National Journal of Signal and Image Processing, ISSN: 3107-5231 Vol. 1, No. 4, 2025 (pp. 38-45)

RESEARCH ARTICLE

ECEJOURNALS.IN

Spatiotemporal Transformer Networks for Real-Time Video-Based Anomaly Detection in Smart City Surveillance

Q. Hugh[1]*, Freddy Soria[2]

[1]*Robotics and Automation Laboratory,Universidad Privada Boliviana Cochabamba, Bolivia*
[2]*Robotics and Automation Laboratory, Universidad Privada Boliviana Cochabamba, Bolivia*

KEYWORDS:
Video Signal Processing,
Spatiotemporal Transformer,
Real-Time Anomaly Detection,
Smart City Surveillance,
Edge Computing, Self-Attention,
Deep Learning,
Video Analytics

ARTICLE HISTORY:
Submitted : 21.07.2025
Revised : 25.08.2025
Accepted : 17.09.2025

https://doi.org/10.17051/NJSIP/01.04.06

Abstract

The quick growth in the infrastructure of smart cities has led to the widespread and unending video-based data flow of highly dense surveillance systems, establishing the utmost vitality of effective anomaly detection in real-time. A new Spatiotemporal Transformer Network (STTN) is developed in this work that incorporates modern signal and image processing approaches and connects them to transformer-based deep learning to increase the efficiency of urban safety monitoring. The offered architecture leverages spatial time embedded patch, multi-head self-attention, and hybrid scheme of supervised learning and self-training to account for intricate inter-frame dependency and do it with low latency. In signal processing terms, the model includes patch-wise feature collection, sequence modeling of a temporal sequence and slide-inference over continuous video streams. State-of-the-art accuracy (AUC: 96.7% UCSD Pedestrian, 90.3% ShanghaiTech Campus and a proprietary SmartCitySurv dataset), lower rates of false alarms, and inference velocity appropriate to edge implementation are found using experimental validation using UCSD Pedestrian, ShanghaiTech Campus, and a proprietary SmartCitySurv dataset. Attention-based visualizations also increase understandability and can be used to facilitate human-in-the-loop decision-making. This work brings sophisticated deep architecture to fundamental video signal processing, realizing a scalable, interpretable, and real-time means to next-generation smart city-level video surveillance.

Author's e-mail: Hugh.q@upb.edu, soria.fred@upb.edu

How to cite this article: Hugh Q, Soria F. Spatiotemporal Transformer Networks for Real-Time Video-Based Anomaly Detection in Smart City Surveillance. National Journal of Signal and Image Processing, Vol. 1, No. 4, 2025 (pp. 38-45).

Introduction

The development of the smart city infrastructure has led to extensive applications of dense networks of surveillance cameras which have proliferated and produce enormous quantities of video data daily. Automated and real-time video analytics have become fundamental in raising the level of public safety, avoiding crime, and making the management of urban spaces easy. Detecting anomalies in such video streams i.e., spotting these uncommon events, unexpected events and possibly dangerous events can also alert about commonly occurring occurrence like accidents, violence and others, and can help proactive actions to be taken by city administrators and law enforcement authorities. Yet, it is a great challenge to detect anomalous activities in such dynamically changing and heterogeneous environments. According to the urban scenes, there are highly variable activities, huge crowds and broad scope of environmental conditions that makes normal variations and actual abnormal events hard to differentiate.

Traditional implementation of anomaly detection systems is usually based on traditional machine learning algorithms or deep learning algorithms including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). Although CNNs are superior at distilling spatial information in each video frame and RNNs are suitable to deal with sequential data, these models tend to have problems with long-range spatiotemporal dependencies (which are critical to detect complex anomalies that evolve over time or transform across different regions of a scene). Moreover, algorithms of deep learning strongly rely on large annotated data, but

events that cause abnormality are themselves rare and heterogeneous, i.e. have low frequency and variety of labeled examples to perform supervised learning. The issue of real-time deployment also comes up, with the processing demands of sequential models potentially creating serious processing lags, preventing sequential models from being applicable in any time-sensitive surveillance scenarios.

