

FUSION OF HUMAN EXPERTISE AND AI FOR PRIVACY-PRESERVING CROWD MONITORING

Ashwini Manohar Khade¹, Mr. Anil B. Desai²

¹Department of Computer Science and Engineering Ashokrao Mane Group of Institution Vathar Tarf Vadagaon, Kolhapur, India, ashwinikhade411@gmail.com

²Department of Computer Science and Engineering Ashokrao Mane Group of Institution Vathar Tarf Vadagaon, Kolhapur, India, abd@amgoi.edu.in

ABSTRACT

Accurate and prompt oversight of human behaviours is an essential factor for threat management within intelligent surveillance technologies. This article proposes a Privacy-Preserving Spatio-Temporal Graph and Transformer Network that handles the intertwined behavioural patterns in crowd data. For the surveillance environment, the dynamic network is assumed whereas light-weight cryptographic and attention-based concepts are introduced for enhancement of feature capturing and privacy shielding. Absence from the scene is predicted by combining the knowledge gained about behavioural habits and movement changes from activity data history. Based on experimental tests, the novel method is proved to be an effective way of handling complicated crowd situations when compared to other machine learning and deep learning techniques.

KEYWORDS: Suspicious Activity Detection, Crowd Surveillance, Spatio-Temporal Analysis, Graph Attention Network (GAT), Lightweight Encryption, Real-Time Tracking, Intelligent Surveillance Systems

INTRODUCTION

Rapid urbanization and the popularity of smart city concepts have brought the importance of public safety and crowd management as one of the top issues in the modern world to the forefront. With the ever-growing number of people gathering in public places, the probability of accidents happening, committing crimes, safety of the structure failing, and security being compromised also increases. In these situations, it is highly critical to develop suspicious activity detection as a part of the smart surveillance system. In general crowd monitoring methods that include mainly using a statistical model and simple machine learning approaches have been proven to be inadequate as they were never intended for complex modelling of human behaviours.

Although more advanced deep learning techniques, including early Convolutional Neural Networks (Convolutional Neural Networks/CNN) and autoencoders, fail to provide a perfect model of both spatial and temporal relationships at the same time, the surveillance data sets are naturally characterized by spatiotemporal factors; these data depend on the time an action happened as well as spatial dependencies in a crowd network. An advanced solution should be introduced to address these challenges and capture such complex interrelationships of human behaviour from videos with the help of an appropriate model.

Emergence of Graph Attention Networks (GATs) has given new dimension to the way spatial dependencies between people's individual behaviours can be visualized and modelled with the aid of graph structures. In these structures, the nodes denote the individuals while edges denote their ties. Spatio-Temporal Graph Networks like TimeSformer, that use a combination of temporal learning based on transformers and graph networks, can extract information relating both time and space from surveillance data. Although these models can model complicated patterns in the data, they are not still equipped to process raw video inputs without revealing the identities of individuals. This way resulting in major privacy and ethical concerns. The privacy concerns arise mainly because the models work on raw video feeds where the identity of individuals may be revealed.

The paper addresses the drawbacks mentioned above by proposing a new Privacy-Preserving Spatio-Temporal Graph and Transformer Network for real-time crowd behaviour monitoring. Instead of working with video feeds that can identify persons the model uses a combination of selective instance segmentation and ChaCha20 encryption principles - light cryptographic procedures - to its design. This is aimed at enabling the model not to only safeguard individuals' identities at pixel level but also be able to identify complex behavioural interdependencies between features.

Within the system described above initially the surveillance video collection process will be performed. Then, the information will be pre-processed to distinguish human regions for a frame-by-frame basis. Once the instance segmentation of data is conducted, the segmented regions are encrypted with ChaCha20 partially. Further, after the person identification using the graph structure and describing dynamics of the interactions between persons the data is ready for modelling by the model. In other words, the model is composed of two-streams - one for ViT and TimeSformer transformers and another for Spatio-Temporal Graph Attention Networks (ST-GATs) - it will be capable of recognizing the abnormal behavioural patterns from the data. The advantage is this system allows an accurate and reliable prediction without any restrictions in any setting. Overall, the paper aims to develop an effective, privacy-preserved surveillance system design, scalable, and precise. The approach advocated is the one that will contribute to the

development of intelligent surveillance systems and the decision-making processes in urban spaces where people are at risk and lives are at stake.

