributions provide a foundation for developing a deep learning based video prediction framework.

Some of the Related works are:

- Y. Zhang et al., “SpatioTemporal Transformer Networks for Long Term Video Prediction,” IEEE

Transactions on Pattern Analysis and Machine Intelligence, 2024.

- M. Chen et al., “Diffusion Models for Video Forecasting,” IEEE Transactions on Neural Networks and Learning Systems, 2025.
- S. Kumar et al., “Adversarial ConvLSTM for Future Frame Prediction,” IEEE Access, 2023.
- L. Wang et al., “Attention Based Deep Video Prediction Models,” IEEE Transactions on Artificial Intelligence, 2024.
- R. Patel et al., “Self-Supervised Learning for Video Forecasting,” IEEE Transactions on Image Processing, 2025.
- H. Lee et al., “Efficient Video Prediction Models for Edge Deployment,” IEEE Internet of Things Journal, 2026..

III. METHODOLOGY

The proposed framework consists of data preparation, spatiotemporal model design, training, evaluation, and deployment.

The dataset used includes benchmark video datasets such as human action sequences and moving object datasets. Each sample consists of a sequence of past frames used to predict subsequent future frames. Frames are resized to a fixed resolution such as 128×128 or 256×256 pixels and normalized.

Data preprocessing includes sequence segmentation, frame stacking, and temporal window selection. Data augmentation

techniques such as random cropping and flipping are applied to enhance generalization.

The model architecture combines convolutional layers for spatial feature extraction with temporal modeling components. A ConvLSTMbased encoder decoder network is implemented as a baseline. To improve long term prediction accuracy, a spatiotemporal transformer module is integrated into the architecture. For enhanced visual realism, an adversarial discriminator network is added to form a generative adversarial framework.

The model receives a sequence of past frames as input and generates predicted future frames. Loss functions include mean squared error for pixel reconstruction, adversarial loss for realism, and perceptual loss to preserve structural consistency. The system workflow is described as follows:

- Video Dataset Collection
- Frame Extraction and Preprocessing
- SpatioTemporal Model Initialization
- Training with Reconstruction and Adversarial Loss
- Future Frame Generation
- Performance Evaluation

Evaluation metrics include Peak Signal to Noise Ratio, Structural Similarity Index, Fréchet Video Distance, and visual qualitative assessment.

IV. RESULTS AND DISCUSSION

Experimental results demonstrate that transformer enhanced architectures outperform baseline ConvLSTM models in long term prediction tasks. The adversarial

loss improves sharpness and reduces blurriness in generated frames.

Quantitative evaluation shows improved PSNR and SSIM scores compared to traditional recurrent models. Fréchet Video Distance analysis confirms better temporal coherence and visual realism.

The model effectively captures object motion and scene dynamics in short and medium term predictions. However, prediction accuracy decreases for highly stochastic or unpredictable motion scenarios.

Lightweight model variants demonstrate suitability for real time deployment, aligning with recent research findings.

Future Frame Prediction

Chosen Loss | No loss chosen | Predict

Fig 1:Input Video

Future Frame Prediction

No file chosen

Predicted Frame



Fig 2:Output (Visualising Results)

V.CONCLUSION AND FUTURE WORK

This paper presented a deep learning based framework for video prediction using spatiotemporal neural networks and generative modeling techniques. The integration of transformer modules and adversarial training enhances prediction quality and temporal consistency.

Future work includes exploring diffusion based video generation models, incorporating multimodal inputs such as optical flow, applying self-supervised pretraining for improved representation learning, and optimizing models for edge computing environments.