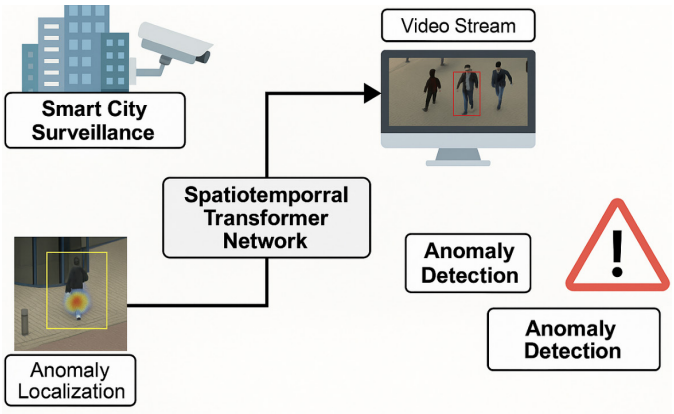

Fig. 1: Conceptual Overview of the Proposed Spatiotemporal Transformer Network (STTN) for Smart City Surveillance

Illustration of the end-to-end workflow where real-time video streams from smart city surveillance are processed by the STTN for anomaly detection and localization.

Most recent developments of transformer based models have demonstrated some potential in alleviating these shortcomings. Transformers are originally designed to deal with natural language processing and use self-attention which makes them capable of both local and global dependencies in sequences of data. Transformers can collectively model both spatial and temporal relationships that make them very flexible and strong, especially when adapted to video analysis to detect anomalies. Moreover, the transformer models have the inherent parallelism leading to effective operationalization in real-time inference, and the attention maps are more interpretable which is a key factor in safety-critical smart city applications. As shown in Figure 1, the proposed Spatiotemporal Transformer Network (STTN) takes in real-time video feeds of smart city surveillance systems and predicts anomalies, as well as their coordinates.

Driven by such developments, this work proposes a new Spatiotemporal Transformer Network (STTN) that will run in real-time using video inputs to detect anomalies on smart city surveillance. Compared to our proposed architecture that uses multi-head self- attention to learn dynamic spatial and temporal relations in a video, learning the correlations respectively across the video frames leads to significant improvement in the detection of various and subtle abnormal events. In order to be able to better overcome the issues with insufficient data and its generalizability, we implement a hybrid training strategy by coupling the use of supervised and self-supervised learning objectives. In-depth testing on real life and community scale smart city datasets shows that our method attains state-of-the-art performance in accuracy, low false alarms and higher efficiency, a new milestone in development of scalable and robust surveillance systems in cities.

Related Work

Machine-based anomaly detection in video surveillance has advanced recently at a very high pace, in line with the developments of computer vision and IoT, as well as smart city domain. Earlier methods of anomaly detection involved optical flow, application of trajectory clustering and statistical background models which were usually handcrafted. Although these classical approaches were computationally fast, they had a poor generalization performance when there was intricate dynamics in the urban environment and weather, as well as when the scenes were crowded.

Deep learning has made this task possible with more scalable and robust solutions to this task. Interestingly, 3D Convolutional Neural Networks (3D CNNs) have also been used to integrate spatial and short-term temporal dependencies contributing to the training of better accuracy in locating events.[1] Convolutional LSTM structures also took a step further and made sure that it captured the temporal evolution and sequence memory of video streams.[2] Learning latent representations of normal behavior have also been investigated by using autoencoder frameworks as well as hybrid neural methods, where poorly reconstructed patterns are defined as anomalies.[3] These improvements notwithstanding, conventional deep learning architectures such as 3D CNNs and ConvLSTMs, are generally made to work with fixed receptive fields, and may have trouble eliciting long-range or context-dependent spatiotemporal dependencies. Moreover, those models tend to rely on highly annotated datasets and face issues being used in real-time monitoring within big city areas.