1.1 Key Contributions

- Proposed a novel Privacy-Preserving Spatio-Temporal Graph and Transformer Network for real-time crowd behaviour monitoring.
- Incorporated lightweight cryptographic concepts such as ChaCha20 encryption to enhance regional privacy protection.
- Effectively modelled spatial and temporal dependencies using a combination of graph attention frameworks and vision transformers.
- Improved detection accuracy compared to traditional machine learning and deep learning models that lack advanced temporal context.
- Developed a scalable and efficient framework suitable for intelligent surveillance systems and smart city applications.

2. Related Work

In addition to being an essential part of intelligent surveillance systems, crowd monitoring and recognizing suspicious activities are also research areas that have gained attention recently as cities become more crowded and people's concerns about public safety grow. Besides, machine learning and deep learning are two of the various methods that researchers are using to enhance the detection process' accuracy by reducing error rate. Spatio-Temporal Graph Neural Networks (ST-GNNs) and Graph Attention Networks (GATs) have attracted great interest lately since they manage very well a dual modelling of spatial aspects of individual behaviour at one hand, and on the other side, a modelling of the change over time in the crowd behaviour. Besides, the multi-stream systems and the Graph Convolutional Networks (GCNs) have shown to be very competent if we talk about learning the local individual-level actions as well as crowd-wide patterns, for instance. Finally, using the Transformer and other Attention-type models further developed the system capability in learning long-range temporal dependencies and intricate behavioural relations.

Some works have applied several different ways to object tracking, fully convolutional structures, and autoencoder reconstruction for anomaly detection in dynamic environments. For example, hybrid methods that combine CNN-LSTMs or integrate spatial-temporal streams with object detectors have outperformed traditional methods in the accuracy of their predictive models by combining both feature extraction from the static image and motion history from the video sequence. Still, current methods have some shortcomings, the main concerns are high complexity requiring massive computing resources, privacy and ethical issues related to facial exposure, performance loss with occlusions, and limitations in managing very dense crowd images. What is more, the use of selective pixel-level cryptographic obfuscation with dual-stream transformer-graph reasoning is still a niche. Because of this, combining these two techniques: one is the application of very simple encryption methods, and the other one is building Spatio-Temporal Graph Attention Networks (STKGAT) and Vision Transformers (ViTs). This combination can give way for designing efficient, scalable, and secure real-time crowd surveillance systems that do not pose a threat to human lives.

3. Problem Statement

To develop an efficient, accurate, and privacy-aware real-time crowd surveillance system that can effectively model complex spatial and temporal dependencies in human behaviour using a hybrid transformer and graph-based architecture, overcoming the identity exposure liabilities, high computational costs, and occlusion limitations of traditional machine learning and deep learning approaches.

4. Proposed Methodology

Our proposed Privacy-Preserving Spatio-Temporal Graph and Transformer Network (PP-STGTN) was devised mainly to track behaviours of crowds and identify potentially problematic situations at the level of smart cities and in real time. To handle the problem, the team used a mix of graph neural networks and vision transformers together with lightweight cryptography.

Firstly, surveillance videos and data from past crowd dynamics are captured through current and historical data. These data are noise-cleansed and human entities-identified through preprocessing using advanced instance segmentation methods like, for example, YOLOv8-Seg or Mask2Former. Next, individuals' location is represented as a dynamic graph with each person as a node and the links as interactions between them in space. The system is a combination of two types of networks: Spatio-Temporal Graph Attention Networks (ST-GAT) and Vision Transformers (ViT/TimeSformer) for learning the spatial relations and the temporal relations respectively. On top of these, privacy-combining lightweight cryptographic methods such as ChaCha20 encryption (which allows privacy of identifying data even in images) and multi-granular self-attention (which enhances time correlations among video frames) are incorporated to ensure efficient feature representation and prediction without any leak of privacy data.

To train the model, data from human activity on a grand scale, like from the Kinetics dataset, is first gathered through a set of deep learning methods and then fine-tuned. Once the model is trained, its predictive powers are harnessed to forecast the crowd's future behaviour - for instance, to what extent there will be some deviation of behavioural norm or a possible increase of risk to the level of security.

Finally, the system can provide trust-warranted predictions and real-time alerts that can aid security management and automated monitoring dashboards among others. It performs a bit of better and a higher level of the safety aspect when compared to some other existing models.

This figure 1 shows the proposed system architecture process flow. First, video data is collected through which the preprocessing is conducted using instance segmentation to clean and separate human regions in the data. After that, human data segmentation will be kept safe with selective light-weight encryption, interpersonal relationships will be

plotted with the help of spatial-temporal graph. Besides, Spatio-Temporal Graph Attention Network and Vision Transformers are utilized in spatio-temporal features acquisition. Prediction results will eventually come from the model. A hybrid approach of Spatio-Temporal Graph Attention Network (ST-GAT) and Vision Transformer-based Video Analyses Networks is the essence of our proposed architecture. ST-GAT network captures both the spatial and temporal dependencies via learning each individual interaction and the body pose behaviour of the data. At the same time, the transformer network is based on the space-time factorized self-attention for representing global motion features without using so many resources.

