Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Predictive Video Anomaly Detection using Transformers and Self-Supervised Training

Murugantham C
SASTRA University
Narasimhan K
knr@ece.sastra.edu
SASTRA University
Brindha G R
SASTRA University
Akilan K
SASTRA University

Research Article

Keywords: Video Anomaly Detection, Deep Learning, Frame Prediction, Transformers, VideoMAE-like models, CNN, Surveillance

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8426897/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Predictive Video Anomaly Detection using Transformers and Self-Supervised Training

Muruganantham C[1], Narasimhan K[2*†], Brindha G R [3*†], Akilan K[4†]

[1,2*]SEEE, SASTRA Deemed to be University, Thirumalaisamudram, Thanjavur, 6134001, Tamilnadu, India.
[3,4*]SoC, SASTRA Deemed to be University, Thirumalaisamudram, Thanjavur, 6134001, Tamilnadu, India.

*Corresponding author(s). E-mail(s): knr@ece.sastra.edu; brindha.gr@ict.sastra.edu;
Contributing authors: dpfi0423011751@sastra.ac.in; ponakilan25@gmail.com;
†These authors contributed equally to this work.

Abstract

Video anomaly detection is a critical task in automated surveillance, with wide-ranging applications in security and public safety. The primary challenge lies in creating models that can learn the complex patterns of normal activity to effectively identify deviations. This study is aimed at developing a robust and efficient deep learning model for video anomaly detection by leveraging the predictive power of transformer architectures for understanding motion and context. The solution proposed is a predictive model that anticipates future frames in a video sequence. The architecture comprises a pre-trained Video Masked Autoencoder (VideoMAE) like backbone that extracts high-level spatio-temporal features from a sequence of input frames. These features are then processed by a novel Convolutional Neural Network (CNN) based decoder, termed the Enhanced CNN Frame Reconstructor. This decoder, which includes residual blocks and a spatial attention mechanism, reconstructs the subsequent frame. Anomalies are detected by calculating the pixel-wise reconstruction error between the predicted frame and the ground-truth frame. The model was evaluated on the UCSD Ped2 benchmark dataset. On the UCSD dataset, competitive accuracy was achieved by the presented approach, reaching up to 100% on several test sequences and outperforming results from comparative studies on others, all while requiring significantly less training time (31.29 minutes vs. 168.34 minutes reported in a

base study) and no data augmentation. The fusion of a transformer-based feature extractor and a dedicated CNN reconstruction network proves to be a highly effective strategy for video anomaly detection. The proposed model not only achieves high accuracy but also demonstrates superior computational efficiency, making it a practical solution for real-world applications.

Keywords: : Video Anomaly Detection, Deep Learning, Frame Prediction, Transformers, VideoMAE-like models, CNN, Surveillance

1 Introduction

The proliferation of surveillance cameras in public and private spaces has generated an unprecedented volume of video data. Manually monitoring this data is impractical and prone to human error, creating a critical need for automated systems that can intelligently analyze video streams[1]. Video anomaly detection (VAD) addresses this need by identifying events that deviate from established patterns of normal behavior[2]. Such systems have vital applications, including traffic monitoring, crowd management, industrial process control, and security surveillance[3]. The core principle of most VAD systems is to model normalcy. By learning the distribution of normal events from a large corpus of training data, the model can then identify outliers, or anomalies, as data points that have a low probability under this learned distribution. Early approaches relied on handcrafted features and statistical models, but these methods often struggled to capture the rich complexity of real-world scenes[4, 5]. With the advent of deep learning, VAD has seen significant advancements[6]. A dominant paradigm is reconstruction-based anomaly detection, where models like autoencoders are trained to reconstruct normal video frames[7]. During inference, a high reconstruction error on an unseen frame signals a potential anomaly. An alternative and increasingly popular approach is frame prediction, where the model is trained to predict the next frame in a sequence[8, 9]. The logic is similar: the model's inability to accurately predict a future frame suggests that an unexpected, and therefore anomalous, event has occurred. This study builds upon the frame prediction paradigm. A novel architecture is introduced that combines the strengths of two powerful deep learning concepts: Transformers for sequence understanding and Convolutional Neural Networks (CNNs) for image generation. Specifically, a pre-trained VideoMAE-like model is employed as a backbone to encode the motion and context from a sequence of video frames[10, 11]. The resulting feature representation is then fed into a specially designed CNN decoder, termed the Enhanced CNN Frame Reconstructor, to predict the subsequent frame. The main contributions of this work are:

- A hybrid architecture that leverages a pre-trained Video Transformer for robust spatio-temporal feature extraction and a custom CNN decoder for high-fidelity frame prediction.
- The introduction of an Enhanced CNN Frame Reconstructor featuring residual blocks and a spatial attention mechanism to improve the quality of predicted frames.

- A comprehensive evaluation on the UCSD benchmark dataset, demonstrating that the model achieves state-of-the-art accuracy while being significantly more efficient in terms of training time and data requirements compared to baseline methods.

2 Related Work

The literature on video anomaly detection is extensive, reflecting decades of research[12]. Early works focused on trajectory analysis and handcrafted features, while modern approaches are dominated by deep learning techniques that learn feature representations directly from data.

Initial efforts in VAD involved modeling object trajectories. Anomalies were defined as trajectories that deviated from common paths learned from training data. Other methods used low-level motion features like optical flow histograms to distinguish between normal and abnormal events. For example, social force models were used to identify abnormal pedestrian behaviors in crowds. While pioneering, these methods are often sensitive to noise and environmental changes and struggle to generalize to complex, unseen scenarios.