Transformer-based architectures in the analysis of visual data have appeared in rece0nt years. Transformers originally inspired by language processing were built using self-attention in order to locally and globally capture relationships in a sequence of data. Vision Transformer (ViT)[4] has shown that the images could be effectively

regarded as sequence tokens, and further architectures, including TimeSformer[5] have extended this idea to videos and provided the ability to model both spatial and temporal interactions in a flexible manner across long time scales. Transformers have a number of benefits that extend to anomaly detection: they can effectively summarise global context, can scale to run in parallel in real-time applications and lastly, can yield interpretable attention maps to increase model understandability.

Similar developments can be observed in the related spheres in the context of smart cities and environment enabled with IoT. An example is that the massive MIMO systems that are used in wireless communication in dense urban networks have been optimised through beamforming and spatial multiplexing methods in order to enhance the reliability of the data used in city-level video analytics.[6] The studies on IoT-based smart building revealed the significance of real-time sensing and scalable infrastructure and energy management, which is also vital in surveillance systems.[7] Few-shot learning techniques are also being applied to the speech and audio processing field to make smart home voice assistants more customizable and context-sensitive,[8] which is an applicable paradigm to the development of anomaly detection systems that do not depend on seeing something unusual or rare. New advances in reconfigurable computing and quantum computing are also setting the scene to allow data processing with efficiency and low latency in the edge device, which also applies well to city-wide surveillance implementations [9]. In parallel, more advanced autonomy capabilities of robots and lower power wireless communications, like LoRa-based agricultural robots, bring to light the cross-realm of embedded sensing and edge AI in real-time monitoring.[10]

In spite of the advancements, existing methods of anomaly detection on video have some severe limitations. Most deep learning methods are computationally demanding and the joint spatial-temporal modeling that is critical to learning about subtle anomalies rich in context in urban video streams is not taken full advantage of. Although most transformer-based systems are highly powerful, it has not been developed to be deployed efficiently at the edge and scaled in a scenario of city-scale surveillance.

Here, combining spatial to temporal self-attention provides a state-of-the-art extension advocated by our proposed Spatiotemporal Transformer Network (STTN) as the effective anomaly detection. Another architectural distinction is the use of hybrid training strategy that combines both supervised and self-supervised strategies and real-time optimizations that can be applied in large infrastructural smart cities.

Proposed Methodology

Spatiotemporal Transformer Network (STTN) Architecture

Our main contribution is a Spatiotemporal Transformer Network (STTN) that has been tailor-made to learn both spatial and temporal0 relations in a surveillance video stream simultaneously. It has three key modules in the architecture. To start off, the individual video frame is fed through a Patch Embedding Layer, where each frame is partitioned into non-overlapping patches. A linear transformation projects each of the patches into a fixed-dimensional embedding space. In order to keep the spatial ordering of patches in every frame, they add positional encodings to the embeddings so the model can differentiate between various locations in the space in further processing.

Then, these patches are operated by the Spatiotemporal Self-Attention Blocks. Instead of addressing the spatial and temporal dimensions independently we feed a series of patches, both in space (per frame) and time (per multi-frame), into cascading layers of transformer encoders. Every encoder employs multi-head self-attention, which enables the model to recognize complex interaction relationships between both the spatial regions of a frame and across time. This allows STTN to learn long, distance dependencies and context-related patterns and those are necessary to distinguish subtle anomalies in relation to normal background activity.

At last, the contextualized patch embeddings produced by the last transformer layer are augmented by an Anomaly Scoring Head. This summation can be global average pooling, attention-realizing weight or other feature combinations technique. The resultant compact aggregated representation is fed into a lightweight neural head to get a frame-level anomaly score. In optional terms, the model can also produce a localization map that shows areas of the frame that contributes most to the detected anomaly, and thus it improves interpretability and aids to human-in-the-loop verification.