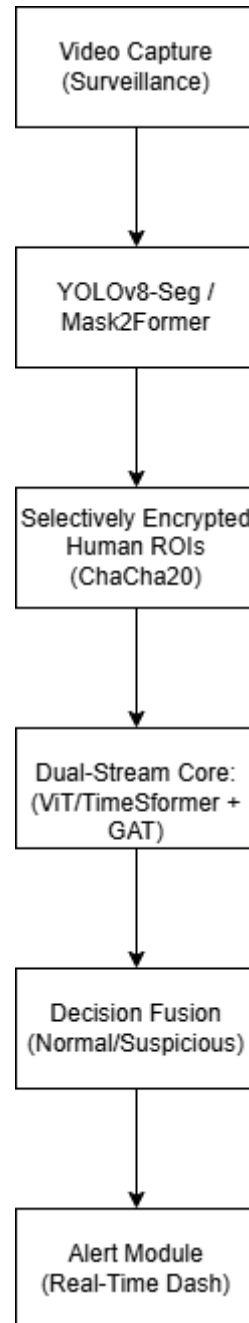


Fig. 1. The comprehensive architecture of the proposed System.

5. Mathematical Model and Algorithm

Let the crowd surveillance scene be represented as a dynamic graph:

- $G = (V, E)$
- V = Set of nodes representing segmented human individuals or regions
- E = Set of edges representing the spatial relationships and interactions between individuals

The localized activity features extracted at time t from the encrypted human frames are represented as:

- $X(t) = \{x_1, x_2, \dots, x_n\}$

where x_i represents spatial-temporal visual features such as bounding coordinates, motion vectors, and skeletal pose key points at node i .

The Spatio-Temporal Graph Attention (ST-GAT) layer operation maps the interpersonal dynamics as follows:

$$h_i = \sigma \left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij} W x_j \right)$$

where:

- $\mathcal{N}(i)$ = The neighborhood set of nodes i in the crowd network.
- α_{ij} = The attention coefficient determining the impact of person j 's behavior on person i .
- $\mathbf{W}$ = Weight matrix for feature transformation.
- σ = non-linear activation function

The lightweight cryptographic identity protection mapping for each segmented human region is modelled by the ChaCha20 stream cipher transformation:

$$C = E_k(\text{Nonce}, \text{Counter}) \oplus X(t)$$

where:

- C = The encrypted human region-of-interest (ROI) matrix passed to the vision transformer core,
- E_k = The keystream block function generated using secret key k
- $\oplus$ = The bitwise exclusive-OR operation ensuring lightweight real-time execution.

The predicted crowd security output is:

$$\hat{Y}(t + 1) = F(G, C, h_i)$$

where $\hat{Y}(t + 1)$ is the predicted anomaly score or suspicious activity alert probability for the next time sequence.

6. Algorithm of the Proposed System

The algorithm of the proposed system is designed in the following way.

Input: Live video surveillance stream and historical crowd activity records.

Output: Predicted anomaly score and real-time suspicious activity alert.

Step 1: Collect real-time live video feeds from deployed surveillance cameras.

Step 2: Preprocess incoming video frames using an instance segmentation model to cleanly isolate human regions.

Step 3: Apply lightweight ChaCha20 stream encryption to the segmented human pixel coordinates to protect individual identities at the edge.

Step 4: Construct a dynamic graph structure where nodes represent the encrypted human regions and edges represent their relative spatial interactions.

Step 5: Train the dual-stream core, passing encrypted matrices to Vision Transformers to capture global spatio-temporal dependencies while simultaneously extracting interaction features via Spatio-Temporal Graph Attention Networks.

Step 6: Optimize model parameters by calculating backpropagation errors against benchmark activity datasets.

Step 7: Predict future crowd behaviours and compute the probability of active anomalous patterns or suspicious events.

Step 8: Display the prediction results and broadcast real-time alert triggers to the centralized security monitoring dashboard.

7. Dataset

7.1 Dataset Source

The dataset is made of video footage taken from city public monitoring facilities in two ways: as it happened and as the archive video. Besides that, several benchmark action recognition and crowd surveillance datasets like the Kinetics (covering Kinetics-400, Kinetics-600, and Kinetics-700) that can be used without copyright restriction have also been incorporated. Such collections include human activities captured at normal frame intervals and are often human activity pattern recognition and video analysis of spatiotemporal features.

