Article

Not peer-reviewed version

Global Temporal Attention-Driven Transformer Model for Video Anomaly Detection

Jie Liu *

Posted Date: 12 May 2025

doi: 10.20944/preprints202505.0838.v1

Keywords: Video anomaly detection; Transformer; global temporal attention; intelligent surveillance

 doi:10.20944/preprints202505.0838.v1

Article

Global Temporal Attention-Driven Transformer Model for Video Anomaly Detection

Jie Liu

University of Minnesota, Minneapolis, USA; jielumn@gmail.com

Abstract: Video anomaly detection is an important task in the field of computer vision and is widely used in scenarios such as intelligent monitoring, security prevention, and behavior analysis. Traditional methods have limitations in dealing with long-term dependencies and modeling global temporal information, making it difficult to accurately identify complex abnormal behaviors. To this end, this paper proposes a video anomaly detection method based on Transformer and global temporal attention mechanism to improve the modeling ability of long-term dependencies. Specifically, the Transformer structure is first used to capture the global relationship between frames, and the global temporal attention mechanism is introduced to optimize the extraction of key time step information. The experiment is conducted on the public dataset and compared with methods such as Conv2D-AE, STAE, ConvLSTM-AE, TSC, MemAE, MNAD-AE, and Stack-RNN. The results show that Ours method achieves the best performance in terms of AUC and EER indicators. In addition, the training loss curve and anomaly score visualization analysis further verify the stability and effectiveness of the model. Future research can explore more efficient attention mechanisms and multimodal fusion to further improve anomaly detection capabilities.

Keywords: video anomaly detection; Transformer; global temporal attention; intelligent surveillance

I. Introduction

In recent years, with the rapid development of computer vision and deep learning technologies, video surveillance systems have been widely deployed in fields such as public safety, industrial production, and traffic management [1]. These systems generate massive volumes of video data on a daily basis, rendering the efficient extraction of anomalous behaviors a critical challenge in intelligent surveillance. Traditional video anomaly detection methods primarily depend on handcrafted feature extraction and rule-based discriminative models, which often struggle to maintain consistent detection performance amidst complex scene variations and vast datasets. Recently, deep learning-based approaches have become mainstream, and the introduction of Transformer architectures has demonstrated significant advantages in modeling long-term dependencies and integrating global information [2]. Nonetheless, effectively capturing long-range dependencies between video frames and incorporating a global temporal attention mechanism to enhance detection accuracy remains a significant research challenge.

In the context of video anomaly detection, anomalies typically refer to behaviors that are rare or deviate from conventional patterns in specific scenarios, such as sudden running within a crowd, vehicles moving in the wrong direction in normal traffic, or equipment failures on industrial production lines. Due to the often transient and localized nature of anomalous events, coupled with a general lack of clear prior information, anomaly detection is inherently more complex than traditional classification or object detection tasks. Compared with conventional methods based on convolutional neural networks (CNNs) or recurrent neural networks (RNNs), Transformer modes, owing to their self-attention mechanism, possess the capability of modeling long-range dependencies on a global scale and capturing intricate temporal dynamics. Consequently, leveraging the robust modeling capabilities of Transformers in conjunction with a global temporal attention mechanism to

 doi:10.20944/preprints202505.0838.v1

improve the robustness and accuracy of video anomaly detection represents a promising direction for further research [3].

Transformer-based approaches have recently driven notable progress in video anomaly detection. Many methods leverage Vision Transformers (ViT) to extract spatial features from video frames, while employing models such as LSTM or TCN to capture temporal dependencies. However, directly applying Transformers to long video sequences remains computationally demanding. Additionally, current methods often focus on short-term temporal segments, overlooking global temporal dependencies and thereby limiting the detection of cross-temporal anomalies. To overcome these limitations, the incorporation of a global temporal attention mechanism has emerged as a promising direction. While traditional methods like RNNs and LSTMs process sequences recursively and struggle with long-range dependencies, Transformers use self-attention to model global relationships more effectively. Nevertheless, applying global attention across all frames can introduce significant computational overhead due to redundant information in video data. Therefore, developing an efficient global temporal attention mechanism that selectively emphasizes key temporal segments while maintaining computational efficiency is essential. Such an approach can enhance the accuracy and robustness of anomaly detection systems, offering improved performance for intelligent surveillance, security, and traffic monitoring applications.