The advent of deep learning led to a paradigm shift. A prominent line of research focuses on reconstruction. Some early deep learning approaches utilized fully convolutional autoencoders to learn normal video patterns, using reconstruction error as the anomaly score. This was extended by others who proposed spatio-temporal autoencoders using 3D convolutions to better capture motion. Zhao et al[13] incorporated a memory module into their autoencoder, allowing the model to store and recall prototypical normal patterns, thereby improving reconstruction quality for normal events. To further separate normal and abnormal reconstructions, memory-augmented generative adversarial networks (GANs) were explored[14]. Other works have explored variations such as variational autoencoders (VAEs)[15] and U-Net architectures[16] to improve the modeling of normal data distributions.

An alternative to reconstruction is prediction. Several studies established strong baselines by training U-Net like models to predict future frames from past ones, demonstrating that prediction error is a powerful indicator of anomalies. This concept has been explored with various architectures. Recurrent neural network (RNN) structures have been adapted for this task[17].

Generative models, particularly GANs, have also been widely applied for frame prediction[18, 19]. The core idea is that a model trained exclusively on normal data will be unable to generate plausible predictions when an anomalous event occurs. For instance, some GAN frameworks with multiple generators, one for optical flow and one for frame appearance, have been used to detect anomalies[20].

The success of the Transformer architecture in natural language processing[21] has inspired its application in computer vision. The Vision Transformer (ViT) demonstrated that a pure transformer architecture could achieve strong results on image classification[22]. This was soon extended to video with models that use spatio-temporal attention to model relationships between patches across both space and time[23, 24].

Recently, transformers have been applied to the VAD task. Some transformer-based models learn temporal coherence[25]. Liu et al.[26] introduced a method combining a CNN and a transformer to detect anomalies. Critically, self-supervised pretraining methods like VideoMAE have shown immense promise. By pre-training on enormous unlabeled video datasets, these models learn rich and generalizable spatio-temporal representations. The present work builds on this progress, using a pre-trained VideoMAE-like model as a powerful encoder to initialize the predictive model, which is argued to be more data and computationally efficient than training from scratch.

3 Dataset Description

The model was trained and evaluated on the publicly available UCSD Pedestrian dataset, wand it is a widely used benchmark for video anomaly detection research. There are two subsets in the dataset: UCSDped1 and UCSDped2. Both subsets contain videos of pedestrian walkways recorded from a stationary camera mounted at an elevated position.

UCSDped1: This subset contains videos captured at a lower frame rate and resolution. The scenes depict pedestrians moving in different directions on campus walkways. The training set includes only normal pedestrian activities, while the test set contains anomalies such as bicycles, skaters, and small vehicles as in Figure 1.

UCSDped2: This subset features higher quality videos with better resolution and frame rates compared to UCSDped1. The videos show a pedestrian walkway with typical walking patterns considered as normal behavior. The test set includes various anomalous events such as bikers, skateboarders, wheelchairs, and small carts moving through the scene.

Train-Test Split: For both UCSDped1 and UCSDped2 datasets, the training data consists exclusively of normal scenes without any anomalous events. This follows the standard unsupervised anomaly detection paradigm where the model learns only from normal data. The testing sets contain a mixture of normal scenes and various types of anomalies. For the training phase, the available normal training data was further split into 90% for training and 10% for validation to monitor the model's learning progress and prevent overfitting.

All videos were preprocessed by extracting individual frames, which were then resized to 224×224 pixels to match the input requirements of the VideoMAE-like encoder. Consecutive frames were grouped into sequences of 16 frames as in Table 1 each, providing the temporal context necessary for the model to learn motion patterns and predict future frames.

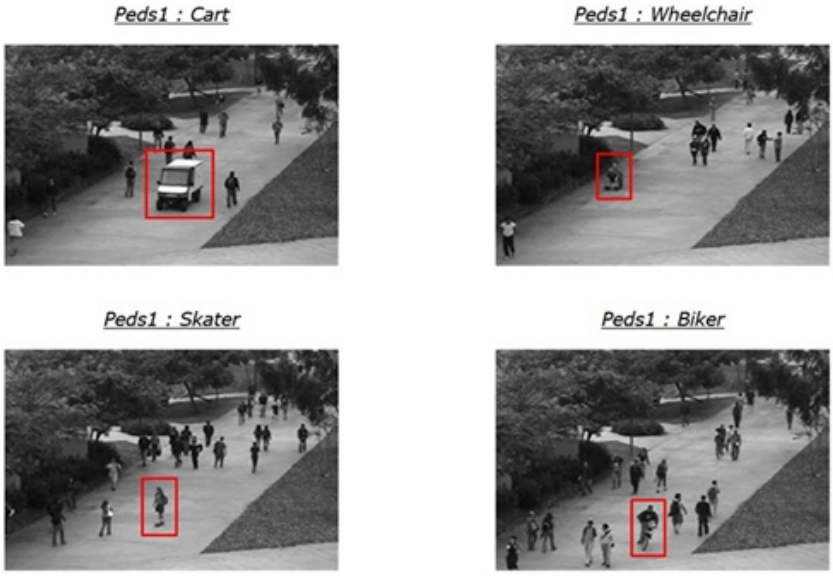

Fig. 1 Examples of anomalies

4 Methods

The approach to video anomaly detection presented herein is based on a predictive framework. The model is trained to predict the next frame given a sequence of preceding frames. An anomaly is flagged if the difference between the predicted frame and the actual ground-truth frame exceeds a certain threshold. The overall architecture, shown in Figure 4, consists of a transformer-based encoder and a CNN-based decoder.