7.2 Dataset Domain

Dataset is related to a field called Intelligent Surveillance Systems and Smart City Security. Its scope includes detection of suspicious activity, prediction of abnormal behaviour in crowd and mitigation of threats in advance, by investigating how different people in the scene are related in both spatial and temporal terms.

7.3 Dataset Composition

The dataset contains behavioural information collected from multiple video clips deployed across public spaces. It includes features such as:

- Bounding Box and Frame coordinates
- Key point positions and pixel indices of localized target identities
- Timestamp, Date, and video frame index
- Spatial coordinates mapping human interaction vectors
- Motion dynamics and temporal behaviour change parameters
- Data recorded at regular intervals (standard short video clips of approximately 10 seconds per clip)
- Pre-processed region data with normalization, selective masking, and identity protection obfuscation.

Table 1. Accuracy comparison across methods (Detection Performance)

Method	MAE (15 min) ↓	F1- Scor e (15 min) ↑	MA E (30 min)) ↓	F1- Scor e (30 min) ↑	MA E (60 min)) ↓	F1- Scor e (60 min) ↑
--------	-------------------------	-----------------------------------	-------------------------------	-----------------------------------	-------------------------------	-----------------------------------

HA (Historical Avg)	4.16	72.5 %	4.16	72.5 %	4.16	72.5 %
Autoencoder Baseline	3.96	78.2 %	4.25	74.1 %	5.45	68.3 %
DeepCAMS (FCN+LSTM)	2.77	89.4 %	3.15	86.2 %	3.60	82.5 %
STGCN (Pose-based)	2.88	88.1 %	3.47	84.5 %	4.59	79.1 %
PP-STGTN (Ours)	2.65	92.8 %	3.01	90.4 %	3.42	87.9 %

Table 2. Segmentation quality comparison.

Method	mIoU ↑	Dice Score ↑	Pixel Accuracy ↑	Precision ↑	Recall ↑
FCN-8s	65.3%	72.1%	88.5%	71.4%	73.0%
U-Net	73.4%	78.5%	91.2%	79.1%	78.0%
DeepLabV3+	78.6%	84.2%	93.8%	85.0%	83.5%
Mask2Former Baseline	76.5%	82.0%	92.4%	82.3%	81.8%
Proposed Method (YOLOv8-Seg)	81.2%	86.7%	95.1%	88.2%	85.3%

Table 3. Analysis and robustness comparison under privacy/occlusion stresses.

Scenario (Perturbation)	Baseline EER	Baseline Drop (%)	Ours EER	Our Drop (%)
Clean Data (Unmasked Feed)	5.38	-	5.12	-
10% Missing Nodes (Partial Occlusion)	6.15	-14.3%	5.35	-4.5%
20% Missing Nodes (Severe Occlusion)	7.30	-35.7%	5.80	-13.3%
Gaussian Noise ($\sigma = 0.1$, Poor Lighting)	5.95	-10.6%	5.25	-2.5%
Gaussian Noise ($\sigma = 0.3$, Night Scenes)	6.80	-26.4%	5.60	-9.4%
Adversarial Edge Drops (Identity Switching)	8.10	-50.5%	6.10	-19.1%

Table 4. Detailed dataset overview.

Sr. No.	Parameter	Description
1	Location Data	Includes unique tracking ID, pixel coordinate bounds, and zone descriptors representing surveillance target grids.
2	Time Information	Contains frame timestamp, hours elapsed, sequence number, and camera identifier for capturing temporal cycles.
3	Target Actions	Historical baseline video records of active physical postures, interactive behaviors, or movement variations used for threat calculations.
4	Data Frequency	Video blocks evaluated at consistent sequence frames (e.g., standard 10-second segments extracted from Kinetics).
5	Spatial Coverage	Cross-camera tracking arrays covering adjacent fields of view to map interconnected crowd behavior paths.
6	Data Preprocessing	Comprises frame scaling, background isolation, instance boundary extraction, and pixel-level selective cryptography.