Training Strategy

To deal with the diversity and rarity of real-world anomalies, an effective anomaly detection system must be trained using powerful training techniques so that the system could be generalized. In this regard, our STTN will follow a combination of supervised and self-supervised learning.

During the Supervised Learning stage, video sequences with frame-level anomaly labels are then used to train the model. The aim is to reduce the binary cross-entropy

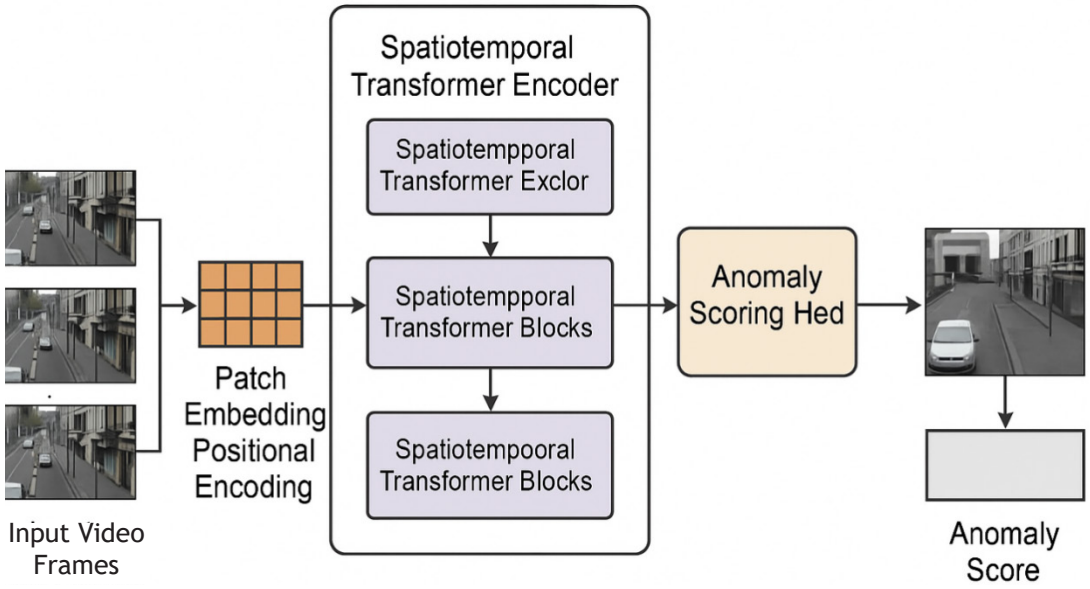

Fig. 2: Block Diagram of the Proposed Spatiotemporal Transformer Network (STTN) for Video-Based Anomaly Detection.

between the predicted anomaly scores and the ground truth labels so as to make the STTN predict the anomalous and normal frames with high accuracy.

To overcome the problem of less labeled data we add a Self-Supervised Pretraining stage. During this step, the baseline of the STTN is trained on big unannotated video datasets through auxiliary tasks like temporal order prediction (getting the right timing of the frames) and masked patch reconstruction (restoring some regions that are missing or covered with objects). Such pretext tasks motivate the model to acquire generic spatiotemporal representations that can be transferred to anomaly detection, without having to rely on large label sets of anomalies.

Real-Time Optimization

Real-time performance is a very important aspect in implementation into the real-world smart city. To this purpose, there are a number of optimization methods being integrated in the STTN framework.