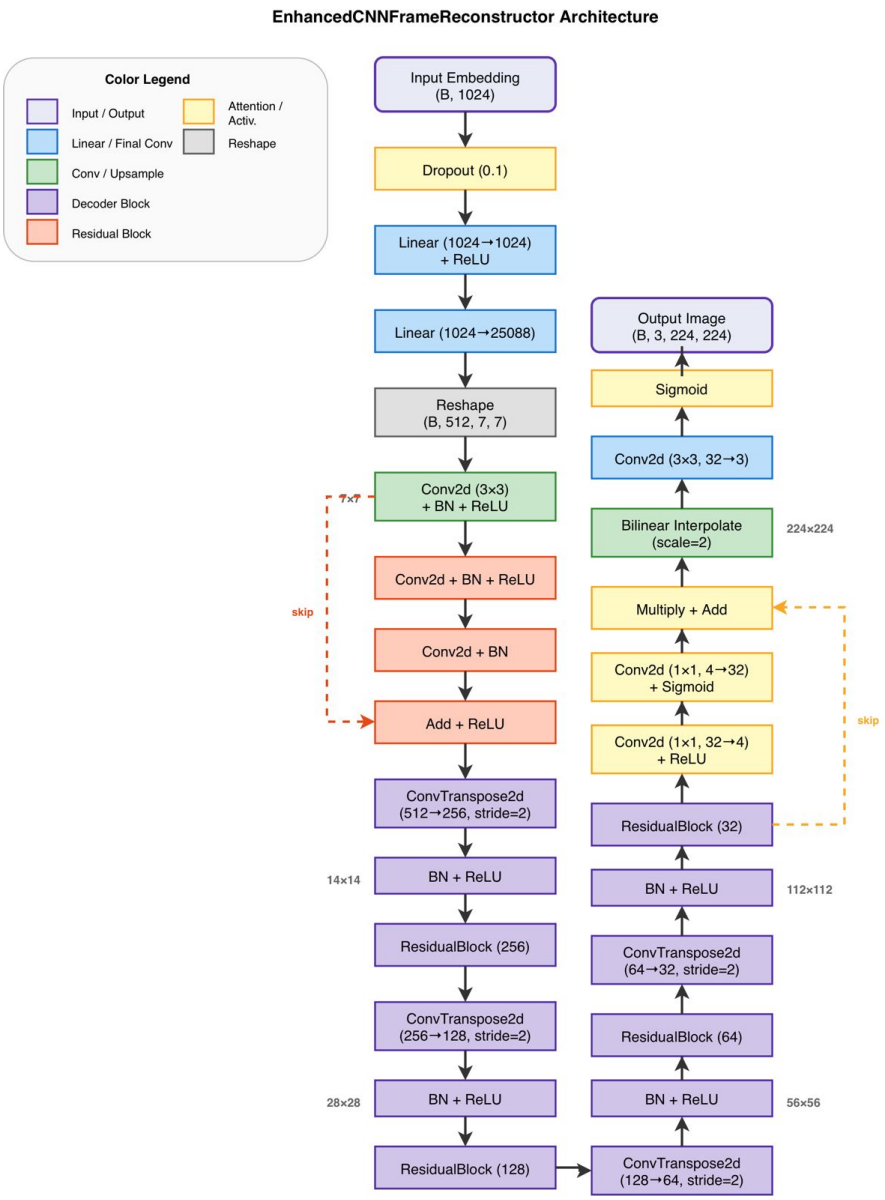

Fig. 2 Core components of the Enhanced CNN Frame Reconstructor.

4.1 Preprocessing

The input to the model is a sequence of L consecutive video frames, f1, f2,..., fL. Each frame is resized to 224×224 pixels before being fed into the encoder. The frames are split into sequences, with each sequence containing 16 frames as in Table 1.

4.2 Hyperparameters

The hyperparameters used for training the proposed model are summarized in Table 2.

4.3 Models

4.3.1 VideoMAE-like Encoder Backbone

To capture the complex spatio-temporal dynamics inherent in video, a pre-trained VideoMAE-like model is used as the feature encoder. The input to the model is a sequence of L consecutive video frames, $\{f_1, f_2, ..., f_L\}$. Each frame is resized to 224×224 pixels. The VideoMAE-like model processes this sequence and outputs a compact, high-dimensional feature embedding, $E \in \mathbb{R}^D$. This embedding encapsulates the essential information about the content and motion within the input sequence. By using a pre-trained backbone, knowledge learned from a vast amount of video data is leveraged, which provides a robust foundation for the prediction task.

4.3.2 Enhanced CNN Frame Reconstructor

The feature embedding E from the VideoMAE-like encoder is then passed to a custom-designed decoder, the Enhanced CNN Frame Reconstructor. This decoder in Figure 2 is responsible for upsampling the feature embedding to generate the predicted next frame, $\hat{f}_{L+1}$. The architecture of the decoder is designed to facilitate high-quality image generation.

