

A Hybrid Vision Transformer–Graph Neural Network Framework for High-Resolution Surveillance Scene Understanding

Abeer Ahmed Ali ¹, Abdullah Ghanim Jaber ², Ali A. Mahmood ², Mohammed Jamal Salim ²

¹ Department of Computer Science, College of Science, University of Dijlah, Baghdad, Iraq

² University of Information Technology and Communications (UoITC), Baghdad, Iraq

ARTICLE INFORMATION

Article History:

Received 11 March 2026

Revised 05 June 2026

Accepted 24 September 2026

Keywords:

Vision Transformer;
Graph Neural Network;
Attention Mechanism;
Multi-modal Feature Fusion;
Relational Scene Understanding;
UA-DETRAC Dataset

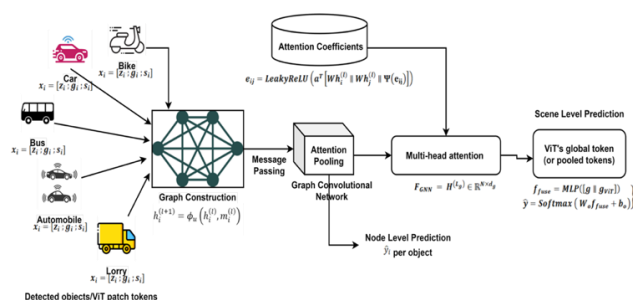
Corresponding Author:

Abdullah Ghanim Jaber,
University of Information
Technology and Communications,
Baghdad, Iraq.
abdullah.ghanim@uoitc.edu.iq

This work is open access under a
[Creative Commons Attribution-Share
Alike 4.0](https://creativecommons.org/licenses/by-sa/4.0/)



ABSTRACT



Real-time processing and accuracy of surveillance systems in a real-world environment are very critical to the understanding of the scene. Current deep learning methods, however, have failed to sufficiently learn the global context and relations between objects from high resolution surveillance information. In this study, a hybrid deep learning framework is proposed, which incorporates the merits of Vision Transformers, Graph Neural Networks, and attention-based fusion to tackle these challenges. The novelty of the research is that they design a single architecture combining global features extraction, and relational reasoning to better interpret the surveillance scene. The proposed model has an improved representation of features, which includes complementary mechanisms in a single model. In terms of method, Vision Transformers are used to use global spatial features from surveillance frames. Relationships between the detected objects are described in a structured graph using the Graph Neural Networks. In complex scenes, these features are adaptively integrated together through an attention guided fusion scheme, which improves the classification and understanding of complex scenes. Experimental results on UA-DETRAC dataset show that the proposed framework achieves higher accuracy, precision, recall, F1 score, mAP and AUC (0.962%, 0.955%, 0.958%, 0.956%, 0.960% and 0.97 respectively). Furthermore, the model's performance is 78 fps making it a compatible model for real-time surveillance systems. Analysis of these results shows that the method is robust and improved consistently compared with the baseline methods in all performance measure. Last but not least, the proposed framework is very efficient and accurate, and is able to achieve a good balance between accuracy and computation efficiency in real-time intelligent surveillance. The results prove its potential application for surveillance, traffic monitoring and security critical environments in smart cities.

Document Citation:

A. A. Ali, A. G. Jaber, A. A. Mahmood, and M. J. Salim "A Hybrid Vision Transformer–Graph Neural Network Framework for High-Resolution Surveillance Scene Understanding," *Buletin Ilmiah Sarjana Teknik Elektro*, vol. 8, no. 5, pp. 1393-1426, 2026, DOI: [10.12928/biste.v8i5.16148](https://doi.org/10.12928/biste.v8i5.16148)

1. INTRODUCTION

The developments in AI and Deep Learning in recent years have enhanced the performance of image analysis, object detection, and surveillance systems. Numerous techniques like CNNs, Vision Transformers (ViTs), and Graph Neural Networks (GNNs) have proven effective in capturing visual features, understanding context, and enabling automated decision-making in complex contexts. Also, real-time surveillance applications are problems that are still open for research, including high-resolution image processing, computational efficiency and relational reasoning.

1.1. Background: AI-Driven Transformation in Security Surveillance

Recent surveillance systems with computer-vision and artificial intelligence (AI) have had a fundamental transformation in the field of infrastructure protection and safety of the people [1]. Autonomous surveillance platforms can be deployed in large-scale environments to continuously monitor and automatically detect suspicious activities, unauthorized access and anomalous behavior with minimal human intervention [2]. The collaboration between advanced deep learning (DL) models and high-resolution imaging technologies is the major cause of this change since it allows detecting and making decisions in real-time under different conditions, such as border crossings, airports, military bases, and smart cities [3]. Although these have been made, the challenge of properly classifying images in complex real-world surveillance is still a challenge and is dynamic [4].