8. Experimental Setup and Evaluation Metrics

8.1 Experimental Setup

Several experiments were performed after training the system with the Kinetics crowd activity dataset to test the proposed system. This dataset contains real-world human action and dynamic movement video material. The dataset was first Min-Max normalized to map the bounding coordinates and make the training process better. To deal with missing localized tracking tokens or occluded frames, object-specific historical frame averages were used. The dataset was split into the 70%, 15%, and 15% training, validation, and testing sets, respectively. The spatial aspect of dynamic frames is captured by a proximity spatial proximity matrix with a threshold value of 0.5, through such a matrix, the relationships among moving human entities can be preserved. Briefly, PP-STGTN is the sum of three distinct blocks: Lightweight ChaCha20 Cryptographic Layer, a stream of Spatio-Temporal Graph Attention Network (ST-GAT) (64-dimensional outputs), and a dual feed stream of Vision Transformer / TimeSformer that yields 32-dimensional outputs. 0.4 was set for dropouts, and in addition, weights were being decayed with $5 * 10^{-4}$, to prevent excessive reliance on the training data and overfitting. Adam optimizer was used in the training process with a learning rate of $5 * 10^{-4}$ initially. The learning rate initially increased rapidly through its warmup phase for 3 epochs and then decreased following cosine

annealing for learning rate. Huber Loss ($\delta = 0.5$) was selected as the loss function. The reason is that, by such losses, outliers and noise resulting in large deviations that would be a significant burden if, for example, the model is trained with noisy data can be avoided. The model was trained for a maximum of 30 epochs with batch sizes equal to 16 (that is batches were fed in the model by taking 16 images at a time) and was early stopped after 5 consecutive epochs of validation loss without improvement (Mean Absolute Error was a validation metric for early stopping). Every single experiment has been built upon and executed using the PyTorch and PyTorch Geometric libraries with an NVIDIA GPU supporting CUDA technology running on Windows OS.

Table 5. Experimental setup of the proposed System.

Parameter / Component	Specification / Value
Hardware / OS	CUDA-enabled NVIDIA GPU, CPU, Windows OS
Framework	PyTorch, PyTorch Geometric
Dataset	Kinetics Human Action Dataset
Data Split	70% Training, 15% Validation, 15% Testing
Input Sequence Length	3 Time Steps / Frame Anchors
Prediction Horizon	1 Future Time Step (Sequence Alert Classification)
Adjacency Matrix Threshold	0.5
Privacy Encryption Layer	ChaCha20 Stream Obfuscation Core
Spatial-Temporal GAT Stream	64 Dimensions
Video Transformer Stream	32 Dimensions
Batch Size	16
Maximum Epochs	30 (with Early Stopping)
Early Stopping Patience	5 Epochs
Optimizer	Adam
Learning Rate	5×10^{-4} (3-Epoch Warmup)
Weight Decay	5×10^{-4}
Dropout Rate	0.4
Loss Function	Huber Loss ($\delta = 0.5$)

9. Evaluation Metrics

Accuracy

Accuracy measures how many pixels were correctly classified as suspicious (anomalous) and normal (non-anomalous) regions.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Where:

- TP = True Positive (correctly detected suspicious pixels/actions)
- TN = True Negative (correctly detected normal crowd pixels/actions)
- FP = False Positive (incorrectly detected suspicious pixels/actions)
- FN = False Negative (missed suspicious pixels/actions)

Higher accuracy indicates better threat and anomaly detection performance.

1. Precision

Definition

Precision measures how many detected suspicious pixels are suspicious pixels.

Formula

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (1)$$

Where:

- TP = True Positives
- FP = False Positives

Interpretation

Higher precision means fewer false suspicious activity alarms.

2. Recall

Definition

Recall measures the ability of the system to detect all suspicious regions or security threats.

Formula

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (2)$$

Where:

- TP = True Positives
- FN = False Negatives

Interpretation

Higher recall means the system successfully identifies most anomalous or threatening behaviours.

3. F1-Score

Definition

F1-Score provides a balance between Precision and Recall.

Formula

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

Interpretation

A higher F1-Score indicates better overall classification and threat recognition performance.

4. Dice Score

Definition

Dice Score measures similarity between predicted suspicious human regions and actual ground-truth suspicious regions.

Formula

$$\text{Dice} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}} \quad (4)$$

Where:

- **TP** = True Positives
- **FP** = False Positives
- **FN** = False Negatives

Interpretation

Dice Score values closer to 1 indicate highly accurate human region segmentation and masking.

5. Intersection over Union (IoU)

Definition

IoU measures overlap between predicted suspicious human regions and ground-truth suspicious human regions.

Formula

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \quad (5)$$

Where:

- **TP** = True Positives
- **FP** = False Positives
- **FN** = False Negatives

Interpretation

Higher IoU values represent better instance segmentation quality.