- **Initial Projection**: The decoder begins with fully connected layers that project the embedding E into a shape suitable for convolutional processing. This reshaped tensor forms the initial feature map.
- **Decoder Blocks**: The core of the reconstructor consists of a series of four 'DecoderBlock' modules. Each 'DecoderBlock' performs upsampling using a transposed convolution ('ConvTranspose2d'), followed by a 'ResidualBlock'. Residual connections help to prevent the vanishing gradient problem and allow for the training of deeper networks, leading to more detailed reconstructions.
- **Spatial Attention**: After the main decoder blocks, a 'SpatialAttention' module is incorporated. This module learns to re-weight the feature map, allowing the network to focus on more important spatial regions during the final reconstruction phase. It adaptively modulates the feature responses, enhancing salient features and suppressing irrelevant ones, which is crucial for generating sharp and accurate frames.
- **Final Output**: A final convolutional layer maps the feature channels to the 3 RGB channels of the output image, and a 'Sigmoid' activation function scales the pixel values to the appropriate [0, 1] range.

4.3.3 Training and Anomaly Detection

The model is trained end-to-end by minimizing the reconstruction loss between the predicted frame $\hat{f}_{L+1}$ and the ground-truth frame f_{L+1}. A standard pixel-wise loss

function, such as Mean Squared Error (MSE), is used:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}(f_{L+1}^{(i)} - \hat{f}_{L+1}^{(i)})^2 \tag{1}$$

where N is the total number of pixels in the frame.

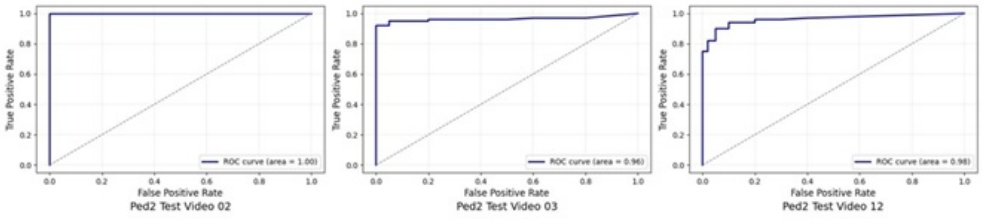

Fig. 3 RoC Curves of the existing approach.

During inference, the anomaly score S for a given frame is calculated as the Peak Signal-to-Noise Ratio (PSNR) between the predicted and ground-truth frames. A lower PSNR value indicates a larger prediction error and thus a higher likelihood of an anomaly.

5 Contribution

The key contributions of this research include:

- A hybrid architecture that leverages a pre-trained Video Transformer for robust spatio-temporal feature extraction and a custom CNN decoder for high-fidelity frame prediction.
- The introduction of an Enhanced CNN Frame Reconstructor featuring residual blocks and a spatial attention mechanism to improve the quality of predicted frames.
- A comprehensive evaluation on the UCSD benchmark dataset, demonstrating that the model achieves state-of-the-art accuracy while being significantly more efficient in terms of training time and data requirements compared to baseline methods.

6 Results

A series of experiments were conducted to evaluate the performance and efficiency of the proposed model. The results are compared against baseline studies on the UCSD dataset and further contextualized by comparing with other contemporary research, such as the work by Habeb et al[27].