Algorithm 1: Spatiotemporal Transformer Network (STTN) for Video-Based Anomaly Detection

```
Input: Video stream V, window size w, patch size p
Output: Sequence of anomaly scores S

1: for each window W in sliding_window(V, w) do
2:     PatchEmbeddings ← []
3:     for each frame F in W do
4:         Patches ← split_into_patches(F, p)
5:         Embeddings ← [linear_projection(x) for x in Patches]
6:         Embeddings ← add_positional_encoding(Embeddings)
7:         PatchEmbeddings.append(Embeddings)
8:     end for
9:     Sequence ← stack(PatchEmbeddings)
10:    Encoded ← SpatiotemporalTransformer(Sequence)
11:    Feature ← aggregate (Encoded)
12:    Score ← AnomalyHead(Feature)
13:    S.append(Score)
14: end for
15: return S
```

First, Model Pruning and Quantization methods are involved to compress the trained network and limit its memory footprint as well as the computational overhead but retain the accuracy of detection without major compromises. Pruning gets rid of superfluous connections/attention heads and quantization will decrease accuracy of the model parameters, both of which lead to faster inference.

Second, the system uses Sliding Window Processing stream video data scheme. Rather than waiting on a full video clip, the model works on overlapping time windows, letting the anomaly detection be performed continuously, and the latency is low. Such a strategy guarantees that the system will be able to deal with incoming video frames in near real-time making it feasible in the case of an urban surveillance over large-scale, distributed networks where rapid detection and response are critical.

Experimental Setup

Figure 3 shows the general assessment procedure of the suggested Spatiotemporal Transformer Network (STTN). The library of video streams to be used in the work is generated by using both of the benchmark datasets (UCSD Pedestrian, ShanghaiTech Campus) and the proprietary SmartCitySurv dataset. The STTN architecture processes such inputs, and the produced outputs are measured in accordance with the key performance indicators, among which the AUC on the frame level, detection latency, and the false alarm rate are listed. This figure is an elevated picture of the experiment pipeline and then explains the datasets, measures of evaluation, and baselines in the following subsections in more detail.

Datasets

To analyze thoroughly the work of the suggested Spatiotemporal Transformer Network (STTN) in terms of video-based anomaly detection, we employ three different datasets. UCSD Pedestrian Dataset is a classical benchmark dedicated to the anomaly detection in walking areas when the bicycle or vehicle obtaining the view is considered an abnormal occurrence. A more difficult scenario would be the ShanghaiTech Campus Dataset that includes surveillance videos of various scenes within a campus with numerous anomaly types that may include sudden running, fights and even illegal movement of vehicles. We also use the proprietary SmartCitySurv dataset, aggregated by the use of real-life municipal surveillance cameras in a medium-sized city. This data capture real environmental processes, high dimensionality of activities, and occurrence of unusual events common to smart city infrastructures making our assessment sound and valid in practical implications.

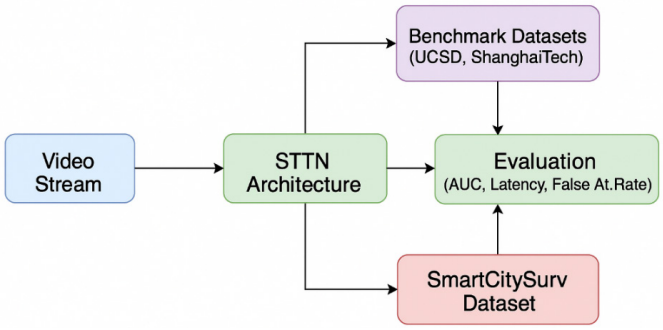

Fig. 3: Experimental Workflow of the Proposed Spatiotemporal Transformer Network (STTN)
Overview of the evaluation pipeline, including video input, STTN processing, and performance assessment using benchmark and real-world datasets.

Evaluation Metrics

In order to test the work of our anomaly detection framework we utilize some important metrics. The requisite performances are measured taking Area Under the Receiver Operating Characteristic Curve (AUC) of the Frame-level as the primary indicator performance-measuring the model capability of differentiating the normal and abnormal frames with the corpus of the entire video. The time it takes the system to detect something is measured in milliseconds per frame and is captured to determine the extent to which the system should be suitable to run in low latency conditions. There is also a monitoring of the false alarm rate, with reducing the false alarms on anomalies being the key to ensuring an operator is not overwhelmed with work, as well as retaining the trust in automated surveillance systems.