II. Related Work

The field of video anomaly detection has witnessed rapid advancement, particularly with the integration of deep learning techniques. Transformer-based models have recently emerged as a powerful alternative to traditional convolutional and recurrent neural networks due to their superior capacity for modeling long-range dependencies and capturing global temporal relationships. Paulraj and Vairavasundaram [4] proposed M2VAD, a weakly supervised framework leveraging multiview and multimodality input with Transformer backbones, laying a foundational path for leveraging attention mechanisms in complex surveillance scenarios. Similarly, Wang and Wu [5] introduced a Transformer with margin learning to tackle video anomaly detection in surveillance footage, demonstrating the advantages of attention-based structures in discerning subtle temporal irregularities.

Beyond video anomaly detection, broader applications of Transformer and attention mechanisms have significantly influenced model architecture design in adjacent areas. Hao et al. [6] applied a hybrid CNN-Transformer model for heart disease prediction, combining spatial and temporal modeling in a health context. Likewise, Xu et al. [7] advanced object detection in medical imagery using cross-scale attention and multi-layer feature fusion in YOLOv8, highlighting the transferability of attention-enhanced models across domains. Li et al. [8] further refined semantic segmentation with an optimized Unet architecture augmented by attention, reflecting the growing consensus on attention's utility in enhancing spatial-temporal comprehension.

In the medical and healthcare imaging domain, Wang et al. [9] developed a transfer learning-based method for breast cancer classification, while He et al. [10] extracted object-level insights in diagnostic imaging using RT-DETR. Although primarily situated in medical applications, the methodologies emphasize the critical role of deep feature extraction and attention modulation, which are directly relevant to anomaly detection's need for subtle pattern differentiation.

Additionally, gesture recognition and HCI research have leveraged deep learning for motion and behavior analysis. Duan [11] explored gesture key point detection using deep learning for intuitive user interfaces, and Shao et al. [12] integrated computer vision techniques to enhance gesture-driven interactions. These applications of motion understanding, though user-centric, share methodological underpinnings with anomaly detection—especially in terms of temporal dynamics and rare event localization.

Supporting this, Wang et al. [13] evaluated interface quality using deep models in HCI, suggesting a growing synergy between perceptual quality assessment and neural feature learning. In more technical pursuits, Liao et al. [14] fine-tuned T5 with knowledge graphs for complex NLP

 doi:10.20944/preprints202505.0838.v1

tasks, and Wu et al. [15] proposed adaptive attention-based enhancements for entity extraction using improved BERT models—both contributing to a broader understanding of Transformer versatility in learning contextualized representations.

Parallel to temporal modeling, sequence prediction and scheduling in edge computing, such as Zhan's LSTM-based micro-module scheduler [16], offers complementary insights into long-range forecasting, aligning conceptually with anomaly prediction in video streams. Lastly, Yan et al. [17] introduced a method to transform multidimensional time series into interpretable event sequences, providing foundational tools for anomaly explanation and model interpretability.

Together, these studies emphasize the central role of Transformers and attention mechanisms across various domains. Their collective contributions—ranging from surveillance to healthcare and HCI—form a robust theoretical and practical foundation for advancing video anomaly detection through enhanced temporal modeling, feature attention, and multimodal integration.

III. Method

In this study, we build a video anomaly detection model based on the Transformer structure and combined with the global temporal attention mechanism to improve the modeling ability of long-term dependencies [18]. The overall model architecture is shown in Figure 1.