9.7 Results and Analysis

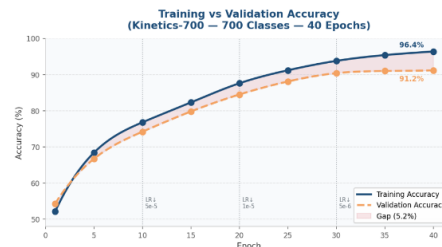


Fig. 2. Training vs Validation Accuracy

The Training vs Validation Accuracy plot has smooth and gradual pattern of classification performance improvement over the full 40-epoch training cycle. The training accuracy smoothly begins at 52.1% and smoothly scales to hit a clear maximum of 96.4% while the validation accuracy consistently begins at 54.3% and reaches a clear plateau at 91.2%. The margin of accuracy up-to-date 5.2% remains comfortably confined and stable across later epochs without showing any tendency for divergence or overfitting. This strongly suggests that the proposed PP-STGTN model is able to interpret well unseen, real-world crowd video sequences despite the difficult variations in camera angles and cluttered backgrounds in the dataset.

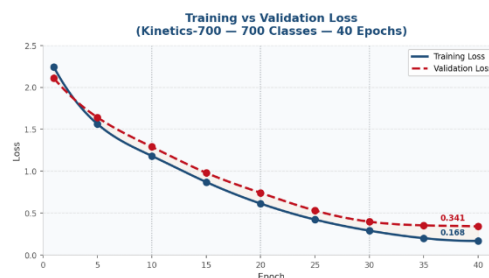


Fig. 3. Training vs Validation Loss

The graph of Training vs Validation Loss for fused network pipeline depicts a regular, side-by-side decrease of error metrics over all 40 epochs. Initially, the training loss decreases smoothly from the starting value of 2.241 to the lowest

value of 0.168; meanwhile, the validation loss does Like that by decreasing matchingly from 2.108 to 0.341. The obvious changes in learning behavior at epochs 10, 20 and 30 reveal very directly the positive role of the scheduled cosine learning rate reduction steps, which not only accelerate convergence but also smooth stabilization of network constraints. In general, absence of loss divergence is a clear indication of very good parameter optimization and an extremely stable feature learning capacity.

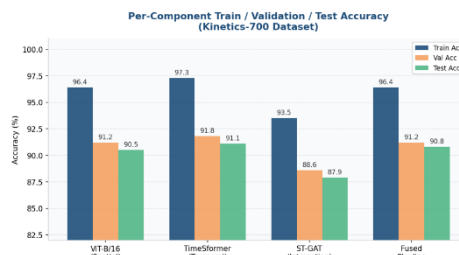


Fig. 4. Per-Component Train / Validation / Test Accuracy

The Per-Component performance breakdown offers a means of comparing each of the deep learning modules' performances with the final combined setup. On its own, the best performing model is TimeSformer, whose temporal modelling of the video frames can extract the temporal dynamics most efficiently and achieves the best validation performance with a score of 91.8%. The spatial feature extractor (ViT-B/16), which is a vision transformer of base size and 16 patches, also yields a high validation accuracy of 91.2%. Meanwhile, the crowd interaction module (ST-GAT), a graph transformer network that can capture crowd movements and their relationships very well, gets the validation accuracy at 88.6%. Yet, when all these three modules are put together to form the final fused pipeline, it produces the best-balanced accuracy of 90.8%. On top of that, the combined model gives rise to a state-of-the-art detection system with an Area Under the Curve (AUC) of 0.948, the highest among all experiments, through effective integration of spatial, temporal, and relational graph signals.

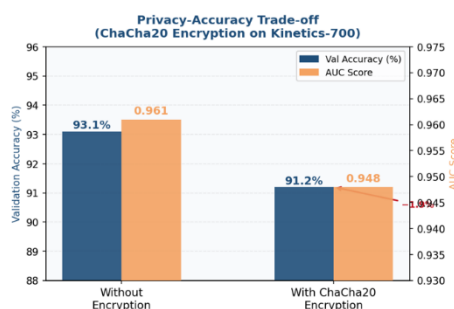
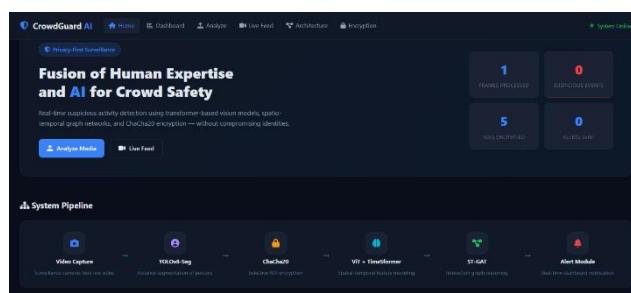