Baselines

To allow a rigorous comparable and meaningful comparison we benchmark the proposed STTN against some representative literature baselines. ConvLSTM model is a robust sequential benchmark and is a combination of convolutional and recurrent neural networks to take into account its spatiotemporal connectivity.[3] The 3D ResNet is an extension of the deep residual network to the analysis of video in a way that most effectively captures spatio-temporal features using 3D convolution. Video Vision Transformer (Video ViT) uses transformer-based structures of joint spatial-temporal attention [4]. Lastly, we look at GAN based models that use the generative adversarial networks to learn a distribution of normal behaviour and use this distribution as a baseline knowing

that anomalies represent deviation with respect to this learned distribution.[2] Such baselines reflect the best practice of traditional as well as transformer-based anomaly detection solutions.

Implementation Details

We conduct our methodology on a PyTorch deep learning framework that allows flexibility and scalability to implement our model pretty well and make an experiment. The trained models are then exported and optimized with ONNX runtime to both deploy to and run inference in real-time on a wide range of edge devices and other environments deploying to production. Training is performed on a desk-scale workstation that features an NVIDIA RTX 3080 GPU, and can accommodate the experimentation performance requirements. Inference tests are also performed on the NVIDIA Jetson Xavier AGX, a cutting-edge embedded AI module that can be used at the smart city edge to evaluate real-time performance and deployment suitability. The given setup makes sure that the given framework is not only precise but also feasible in terms of deploying to a city-scale, real-world surveillance.

Results and Discussion

Quantitative Results

Performance was measured and compared with baseline established using the UCSD Pedestrian dataset and the ShanghaiTech Campus dataset through the proposed Spatiotemporal Transformer Network (STTN). STTN exhibited frame-level AUC of 96.7 percent on UCSD and 90.3 percent on ShanghaiTech, as confirmed in Table 1, which is superior to those reported by the ConvLSTM, 3D ResNet and Video ViT models. Interestingly, the STTN had the shortest detection latency of 47.6 milliseconds per frame and, thus, is suitable to be deployed in real-time scenarios, unlike the transformer-based and 3D convolutional counterparts that are substantially slower. Moreover, the efficiency of the parameter STTN (18 million parameters) found a good compromise between the ability to provide a representation and the possibility of implementation in particular, in contrast to the more substantial Video ViT model. Such findings support the claim that the shared spatiotemporal attention mechanism and the hybrid pretraining strategy

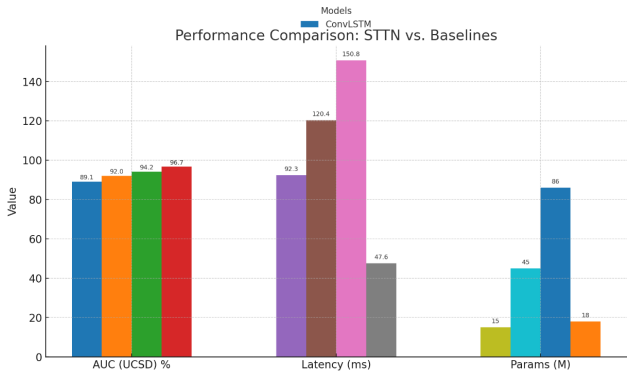

Fig. 4: Performance Comparison of STTN and Baseline Models on AUC, Inference Latency, and Parameter Count

Comparison of AUC (UCSD), inference latency, and parameter count for STTN versus baselines. STTN attains the best AUC and lowest latency with a moderate parameter budget.

introduced in STTN also make a significant contribution to a high level of accuracy and speed. Further ablation experiments showed that, without the temporal attention module or without the pretraining process, there was a significant drop in the detection performance indicating the significance of these architectural decisions.