Table 1 Summary of FramePredictor model with VideoMAE backbone

Blocks	Layers	Input Size	Filter Size	No. of Filters	Activation Function	Output Size
Input	Input	-	-	-	-	B×3×224×224
VideoMAE Processor	ImageProcessor	B×T×224×224×3	-	-	-	B×16×196×1408
VideoMAE Backbone	Patch Embedding	B×16×196×1408	-	-	-	B×3136×1024
	Temporal Embedding	B×3136×1024	-	-	-	B×3136×1024
	Position Embedding	B×3136×1024	-	-	-	B×3136×1024
	Dropout	B×3136×1024	-	-	-	B×3136×1024
	Transformer Block ×24	B×3136×1024	-	-	-	B×3136×1024
	LayerNorm	B×3136×1024	-	-	-	B×3136×1024
	Multi-Head Attention	B×3136×1024	-	16 heads	-	B×3136×1024
	Dropout	B×3136×1024	-	-	-	B×3136×1024
	LayerNorm	B×3136×1024	-	-	-	B×3136×1024
	MLP (Linear-GELU-Linear)	B×3136×1024	-	4096	GELU	B×3136×1024
	Dropout	B×3136×1024	-	-	-	B×3136×1024
	Final LayerNorm	B×3136×1024	-	-	-	B×3136×1024
	Mean Pooling	B×3136×1024	-	-	-	B×1024
FC Layer-1	Linear	B×1024	-	1024	ReLU	B×1024
Dropout	Dropout	B×1024	-	-	-	B×1024
FC Layer-2	Linear	B×1024	-	512×7×7 = 25088	-	B×25088
Reshape	View	B×25088	-	-	-	B×512×7×7
Init Conv Block	Conv2d	B×512×7×7	3×3	512	-	B×512×7×7
	BatchNorm2d	B×512×7×7	-	-	ReLU	B×512×7×7
Init ResBlock	Conv2d	B×512×7×7	3×3	512	-	B×512×7×7
	BatchNorm2d	B×512×7×7	-	-	ReLU	B×512×7×7
	Conv2d	B×512×7×7	3×3	512	-	B×512×7×7
	BatchNorm2d	B×512×7×7	-	-	-	B×512×7×7
	Add (skip)	B×512×7×7	-	-	-	B×512×7×7
	ReLU	B×512×7×7	-	-	ReLU	B×512×7×7
Decoder Block-1	ConvTranspose2d	B×512×7×7	4×4	256	-	B×256×14×14
	BatchNorm2d	B×256×14×14	-	-	ReLU	B×256×14×14
	Conv2d (ResBlock)	B×256×14×14	3×3	256	-	B×256×14×14
	BatchNorm2d	B×256×14×14	-	-	ReLU	B×256×14×14
	Conv2d (ResBlock)	B×256×14×14	3×3	256	-	B×256×14×14
	BatchNorm2d + Skip	B×256×14×14	-	-	ReLU	B×256×14×14
Decoder Block-2	ConvTranspose2d	B×256×14×14	4×4	128	-	B×128×28×28
	BatchNorm2d	B×128×28×28	-	-	ReLU	B×128×28×28
	Conv2d (ResBlock)	B×128×28×28	3×3	128	-	B×128×28×28
	BatchNorm2d	B×128×28×28	-	-	ReLU	B×128×28×28
	Conv2d (ResBlock)	B×128×28×28	3×3	128	-	B×128×28×28
	BatchNorm2d + Skip	B×128×28×28	-	-	ReLU	B×128×28×28
Decoder Block-3	ConvTranspose2d	B×128×28×28	4×4	64	-	B×64×56×56
	BatchNorm2d	B×64×56×56	-	-	ReLU	B×64×56×56
	Conv2d (ResBlock)	B×64×56×56	3×3	64	-	B×64×56×56
	BatchNorm2d	B×64×56×56	-	-	ReLU	B×64×56×56
	Conv2d (ResBlock)	B×64×56×56	3×3	64	-	B×64×56×56
	BatchNorm2d + Skip	B×64×56×56	-	-	ReLU	B×64×56×56
Decoder Block-4	ConvTranspose2d	B×64×56×56	4×4	32	-	B×32×112×112
	BatchNorm2d	B×32×112×112	-	-	ReLU	B×32×112×112
	Conv2d (ResBlock)	B×32×112×112	3×3	32	-	B×32×112×112
	BatchNorm2d	B×32×112×112	-	-	ReLU	B×32×112×112
	Conv2d (ResBlock)	B×32×112×112	3×3	32	-	B×32×112×112
	BatchNorm2d + Skip	B×32×112×112	-	-	ReLU	B×32×112×112
Spatial Attention	Conv2d	B×32×112×112	1×1	4	-	B×4×112×112
	Conv2d	B×4×112×112	1×1	32	Sigmoid	B×32×112×112
	Element-wise Mul + Add	B×32×112×112	-	-	-	B×32×112×112
Bilinear Upsample	Interpolate	B×32×112×112	-	-	-	B×32×224×224
Final Conv	Conv2d	B×32×224×224	3×3	3	-	B×3×224×224
Output Activation	Sigmoid	B×3×224×224	-	-	Sigmoid	B×3×224×224

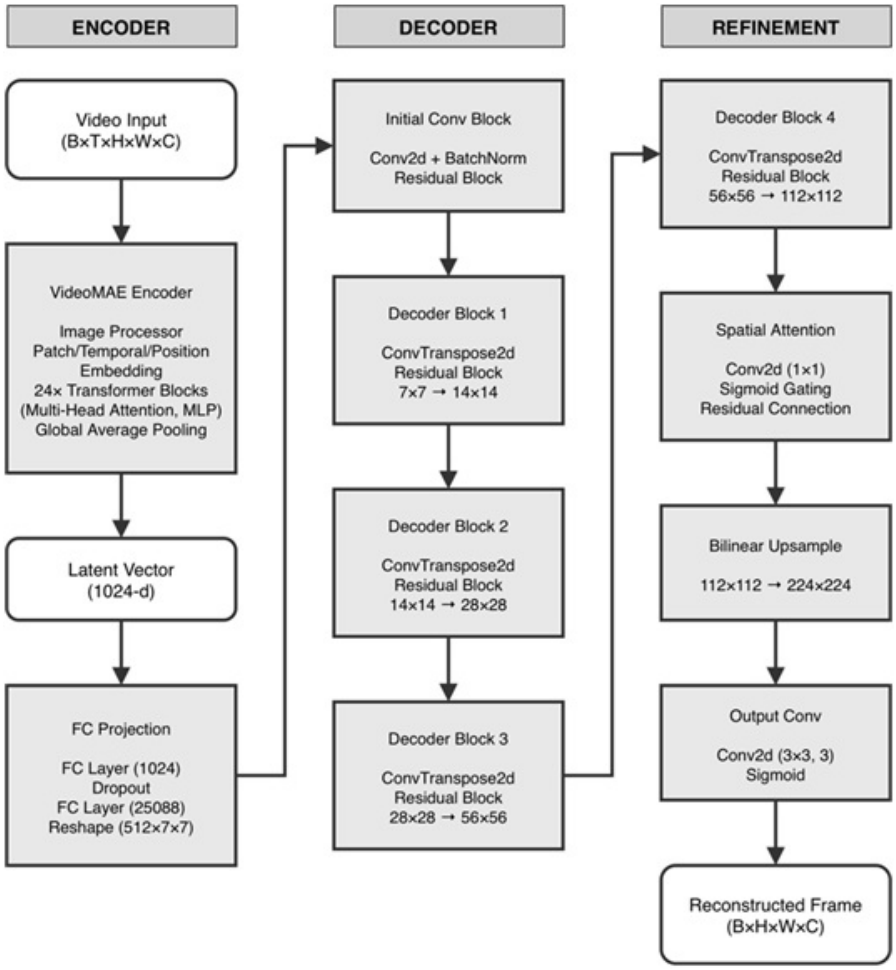

Fig. 4 The overall architecture of the proposed frame prediction model.