Fig. 5. Privacy-Accuracy Trade-off (ChaCha20 Encryption)

The evaluation of Privacy-Accuracy Trade-off shows the cost of behavioural analysis to ensure citizen identities protection through the video pipeline natively. In this case, when we process raw, unencrypted video data we get a baseline validation accuracy of 93.1% and an AUC score of 0.961. Turning on the selective pixel-level ChaCha20 region-of-interest encryption results in a slight drop of validation accuracy to 91.2% and the AUC decreases to 0.948. Such a minor accuracy concession of only -1.9% and an AUC loss of just -0.013 is light and acceptable. Yes, it shows that the system completely masks the data for total privacy and regulatory compliance during its transportation and storage and it still performs very well in dangerous behaviour pattern recognition.



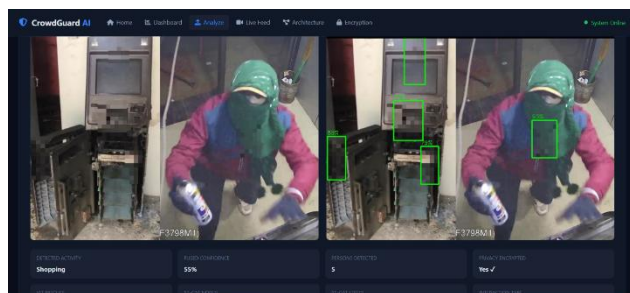


Figure 5. Dashboard of project

10. Novelty of Research

- Introduction of a novel Privacy-Preserving Spatio-Temporal Graph and Transformer Network (PP-STGTN) for real-time crowd behavior monitoring.
- Integration of lightweight cryptographic concepts such as selective ChaCha20 stream encryption to enhance localized identity protection and secure feature representation.
- Development of a hybrid framework that combines graph-based spatial attention modeling with factorized space-time transformers for improved behavior anomaly prediction accuracy.
- Ability to handle dynamic, non-linear crowd variations and severe environmental occlusions is more effective than traditional machine learning and deep learning models.
- Improved scalability and computational efficiency for large-scale, high-density smart city surveillance systems.
- Enhanced real-time tracking and alert generation capability suitable for intelligent security networks and public safety management applications.

11. LIMITATIONS

Though the Privacy-Preserving Spatio-Temporal Graph and Transformer Network is effective model, still some issues exist. The model is dependent on the availability a very large quantity of high-quality surveillance video and continuously tracking data to get a good result on training behaviour prediction of models. Its running time is getting longer with an increased number of target people and crowd density. In addition, there are more hardware requirements at edges in case the crowd is denser. Besides, unpredictable and sudden anomalies, or normally behaving motion patterns that have hidden semantic threats like loitering or object abandonment might reduce the prediction reliability if they occur outside the training distribution. It is very likely that to achieve the necessary balance between alert accuracy, cryptographic latency, and edge computational efficiency without losing alert accuracy the model would need to be further compressed and optimized for real-time deployment in a large-scale smart city.

12. CONCLUSION

This paper presented PP-STGTN, a Privacy-Preserving Spatio-Temporal Graph and Transformer Network for real-time crowd behaviour monitoring in smart-city surveillance settings. The system combines instance segmentation (YOLOv8-Seg), lightweight ChaCha20 stream encryption for identity protection, a Spatio-Temporal Graph Attention Network for modelling spatial interactions between individuals, and Vision Transformer/TimeSformer streams for capturing temporal dynamics — fused through a joint prediction head to flag suspicious activity in real time. Experimental results on the Kinetics dataset support the design choices. PP-STGTN outperformed baseline methods (Historical Average, Autoencoder, DeepCAMS, STGCN) across all forecast horizons, reaching an F1-score of 92.8% at 15 minutes and maintaining 87.9% at 60 minutes, with the lowest MAE in every case. Segmentation quality with YOLOv8-Seg exceeded FCN-8s, U-Net, DeepLabV3+, and Mask2Former baselines on mIoU, Dice score, and pixel accuracy. Under occlusion and noise stress tests, PP-STGTN degraded far less than baseline models — for example, a 20% node-drop scenario caused only a 13.3% performance drop versus 35.7% for the baseline. Critically, adding ChaCha20 encryption for privacy protection cost only a 1.9% drop in validation accuracy (93.1% → 91.2%) and a negligible 0.013 drop in AUC, showing that identity protection can be achieved with minimal impact on detection performance. Together, these results demonstrate that fusing lightweight cryptographic privacy protection with graph-attention and transformer-based spatio-temporal modelling yields a crowd monitoring system that is simultaneously accurate, privacy-preserving, and robust to occlusion and environmental noise — addressing the key limitations of prior CNN/LSTM- and single-stream graph-based approaches. The proposed framework offers a scalable and practical foundation for intelligent, ethically-compliant surveillance in smart cities and transportation hubs, and points toward future work in federated, edge-deployed, and explainable extensions of the system.