1.2. Challenges in High-Resolution Image Classification

The variety of spatial and contextual cues that high-resolution images provide are essential to the process of identifying dangers, and they pose significant modeling and computing problems. The huge amount of pixel-scale data increases processing and memory requirements and real-time categorization is computationally expensive [5]. The traditional convolutional neural networks (CNNs) that dominate the industry often become less accurate when scaled to operate with such large inputs and a trade-off between speed and resolution is needed [6]. Other environmental factors, including limited illumination, visibility, crowd density, occlusions, and novel objects, also decrease system reliability [7]. The difficulties mentioned above show that the strengths required in real security situations with high stakes are not the same as those obtained in most DL techniques in the lab [8].

1.3. The Need for Robust Autonomous Security Surveillance Systems

The consequences of such delays or misclassifications could be very serious in such a discipline where the security issue is a concern [9]. The safety surveillance in urban areas requires real-time recognition of anomalies in crowded areas, border security requires a high level of accurate identification of suspicious activities, and airports need to be able to promptly identify unwanted conduct [10]. Continuous monitoring, operator fatigue and cognitive limitations are a challenge to manual surveillance and are driving the development of intelligent automated surveillance systems. Consequently, an increased demand to have independent systems that are precise, scalable and pliable has arisen [11]. In order to respond appropriately to novel threats in the diversity of the uncertain context, high-resolution inputs should be rapidly ingested by the systems and the system should retain some level of context awareness [12].

1.4. Impact of Deep Learning Architectures on Image Classification

The use of deep learning has driven spectacular progress in image classification within the past decade. During the initial applications, CNNs established new performance metrics and were found to be better at local feature extraction [13]. Video-based surveillance has now been enhanced with time modeling features given by recurrent neural networks (RNNs) and Long Short-Term Memory (LSTM) networks, allowing them to track the dynamic events [14]. Attention mechanisms could boost classification accuracy by several folds because they allowed models to focus on significant parts of an image instead of looking at all input pixels in an equal manner [15]. More recently, with the help of global self-attention, an effective model of long-range spatial dependencies, Vision Transformers (ViTs) have presented a potential in the high-resolution input processing [16]. ViTs alone are often not sufficient in surveillance, where interactions between people, vehicles or objects at the isolated acquisition point may contain more information than individual appearances [17].

1.5. Graph Neural Networks: Contextualizing Surveillance Intelligence

To overcome the limitations of classical topologies, one can use the category of networks that are able to simulate relationships between objects called Graph Neural Networks (GNNs) [18]. GNNs can also be extended to go beyond the object recognition to scene understanding by representing the scenes under

surveillance as graphs where nodes correspond to the detected objects and edges represent the location or surrounding concepts of these objects [19]. This is important for detecting complex security scenarios such as traffic violations, crowd behavior or organized suspicious activity [20]. Graph neural networks (GNNs) combined with other global feature extractors such as ViTs to capture fine-grained object information, scene structure and object relationships [21].

1.6. Toward a Hybrid Framework: Surveillance Hybrid Deep Learning (Surv-HDL)

In this paper, we present the Surveillance Hybrid Deep Learning Framework (Surv-HDL) that was designed to overcome these drawbacks. Surv-HDL has three complementary components:

- High-resolution images can be used to extract global spatial features using Vision Transformers (ViTs).
- Graph Neural Networks (GNNs): To capture the contextual dependency and reasoning between identified things.
- Attention mechanisms that highlight relevant parts and ignore irrelevant and misleading information.

These modules are offering a complete solution from Surv-HDL which improves the accuracy of the classification, flexibility to adjust settings and ability to deal with difficult situations, including low visibility and occlusion. The images used in this study are the high-resolution surveillance images (960x540) of the UA-DETRAC dataset, and these images are not a 4K (3840x2160) or 8K (7680x4320) standards, but they do contain a lot of information when compared to traditional surveillance video inputs with lower resolutions, and are hard to analyze in real time because of object density, occlusion and computational complexity.

1.7. Motivation and Problem Statement

Today's security settings demand surveillance solutions which are accurate, scalable, and real-time. Although high-resolution imagery can aid in scene understanding, existing deep learning models have computational limitations and may not be as flexible for large-scale surveillance applications. The autonomous surveillance system should be capable of effectively processing the high-resolution images and should be capable of accurately modelling the contextual relationships between the objects detected.

CNNs have proven useful in extracting local features but are less efficient at capturing long-range dependencies and can lose fine-grained details, as a result of successive downsampling. To overcome this, vision Transformers (ViTs) learn global relationships between image patches using self-attention mechanisms to achieve richer scene understanding. But the structural relationships between objects are not explicitly represented in ViTs. This is overcome by a graph representation of the surveillance scene, using nodes to represent entities detected in the scene and edges to model spatial or semantic relationships. This is addressed by the Graph Neural Network (GNN), which is used to represent the surveillance scene as a graph, with nodes representing entities detected in the scene and edges representing spatial or semantic relationships.