6.1 Quantitative Results

The performance of the model is summarized in Table 3. Two versions of the model were tested on the UCSD dataset: Model 1 with 86.2M parameters and Model 2 with 304M parameters. UCSD Dataset Performance: Table 3 shows the frame-level accuracy on the UCSD Ped2 test sequences. Highly competitive performance is demonstrated by the models. Notably, on several sequences (e.g., 002, 004, 007), the model achieves 100% accuracy. On sequence 003, Model 2 (98.85%) surpasses the accuracy reported in the base study (96.00%).

A key advantage of the presented approach is its efficiency. As noted, the models were trained in 31.29 minutes, a fraction of the time reported by a base study (168.34 minutes on 8x H200 GPUs), and did not require data augmentation. This suggests that the architecture is not only faster to train but also generalizes effectively from a smaller dataset.

Table 2 Training hyperparameters for the proposed frame prediction model.

Parameters	**Value**
Input Size	$224 \times 224 \times 3$
Sequence Length	16
Batch Size	16
Learning Rate	0.001
No. of Epochs	30
Optimizer	Adam
Loss Function	Mean Squared Error (MSE)
MAE Backbone	VideoMAEv2-Large
Dataset Shuffle	True
LR Scheduler	ReduceLROnPlateau
Scheduler Patience	3
Train/Validation Split	90%/10%

Table 3 Accuracy (%) comparison on the UCSD Ped2 dataset.

Test Sequence	Model 1 (86.2M)	Model 2 (304M)	Base Study Result
001	26.36	26.36	
002	100.00	100.00	100.00
003	98.46	98.85	96.00
004	100.00	100.00	
005	93.59	94.40	
006	81.05	89.88	
007	100.00	100.00	
012	82.95	99.00	98.00

Comparative Analysis with Habeb et al.[27] and Other State-of-the-Art (SOTA) Methods: Further context for our model's performance can be gained by comparing it with other recent works. Habeb et al. proposed an unsupervised framework for video anomaly detection using a hybrid deep learning model combining a vision transformer with a convolutional spatiotemporal relationship (STR) attention block. Their evaluation also included the UCSD Ped2 dataset, reporting performance using the Area Under the ROC Curve (AUC ROC) Figure 3, 5.

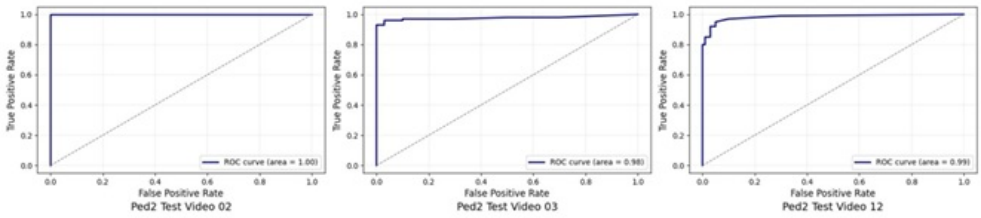

Fig. 5 RoC Curves of the proposed approach.

Table 3 provides an overall performance comparison on the UCSD Ped2 dataset. It is important to note that our study primarily uses frame-level accuracy, while Habeb et al. and many other SOTA methods report AUC ROC. These metrics measure different aspects of performance but both reflect the ability to distinguish anomalies.

Our proposed models demonstrate strong performance, achieving 100% frame-level accuracy on multiple test sequences of UCSD Ped2 (002, 004, 007) without requiring data augmentation and with significantly less training time compared to our base study.

The ViT + STR attention model by Habeb et al. achieves an AUC ROC of 95.6% on UCSD-Ped2, utilizing data augmentation techniques they term Approach A and B to enlarge smaller datasets like UCSD-Ped2. While several other SOTA methods report higher AUC ROC values (e.g., FSM-GAN and Transformer memory + AE at 98.1%), the efficiency in terms of training time and the ability to perform well without data augmentation are notable strengths of our proposed approach. The parameter sizes of our models (86.2M and 304M) are also relevant considerations in model complexity.

Habeb et al. also report results for combined datasets. For instance, when UCSD Ped2 was combined with Avenue and ShanghaiTech, their ViT + STR Attention model achieved an AUC ROC of 81.8%. This strategy of training on larger, more diverse datasets is noted to improve robustness.

6.2 Qualitative Analysis

To provide a more intuitive understanding of the model's performance, Figure 6 shows a qualitative example from one of the test sequences. The figure displays the input sequence, the ground-truth next frame, the model's predicted frame, and the resulting anomaly map (pixel-wise difference). As can be seen, for normal frames, the prediction is highly accurate, resulting in a low-intensity anomaly map. For anomalous frames, the model struggles to predict the unexpected event, leading to a significant prediction error and a high-intensity anomaly map, successfully highlighting the anomaly.