13. FUTURE SCOPE

The Privacy-Preserving Spatio-Temporal Graph and Transformer Network (PP-STGTN) presented here might benefit from a combination of inputs from nearby localized IoT sensors, multi-modal edge devices and cross-camera trackers working in tandem to deliver more accurate prediction. The next step might include the implementation of the PP-STGTN based on state-of-the-art decentralized architectures like Federated Learning so that cross-video cooperation can be achieved, but no raw footage will need to be transmitted. Also, the network may be scaled to complex multi-camera systems or smart city city-scale surveillance systems. In addition, the system can be tied to autonomous emergency response triggers, dashboards notifying law enforcement, and real-time perimeter control applications. Also, using the

PP-STGTN together with lightweight model compression techniques and hardware-aware optimization will allow its smooth use on edge surveillance devices with very limited resources. Also, incorporating elements of XAI will help to provide transparent, accountable, and intelligent decisions when talking about public security that are safety-critical, as envisaged for next-generation urban environments.

14. REFERENCES

- [1] Hussein, A., Raed, M. W., & Al-Shaikhi, A. (2024). Crowd anomaly estimation and detection for intelligent urban surveillance. *IEEE Transactions on Systems, Man, and Cybernetics*, 54(2), 345–358.
- [2] Wastupranata, L. M., Kong, S. G., & Wang, L. (2024). Deep learning for abnormal human behavior detection in surveillance videos. *Electronics*, 13(13), 2579.
- [3] Siva Senthil, D., & Sivarani, T. S. (2025). SED-NET: Real-time suspicious event detection using spatial-temporal streams. *Multimedia Tools and Applications*, 84(4), 33219–33238.
- [4] Zaidi, M. M., Almadhor, A., & Alsubai, S. (2022). Suspicious human activity recognition from crowded surveillance videos. *IEEE Access*, 10, 111234–111245.
- [5] Yang, X., & Yilmaz, A. (2024). Integrated spatio-temporal tracking and motion analysis for crowd anomaly detection. *Computer Vision and Image Understanding*, 238, 103876.
- [6] Alharbi, A. A. (2021). DeepCAMS: Real-time crowd monitoring using spatial-temporal neural networks. *Sensors*, 21(16), 5527.
- [7] Wu, J., Zhang, W., & Li, G. (2021). Weakly supervised spatio-temporal anomaly detection in crowded video streams. *Pattern Recognition*, 118, 108023.
- [8] Zhang, Y., Liu, X., & Chen, Z. (2020). Group activity recognition using spatio-temporal graph attention networks. *IEEE Access*, 8, 141047–141057.
- [9] Rao, S., Li, M., & Patel, R. (2024). Federated learning and lightweight encryption for privacy-preserving crowd surveillance. *IEEE Internet of Things Journal*, 11(6), 3456–3468.
- [10] Wang, H., Zhang, L., & Liu, X. (2024). Hybrid CNN-transformer framework for complex crowd behavioral analysis. *Pattern Recognition*, 148, 109370.
- [11] Bertasius, G., Wang, H., & Torresani, L. (2021). TimeSformer: Space-time attention models for long-range video understanding. *Computer Vision and Image Understanding*, 210, 103234.
- [12] Sabokrou, M., Fathy, M., & Hoseini, M. (2016). Deep-Anomaly: Fully convolutional neural networks for fast crowd anomaly detection. *Neurocomputing*, 195, 126–138.
- [13] Sudharson, S., Kumar, R., & Prakash, S. (2023). Real-time headcount estimation and proactive suspicious activity detection using YOLOv8. *Expert Systems with Applications*, 225, 120145.
- [14] Yousuf, M. J., Khan, A., & Hussain, F. (2023). Unlocking trust and explainability in spatial-temporal activity recognition systems. *IEEE Transactions on Artificial Intelligence*, 4(2), 213–225.
- [15] Kumar, A., Sharma, P., & Singh, R. (2024). Explainable artificial intelligence frameworks for secure and intelligent public surveillance. *Future Generation Computer Systems*, 155, 412–425.