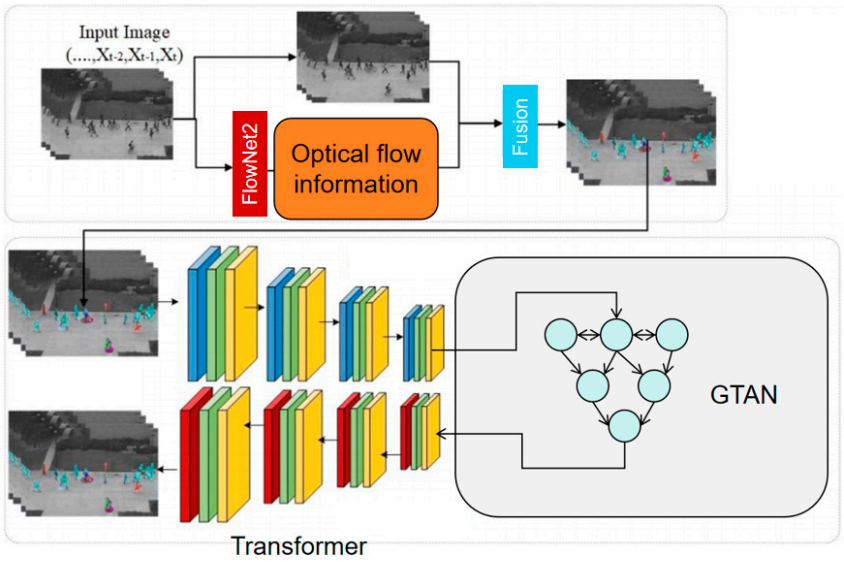

Figure 1. Model network architecture.

Let the video sequence be $X = \{x_1, x_2, ..., x_T\}$, where $x_t \in R^{H \times W \times C}$ represents the t-th frame, and H,W,C represents the height, width, and number of channels of the frame. First, we use a pre-trained CNN (such as ResNet or ViT) to extract the spatial features of the video frame and obtain the feature representation $F_t \in R^d$. For the entire video sequence, we can represent its spatial feature matrix as $F = \{F_1, F_2, ..., F_T\}$, where $F \in R^{T \times d}$. In order to further extract temporal information, we introduce the temporal feature encoding based on Transformer and combine it with the position encoding to maintain the order information of the frame sequence, that is:

$$F_t^{'} = F_t + PE(t)$$

where $PE(t)$ uses the standard sine-cosine position encoding format:

$$PE(t,2i) = \sin(\frac{t}{10000^{2i/d}})$$

 doi:10.20944/preprints202505.0838.v1

$$PE(t,2i+1)=\cos(\frac{t}{10000^{2i/d}})$$

Then, we use the Transformer's self-attention mechanism to model the global temporal dependency between frames. The core of the self-attention mechanism is to calculate the weighted relationship between the query, key, and value. The calculation formula is as follows:

$$Q=W_Q F\ , K=W_K F\ , K=W_V F$$

$$A=soft\max(\frac{QK^T}{\sqrt{d_k}})$$

$$Z=AV$$

Among them, W_Q , W_V , W_K is the trainable parameter matrix and d_k is the dimension of attention calculation. Since the computational complexity of the standard Transformer is $O(T^2\mathrm{d})$, the computational cost is high when the video sequence is long, so we further introduce the global temporal attention mechanism and add a cross-time window optimization strategy to the self-attention calculation. Specifically, we introduce a global temporal weight matrix $G\in R^{T\times T}$ to enhance the modeling ability of long-term dependencies:

$$A`=soft\max(\frac{QK^T+G}{\sqrt{d_k}})$$

Among them, G is generated by a learnable Global Temporal Awareness Network (GTAN):

$$G_{i,j}=\sigma(W_g(F_i-F_j)^2)$$

W_g is a trainable parameter and σ is a nonlinear activation function (such as ReLU). In this way, the model can adaptively adjust the attention weight according to the length of the time interval, so that the correlation information between long-distance frames can be preserved while avoiding the information loss caused by the short time window method.

$$F'_t=f_\theta(F_t)$$

$$S_t=\|F_t-F'_t\|_2$$

Among them, f_θ is the autoencoder and F'_t is the reconstructed feature. If S_t exceeds the set threshold, the frame is judged to be abnormal. Finally, we use statistical methods (such as mean deviation or quantile analysis) based on the distribution of abnormal scores of the entire video sequence to determine the final result of anomaly detection. Through the above method, we have improved the accuracy and robustness of video anomaly detection while ensuring computational efficiency.