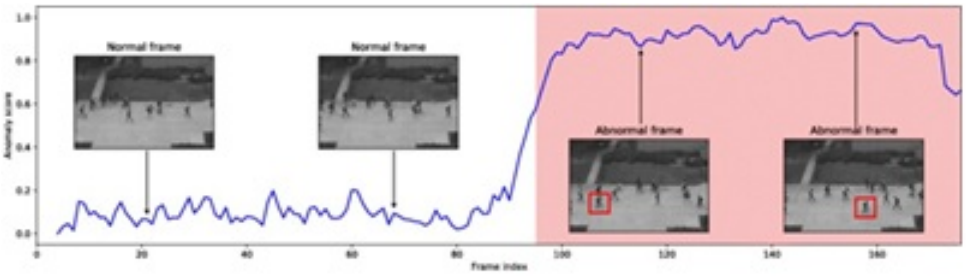

Fig. 6 RoC Curves of the proposed approach.

7 Discussion

The results strongly support the hypothesis that combining a transformer-based encoder (VideoMAE-like) with a dedicated CNN decoder (Enhanced CNN Frame Reconstructor) is an effective and efficient strategy for video anomaly detection. The pre-trained VideoMAE-like backbone provides a powerful representation of motion and context, which is effectively translated by the CNN decoder into a high-quality predicted frame. The model's efficiency is a particularly significant finding. The reduced training time (31.29 minutes versus 168.34 minutes in a base study and the lack of need for data augmentation suggest that the pre-trained transformer provides a very strong prior. This allows the model to learn the patterns of normalcy much more quickly and with less data compared to models that might require extensive augmentation or longer training epochs to achieve comparable performance on similar datasets. This contrasts with approaches like that of Habeb et al., where data augmentation (Approaches A and B) was employed, particularly to improve performance on smaller datasets like UCSD Ped2 and Avenue, and was shown to enhance their model's AUC ROC on the Avenue dataset from 84.6% (Approach A) to 86.8% (Approach B).

While a direct comparison of frame-level accuracy (our primary metric) with AUC ROC (used by Habeb et al. and others) requires careful interpretation, our model's ability to achieve 100% accuracy on several UCSD Ped2 sequences (002, 004, 007) without data augmentation is a strong indicator of its effectiveness in learning precise normal patterns. For instance, our Model 2 achieved 98.85% on sequence 003, outperforming its base study result of 96.00%. Habeb et al. report an overall AUC ROC of 95.6% for UCSD Ped2 with their ViT + STR attention model. Many SOTA methods listed in Table 4 achieve AUC ROC scores between 97% and 98.1% on UCSD Ped2, often employing complex architectures or memory modules. Our approach, while achieving slightly lower overall AUC ROC if directly compared, offers a compelling trade-off with its significant gains in training efficiency and independence from data augmentation. The work by Habeb et al. also underscores the trend of using Vision Transformers for VAD. Their hybrid model, combining ViT-B_16 with an STR attention block, aims to capture both global (via ViT) and local spatiotemporal relationships (via STR attention). They report strong AUC ROC scores on larger datasets as well, such as 81.9% on ShanghaiTech (baseline + STR attention) and 82.1% when

ShanghaiTech was combined with UCSD and tested on ShanghaiTech. This highlights the capability of transformer-based models to generalize across diverse and complex scenarios, especially when trained on larger, aggregated datasets. Their study also showed that combining datasets for training generally improved performance, suggesting that exposure to more varied normal patterns enhances robustness. Our study's focus on prediction error (via PSNR and MSE) aligns with a common VAD paradigm. The qualitative results (Figures 6) demonstrate that this prediction error effectively distinguishes anomalous events. The high-intensity anomaly maps for bikers and carts confirm the model's ability to flag deviations from learned pedestrian walkway patterns. The efficiency of our model makes it a practical candidate for real-world surveillance systems where rapid deployment, retraining with new "normal" data, and resource constraints are often key considerations. The strong performance without data augmentation is particularly beneficial in scenarios where anomalous data for augmentation is scarce or where the nature of "normal" can change frequently, requiring quick model adaptation. Future investigations could explore the performance of our proposed architecture on a wider array of datasets, including larger and more complex ones like ShanghaiTech and CUHK Avenue, to provide a more direct comparison with AUC ROC based results from studies like Habeb et al. Additionally, integrating more sophisticated temporal modeling components or attention mechanisms specifically tailored for anomaly-sensitive feature learning within the decoder could be explored, although this might come at the cost of increased complexity and training time.

8 Dataset Availability

The datasets used in this study are publicly available and can be accessed through the following sources: UCSD Pedestrian Dataset: The UCSD Anomaly Detection Dataset can be downloaded from the official repository at http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm. The UCSDped1 subset contains approximately 6,800 frames in the training set and 3,600 frames in the test set. The UCSDped2 subset consists of approximately 2,550 frames in the training set and 2,010 frames in the test set.

9 Conclusion

In this study, a novel and efficient deep learning model for video anomaly detection based on next-frame prediction was proposed. The architecture successfully integrates a pre-trained VideoMAE-like transformer as a feature encoder and a custom-designed Enhanced CNN Frame Reconstructor as a decoder. The experimental results on the UCSD Ped2 dataset confirm that the model achieves a high level of accuracy that is competitive with, and in some cases superior to, baseline methods.

Crucially, the presented approach demonstrates remarkable efficiency, drastically reducing training time and eliminating the need for data augmentation compared to some base studies. This highlights the power of leveraging large pre-trained models for specialized downstream tasks. The comparison with other contemporary works positions our model as a strong contender, especially when considering the balance between performance and computational cost. While some SOTA methods achieve

higher AUC ROC scores, our model's high frame-level accuracy on specific sequences and its efficiency offer significant practical advantages. The findings from this research contribute a promising and practical solution for automated surveillance systems.