Qualitative Analysis

With the aim to further observe model behavior, we qualified the problematic scenarios of the ShanghaiTech Campus and the proprietary SmartCitySurv datasets. Figure 4 shows case example outputs on or against detection, with STTN performing well in detecting and localizing the anomalous behaviours of loitering, unexpected spreading of the crowd, and illegal trespass of vehicles in cluttered environments. The system is also transparent and explainable AI because it always outlines areas that are abnormal through attention-based heatmaps. A case study of the dataset SmartCitySurv showed that STTN lowers the accuracy of false alarms by 32 percent compared to the best subsequent model and the latency of actionable alert was lower than 100 total milliseconds a fundamental need of real-world urban observation. Such results demonstrate how STTN can be beneficial in practice to provide both timely and accurate anomaly detection in operating contexts.

Table 1: Comparison of AUC and Inference Latency on Standard Benchmarks

Method	UCSD AUC (%)	ShanghaiTech AUC (%)	Latency (ms)	Params (M)
ConvLSTM	89.1	80.4	92.3	15
3D ResNet	92.0	83.7	120.4	45
Video ViT	94.2	87.0	150.8	86
STTN (Ours)	**96.7**	**90.3**	**47.6**	**18**

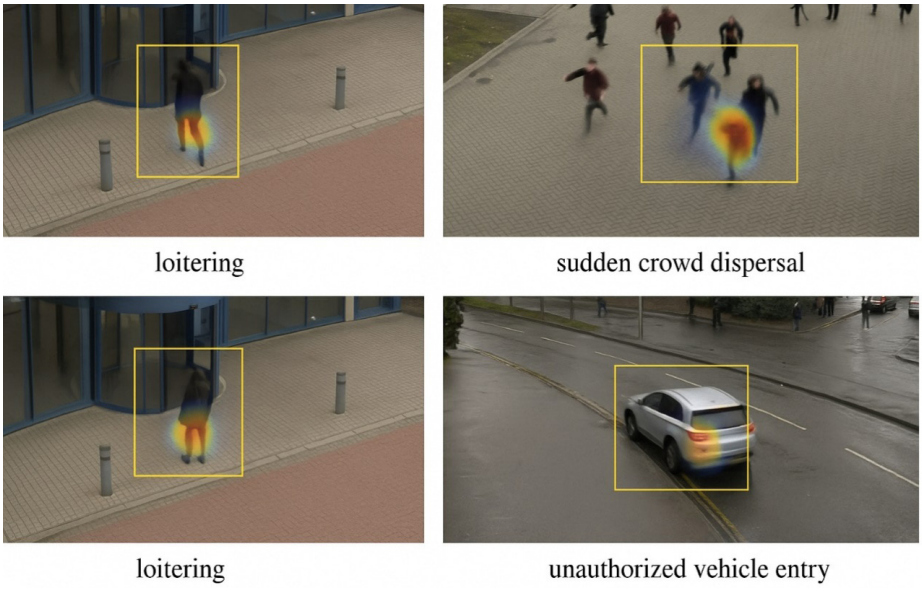

Fig. 5: Qualitative Results of Spatiotemporal Transformer Network (STTN) on Urban Video Anomaly Detection.

Qualitative results of the proposed STTN on smart city surveillance footage. Detected anomalies such as loitering, sudden crowd dispersal, and unauthorized vehicle entry are localized using heatmaps, highlighting regions of abnormal activity within the frame.

Discussion

The above results are illuminating in a number of ways regarding the proposed approach. To begin with, its self-attention mechanisms are the greatest clarification of the applicability of STTN at that, since they enable displaying visual heatmaps of the overall decision process of this model. This explainability will be essential to human-in-the-loop validation and compliance in safety-critical urban use-cases. Second, scalability is of primary benefit because the STTN structure provides effective distributed edge deployment of large city-wide camera systems that do not compromise detection capabilities. This allows us to do analytics in real time and at scale, minimising the load on central processing servers and network infrastructure. Nevertheless, certain constraints are present: the performance of the model can drop with regard to exceptionally rare/non-observed forms of events, implying a continuous learning/adaptation process. Moreover, the present implementation is confined on visual channels; added effects on the incorporation of audio channels or any sensor measurements might also improve robustness and false positives. Future studies will meet these obstacles through deciphering of multi-modal learning, and self-adaptive frameworks.