IV. Experiment

A. Datasets

The ShanghaiTech Anomaly Detection Dataset is a widely used benchmark for video anomaly detection, designed for real-world surveillance scenarios. It comprises videos from 13 diverse

 doi:10.20944/preprints202505.0838.v1

scenes—including campuses, streets, and malls—featuring both normal activities (e.g., walking, cycling) and anomalies (e.g., fighting, falling, vehicles moving against traffic), making it a challenging testbed for model evaluation [19]. The dataset includes 330 training videos containing only normal events and 107 test videos with both normal and anomalous behavior, reflecting unsupervised detection settings. Videos are provided as frame sequences with stable frame rates and high resolution, and test videos are annotated frame-by-frame to support precise temporal evaluation. Despite its utility, the dataset presents challenges such as imbalanced anomaly distribution and scene-specific events, which hinder generalization. To address this, optical flow information is extracted and used in our experiments, as shown in Figure 2.

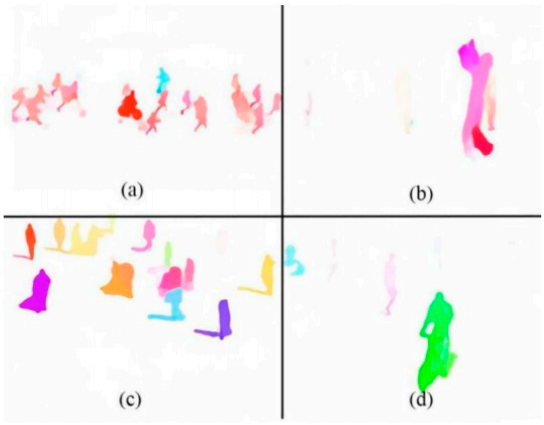

Figure 2. Partial data optical flow image information.

B. Experimental Results

To further validate the effectiveness of the proposed method, we conducted systematic comparative experiments against several state-of-the-art video anomaly detection approaches. These include Conv2D-AE, a convolutional autoencoder-based method; STAE (Spatio-Temporal Autoencoder), which integrates spatial and temporal feature learning; ConvLSTM-AE (Convolutional LSTM Autoencoder), which incorporates sequential modeling; TSC (Temporal Shift Clustering), a temporal self-supervised learning approach; MemAE (Memory-Augmented Autoencoder), which leverages a memory module for anomaly detection; MNAD-AE (Masked Normality Autoencoder), known for its adaptability to anomaly detection; and Stack-RNN, a recurrent neural network-based method. Each of these approaches represents a unique strategy in video anomaly detection, focusing on different aspects of feature extraction and anomaly identification [20].

To comprehensively evaluate the performance of these methods, we adopt the Area Under Curve (AUC) as the primary evaluation metric, which measures the overall detection capability of each model. Additionally, we compute the Equal Error Rate (EER) to assess the balance and robustness of the detection results [21]. By conducting experiments on the ShanghaiTech dataset, we can intuitively analyze the detection capability of the proposed method across different anomaly scenarios and verify the advantages of incorporating the Transformer architecture with a global temporal attention mechanism for anomaly detection tasks. The experimental results are shown in Table 1.

Table 1. Experimental results.