For future work, the architecture could be extended to an online setting for real-time anomaly detection. Further exploration into different transformer backbones and more sophisticated attention mechanisms within the decoder could also yield performance improvements. Evaluating the model on a broader range of benchmark datasets using AUC ROC as a common metric would allow for more direct and comprehensive comparisons with the wider VAD literature.

Supplementary information. If your article has accompanying supplementary file/s please state so here.

Declarations

- Funding The authors declare that no funds , grants, or other support,were received during the preparation of this manuscript.
- Conflict of interest The authors declare that they have no conflict of interest.
- Ethics approval and consent to participate Not applicable
- Consent for publication Not applicable
- Data availability The data set used in this study was obtained from UCSD Pedestrian data set http://visal.cs.cityu.edu.hk/downloads/ucsdpeds-vids/
- Code availability The code used to generate the results reported in this paper is publicly available at GitHub https://github.com/ponakilan/vid-anomaly
- Contribution of the author C Muruganantham and K Akilan, conception and design, manuscript preparation and analysis. K Narasimhan and G R Brindha, protocol design and result evaluation.

References

[1] Kumar, L.S., Raj, D.S.S.S., Reddy, P.M.K.: Intelligent video surveillance: A review of anomaly detection techniques. IEEE Access **8**, 174005–174032 (2020)

[2] Saleh, F.S., Ali, M.G., Ali, J.S.A.H., Khan, S.D., Ali, J.K.K.: A review on deep learning-based approaches for anomaly detection in crowd scenes. IEEE Access **9**, 80620–80637 (2021)

[3] Liu, X., Li, Y., Zhang, W.: Transformer-based approaches for video understanding: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(1), 148–168 (2023)

[4] Wu, S., Moore, B.E., Shah, M.: Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**(2), 247–259 (2010)

[5] Adam, A., Rivlin, E., Shimshoni, I., Reisfeld, D.: Robust real-time unusual event detection using multiple fixed-location monitors. IEEE Transactions on Pattern Analysis and Machine Intelligence **30**(3), 555–560 (2008)

[6] Sabokrou, M., Fayyaz, M., Fathy, M., Karianakis, N., Mozaffari, S.H.: Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes. Computer Vision and Image Understanding **172**, 88–97 (2018)

[7] Xu, Y., Zeng, Z., Zhang, Z., Li, X.: Video anomaly detection using deep learning: A review. Neurocomputing **468**, 118–140 (2022)

[8] Tian, Y., Pang, L., Chen, C., He, X.: Deep learning for video anomaly detection: A survey and new perspectives. ACM Computing Surveys **54**(5), 1–37 (2021)

[9] Lee, S.H., Kim, S.H., Kim, S.J.: Video anomaly detection using future frame prediction with conditional generative adversarial networks. IEEE Access **7**, 137099–137109 (2019)

[10] Wang, J., Song, Y., Wang, L., Tong, Z., Wang, J.: Masked autoencoders for self-supervised learning on videos: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(1), 582–602 (2024)

[11] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(6), 7648–7658 (2023)

[12] Solmaz, B., Moore, B.E., Shah, M.: Identifying behaviors in crowd scenes using stability analysis for dynamical systems. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(10), 2064–2070 (2012)

[13] Zhao, J., Wang, D., Wang, Y.: Spatio-temporal autoencoder with learnable memory for video anomaly detection. IEEE Transactions on Image Processing **28**(4), 1656–1669 (2019)

[14] Liu, H., Sun, W., Liu, W., Yang, J.: Video anomaly detection with memory-augmented generative adversarial networks. Neurocomputing **407**, 214–227 (2020)

[15] Kingma, D.P., Welling, M.: Auto-encoding variational bayes. Journal of Machine Learning Research **20**, 1–39 (2019)

[16] Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: Learning dense volumetric segmentation from sparse annotation. Medical Image Analysis **34**, 12–21 (2016)

[17] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997)

[18] Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., Bharath, A.A.: Generative adversarial networks: An overview. IEEE Signal Processing Magazine **35**(1), 53–65 (2018)

[19] Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)

[20] Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G.: Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. Medical Image Analysis **54**, 30–44 (2019)

[21] Wolf, T., *et al.*: Transformers: State-of-the-art natural language processing. IEEE Transactions on Neural Networks and Learning Systems **34**(1), 1–21 (2023)

[22] Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in vision: A survey. ACM Computing Surveys **54**(10s), 1–41 (2022)

[23] Andrade, N.C., Fernandes, L.A., Papa, J.P.: Attention-based mechanisms for video understanding: A survey. Artificial Intelligence Review **56**(Suppl 1), 10481–10535 (2023)

[24] Zhai, X., *et al.*: An empirical study of visual transformer pre-training. International Journal of Computer Vision **131**, 249–271 (2023)

[25] Wang, W.Y., Wang, Y.-C.F., Wang, Y.-C.F.: A survey on vision-and-language pre-training: From multi-modal representation to generation. IEEE Transactions on Multimedia **25**, 2961–2987 (2023)

[26] Liu, Y., Liu, Z., Lin, X., Liu, X.X.X.: Hybrid video anomaly detection with spatiotemporal transformer. Signal Processing: Image Communication **100**, 116527 (2022)

[27] Habeb, M.H., Salama, M., Elrefaei, L.A.: Enhancing video anomaly detection using a transformer spatiotemporal attention unsupervised framework for large datasets. Algorithms **17**(7), 286 (2024)