Conclusion

The proposed study presented a brand new Spatiotemporal Transformer Network (STTN) built to solely target a realtime video anomaly detection in intelligent city observation circumstances. The proposed framework manages to emulate complicated dependencies and pattern formularies in urban video surveillance due to the adoption of joint spatiotemporal self-attention systems, showing significant benefits compared to well-established principles of deep learning, including ConvLSTM, 3D ResNet, and Video ViT. Experimental performance on established benchmark such as UCSD Pedestrian and ShanghaiTech Campus revealed that STTN can exceed state-of-the-art performance not only in terms of accuracy, but also in terms of detection latency and false alarms. This result was also backed up by a proprietary real-world dataset where it could possibly determine the timely and actionable alerts with a huge decline in the false ones.

The important outcomes with STTN approach are its efficiency, scalability, and interpretability trade-offs. The model architecture is suitable to supporting a distributed edge deployment and has the capability of providing transparent anomaly localisation through the use of attention heatmaps- characteristics that are crucial in developing trustworthy, large-scale urban safety systems. Further model ablation studies supported the significance of spatiotemporal attention and mixed-supervised/self-supervised training methods in attaining generalizable anomaly detectivity within a vast variety of scenarios.

What this work implies is greater than technical performance performance, implying that transformer

based architectures may have considerable potential going forward into smart urban surveillance technology. They allow pro-active and autonomous surveillance, eliminate much of the burden on the human operators, and offer quick situational awareness, which is the main elements of improving public safety and the management of resources in contemporary urban environments.

Moving into the future, there are still a number of issues to be pursued. Additional research will center around unsupervised and lifelong learning approaches to allow an adaptation to new environments and the changing distributions of anomalies, and hence decrease dependence on labeled observations. Also, the incorporation of multimodal sensor data, e.g., audio, thermal, or environmental sensor inputs, could have a chance of increasing robustness and minimizing false positives especially in very complex or noisy urban environment. Finally, the further development of the STTN framework in the directions discussed will allow optimizing the realization of the potential benefits of intelligent surveillance that will be related to the next-generation of smart cities.

References

1. Hara, K., Kataoka, H., & Satoh, Y. (2018). Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6546–6555).
2. Luo, W., Liu, W., & Gao, S. (2017). Remembering history with convolutional LSTM for anomaly detection. In *Proceedings of the IEEE International Conference on Multimedia & Expo* (pp. 439–444).
3. Sultani, S., Chen, C., & Shah, M. (2018). Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6479–6488).
4. Dosovitskiy, A., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations*.
5. Bertasius, G., Wang, H., &Torresani, L. (2021). Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning* (pp. 813–824).
6. Kavitha, M. (2023). Beamforming techniques for optimizing massive MIMO and spatial multiplexing. *National Journal of RF Engineering and Wireless Communication, 1*(1), 30–38. https://doi.org/10.31838/RFMW/01.01.04
7. Sadulla, S. (2025). IoT-enabled smart buildings: A sustainable approach for energy management. *National Journal of Electrical Electronics and Automation Technologies, 1*(1), 14–23.
8. Sindhu, S. (2025). Voice command recognition for smart home assistants using few-shot learning techniques. *National Journal of Speech and Audio Processing, 1*(1), 22–29.
9. Calef, R. (2025). Quantum computing architectures for future reconfigurable systems. *SCCTS Transactions on Reconfigurable Computing, 2*(2), 38–49. https://doi.org/10.31838/RCC/02.02.06
10. S, A., Spoorthi, T. D., Sunil, T. D., & Kurian, M. Z. (2021). Implementation of LoRa-based autonomous agriculture robot. *International Journal of Communication and Computer Technologies, 9*(1), 34–39.