Model	AUC	EER
Conv2D-AE	72.4	31.2
STAE	74.8	29.5

 doi:10.20944/preprints202505.0838.v1

ConvLSTM-AE	76.1	28.3
TSC	75.6	28.8
MemAE	78.2	26.7
MNAD-AE	79.1	25.6
Stack-RNN	77.5	27.2
Ours	80.5	24.5

The experimental results show that different models vary in video anomaly detection performance based on AUC and EER metrics. Traditional autoencoder-based methods, such as Conv2D-AE and STAE, struggle with complex scenarios, achieving low AUCs (72.4, 74.8) and high EERs (31.2, 29.5) due to weak temporal modeling. ConvLSTM-AE and TSC improve performance slightly (AUCs of 76.1, 75.6; EERs of 28.3, 28.8) but fail to capture long-term dependencies. Memory-augmented models (MemAE, MNAD-AE) achieve better results (AUCs of 78.2, 79.1; EERs of 26.7, 25.6) but remain limited in global temporal modeling. Stack-RNN improves temporal modeling but suffers from the gradient vanishing issue, capping its AUC at 77.5 with an EER of 27.2. In contrast, the proposed Transformer-based method achieves the best performance (AUC 80.5, EER 24.5), demonstrating that global self-attention effectively captures long-range dependencies and enhances anomaly detection. Compared to MNAD-AE, it improves AUC by 1.4% and reduces EER by 1.1%, significantly reducing false positives and false negatives. Figure 3 presents the loss function decline for Conv3D-AE.

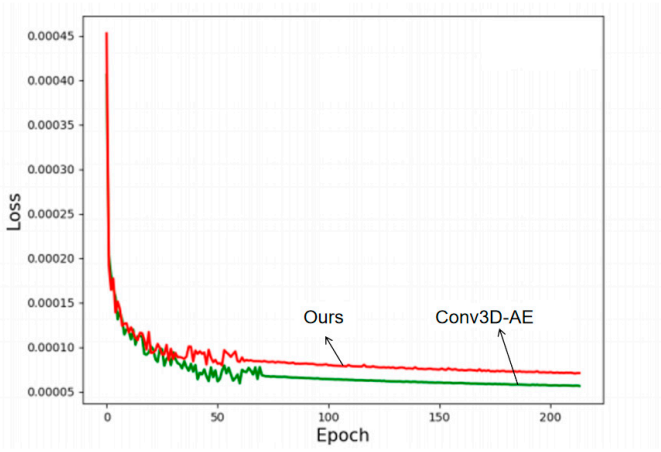

Figure 3. Loss function drop graph.

As shown in Figure 4, the proposed method (green) converges faster and more stably than Conv3D-AE (red). Conv3D-AE shows a higher initial loss, slower decline, and greater fluctuations, indicating less efficient optimization.

By around 50 epochs, our method stabilizes with minimal variation, reflecting stronger generalization. The lower final loss suggests better reconstruction and anomaly discrimination, thanks to the Transformer and global temporal attention mechanisms, which capture long-range dependencies more effectively than Conv3D-AE's local temporal modeling.

 doi:10.20944/preprints202505.0838.v1

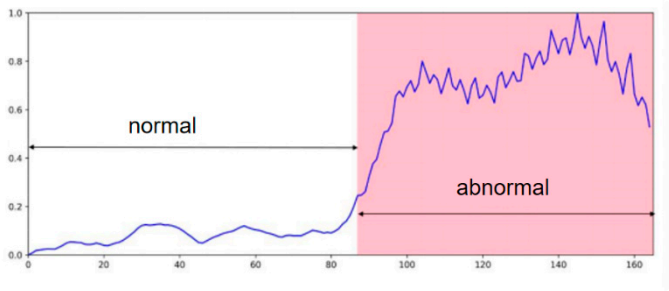

Figure 4. Visual analysis of normal and abnormal scores.

From the visualization analysis results in Figure 4, the proposed method can clearly distinguish between normal and abnormal events, and the anomaly score increases significantly when abnormal behavior occurs. In the normal area, the anomaly score always remains at a low level with a small fluctuation, indicating that the model has high stability in detecting normal behavior and is not easily disturbed by noise. When abnormal behavior begins to appear (around the 80th frame), the anomaly score gradually rises and maintains a high value in the abnormal area, reflecting that the model can effectively perceive the occurrence of abnormal events, and the trend of the anomaly score is highly consistent with the time of occurrence of the actual abnormal behavior.

In the abnormal region, the score maintains a high level with limited fluctuations, demonstrating the model's ability to consistently detect and track ongoing anomalies. This is critical, as many abnormal events span multiple frames. Peaks in the score may correspond to more severe anomalies, further highlighting the model's sensitivity to event intensity. Compared with the possible false positives or false negatives of traditional methods, this paper performs well in long-term dependency modeling through the Transformer structure and global temporal attention mechanism, so that the anomaly score curve can accurately reflect the evolution of abnormal behavior. Especially in the transition stage between normal and abnormal, the steady rise of the anomaly score indicates that the model can capture the potential abnormal trend in advance rather than abruptly rising when the anomaly occurs, which reflects the effectiveness and robustness of the model in time series modeling. Overall, this paper shows strong stability and accuracy in anomaly detection tasks and can provide reliable anomaly recognition capabilities for practical applications.

V. Conclusion

This paper proposes a Transformer-based video anomaly detection method with global temporal attention to address the limitations of existing approaches in capturing long-term dependencies and global temporal information. By leveraging self-attention mechanisms, the model effectively captures frame-wise correlations, while global temporal attention enhances feature extraction over extended time spans. Experimental results on the dataset demonstrate that the proposed approach outperforms state-of-the-art methods in AUC and EER, achieving superior detection accuracy, faster convergence, and lower false alarm rates. Compared to convolutional autoencoder-based and recurrent models, our method exhibits stronger temporal modeling capabilities, overcoming the limitations of local feature extraction. Additionally, it surpasses memory-augmented models like MemAE and MNAD-AE by adopting a more flexible temporal modeling strategy, improving the precision of anomaly detection. Despite these advancements, challenges remain, including high computational complexity and reliance on unsupervised learning assumptions. Future research could explore efficient sparse attention mechanisms, semi-supervised strategies, and applications in dynamic camera environments such as drone surveillance and dashcam analysis. Furthermore, integrating multimodal data and reinforcement learning could enhance adaptability, advancing intelligent surveillance for security monitoring, behavior analysis, and smart transportation.

 doi:10.20944/preprints202505.0838.v1

References

1. A. Hussain, W. Ullah, N. Khan, et al., "TDS-Net: Transformer enhanced dual-stream network for video anomaly detection", Proceedings of the 2024 Expert Systems with Applications Conference, pp. 124846, 2024.
2. M. H. Habeb, M. Salama and L. A. Elrefaei, "Enhancing video anomaly detection using a transformer spatiotemporal attention unsupervised framework for large datasets", Proceedings of the 2024 Algorithms Symposium, vol. 17, no. 7, pp. 286, 2024.
3. T. M. Tran, D. C. Bui, T. V. Nguyen, et al., "Transformer-based spatio-temporal unsupervised traffic anomaly detection in aerial videos", Proceedings of the 2024 IEEE Conference on Circuits and Systems for Video Technology, 2024.
4. S. Paulraj and S. Vairavasundaram, "M2VAD: multiview multimodality transformer-based weakly supervised video anomaly detection", Proceedings of the 2024 Image and Vision Computing Conference, vol. 148, pp. 105139, 2024.
5. D. Wang and K. Wu, "Anomaly detection in surveillance videos using Transformer with margin learning", Proceedings of the 2024 Multimedia Systems Conference, vol. 30, no. 5, pp. 250, 2024.
6. R. Hao, Y. Xiang, J. Du, Q. He, J. Hu and T. Xu, "A Hybrid CNN-Transformer Model for Heart Disease Prediction Using Life History Data", Proceedings of the 2025 arXiv Machine Learning and Healthcare Applications Conference, arXiv:2503.02124, 2025.
7. T. Xu, Y. Xiang, J. Du and H. Zhang, "Cross-Scale Attention and Multi-Layer Feature Fusion YOLOv8 for Skin Disease Target Detection in Medical Images", Proceedings of the 2025 Journal of Computer Technology and Software Conference, vol. 4, no. 2, 2025.
8. X. Li, Q. Lu, Y. Li, M. Li and Y. Qi, "Optimized Unet with Attention Mechanism for Multi-Scale Semantic Segmentation", Proceedings of the 2025 arXiv Computer Vision and AI Conference, arXiv:2502.03813, 2025.
9. W. Wang, Y. Li, X. Yan, M. Xiao and M. Gao, "Breast cancer image classification method based on deep transfer learning," Proceedings of the International Conference on Image Processing, Machine Learning and Pattern Recognition, pp. 190-197, 2024.
10. W. He, Y. Zhang, T. Xu, T. An, Y. Liang and B. Zhang, "Object Detection for Medical Image Analysis: Insights from the RT-DETR Model", Proceedings of the 2025 arXiv Medical Imaging and Deep Learning Conference, arXiv:2501.16469, 2025.
11. S. Duan, "Deep Learning-Based Gesture Key Point Detection for Human-Computer Interaction Applications", Proceedings of the 2025 Transactions on Computational and Scientific Methods Conference, vol. 5, no. 1, 2025.
12. F. Shao, T. Zhang, S. Gao, Q. Sun and L. Yang, "Computer Vision-Driven Gesture Recognition: Toward Natural and Intuitive Human-Computer Interaction", Proceedings of the 2024 arXiv Computer Vision and HCI Conference, arXiv:2412.18321, 2024.
13. S. Wang, R. Zhang, J. Du, R. Hao and J. Hu, "A Deep Learning Approach to Interface Color Quality Assessment in HCI", Proceedings of the 2025 arXiv Human-Computer Interaction Conference, arXiv:2502.09914, 2025.
14. X. Liao, B. Zhu, J. He, G. Liu, H. Zheng and J. Gao, "A Fine-Tuning Approach for T5 Using Knowledge Graphs to Address Complex Tasks", Proceedings of the 2025 arXiv NLP and Knowledge Engineering Conference, arXiv:2502.16484, 2025.
15. L. Wu, J. Gao, X. Liao, H. Zheng, J. Hu and R. Bao, "Adaptive Attention and Feature Embedding for Enhanced Entity Extraction Using an Improved BERT Model", Proceedings of the 2025 AI for NLP Conference, 2025.
16. J. Zhan, "Elastic Scheduling of Micro-Modules in Edge Computing Based on LSTM Prediction", Proceedings of the 2025 Journal of Computer Technology and Software Conference, vol. 4, no. 2, 2025.
17. X. Yan, Y. Jiang, W. Liu, D. Yi, and J. Wei, "Transforming Multidimensional Time Series into Interpretable Event Sequences for Advanced Data Mining", arXiv preprint, arXiv:2409.14327, 2024.
18. S. M. Rahimpour, M. Kazemi, P. Moallem, et al., "Video anomaly detection using transformers and ensemble of convolutional auto-encoders", Proceedings of the 2024 Computers and Electrical Engineering Conference, vol. 120, pp. 109879, 2024.

 doi:10.20944/preprints202505.0838.v1

19. C. Tao, C. Wang, S. Lin, et al., "Feature reconstruction with disruption for unsupervised video anomaly detection", Proceedings of the 2024 IEEE Transactions on Multimedia Conference, 2024.
20. K. Biradar, D. K. Tyagi, R. B. Battula and J. Y. Jung, "Robust Anomaly Detection through Transformer-Encoded Feature Diversity Learning", Proceedings of the 2024 Asian Conference on Computer Vision, pp. 115–128, 2024.
21. H. Kim, C. H. Lee and C. Hong, "VATMAN: Video Anomaly Transformer for Monitoring Accidents and Nefariousness", Proceedings of the 2024 IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–7, 2024.