Although they are all powerful in their own right, none of the architectures are effective in all three aspects, local feature learning, global context understanding and relational reasoning. This encourages the creation of a hybrid solution that combines these complementary technologies, yet remains efficient in terms of computing needs for real-time surveillance tasks.

1.8. Problem Statement

These limitations can limit performance in some video surveillance applications, such as occlusions, dense object interactions, and varying illumination conditions. Therefore, a framework that can capture the global context of the scene and model the relationships between the objects at the object level, and can be applied in real-time decision making for high-resolution surveillance applications is needed.

The main contributions of the research are,

- To tackle this classification problem in high-resolution surveillance images, a novel Surveillance Hybrid Deep Learning Framework (Surv-HDL) is proposed by combining Vision Transformers (ViTs), Graph Neural Networks (GNNs) and attention-based feature fusion.
- A spatial graph based relational reasoning module is designed to model spatial interactions and context-dependency of detected objects in complex surveillance scenarios.
- The multi-level attention fusion mechanism is designed to effectively fuse the global scene representation with the object-level relational features, which results in better feature discrimination and context understanding.
- An efficient pre-processing and patch-tokenization pipeline is developed to enable high-resolution surveillance images analysis under different environmental conditions.

- The proposed approach is validated by comprehensive experiments on the UA-DETRAC dataset, with 96.2% accuracy, 96.0% mAP, 0.97 AUC and real-time performance of 78 FPS.

Surv-HDL proposes a holistic framework combining global scene understanding, graph-based relational reasoning and multi-level attention-guided feature prioritization for high-resolution surveillance image classification, different from the current hybrid ViT–GNN models mainly based on feature fusion or object representation learning. It is not limited by occlusion, illumination changes, and dense traffic conditions, and simultaneously models the scene-level context and object-level interactions with a dedicated cross-attention fusion mechanism, while processing data in real-time. The special design and implementation of Surv-HDL make it superior to the current ViT–GNN models in real-time surveillance scenarios and allow for the fusion of contextual reasoning and attention.

1.9. Research Question

RQ1: What are the improvements in terms of both accuracy and real-time performance in high-resolution surveillance scene understanding when using a hybrid Vision Transformer–Graph Neural Network framework with attention-based fusion compared to current CNN, Transformer, and GNN based methods?

Current CNN-based approaches struggle to capture larger-scale contextual information because of the hierarchical downsampling, and Vision Transformers have high computational complexity for high-resolution surveillance data. Although Graph Neural Networks are adept at capturing relations between objects, they can be complex and expensive to compute, with an inability to be effectively deployed in real-time. These limitations highlight the importance of developing a single hybrid solution that balances accuracy, contextual understanding and computational efficiency.

The proposed Surv-HDL framework yields the following creative research contributions: (i) A structured graph-based modeling approach is presented that leverages the global feature extraction capabilities of Vision Transformers, relational reasoning capabilities of Graph Neural Networks and adaptive feature fusion capabilities of attention mechanisms; (ii) An efficient fusion strategy is designed to balance the accuracy-real-time performance trade-off for high-resolution video analysis; and (iii) Extensive experiments using the UA-DETRAC dataset demonstrate the superiority of the proposed framework in terms of accuracy, precision, recall, F1-score, mAP, AUC and inference speed when compared with the existing state-of-the-art methods.

The structure of the article is the following: in Section 1 the context and the research-related challenges and necessity to develop a hybrid surveillance model are introduced. Section 2 outlines the related work, focusing on the existing approaches and gaps in research. Section 3 provides the proposed methodology, which includes preprocessing, Vision Transformer (ViT), Modeling based on Graph Neural Network (GNN), attention-based fusion, and dual-output classification. Section 4 gives results and discussion with benchmark analysis. Section 5 brings the report to an end by outlining findings, limitations, and future research directions.

2. RELATED WORK

Abba *et al.* [22] suggested a real-time surveillance system that uses deep learning techniques, component labeling, background subtraction, and approximate median filtering for object identification, tracking, and recognition. On the MOT15, MOT16, and MOT17 (Multiple Object Tracking) datasets, the system achieves exceptional accuracy and precision because of its Python implementation and integration with C# for user-friendly program development. Although the results demonstrate increased monitoring efficacy, YOLO (You Only Look Once) -based techniques will likely be integrated in the future due to their limitations, which include scalability in crowded environments and adaptation to a variety of conditions.

Sukhavasi *et al.* [23] offered a thorough analysis of surveillance applications based on CMOS Image Sensors (CIS) in a variety of fields, including intelligent monitoring, space observation, aerial defense, agricultural, and driver assistance for automobiles. Methods concentrate on examining design features such as processing technology, dynamic range, frame rate, resolution, and signal-to-noise ratio. The findings show that over the past ten years, CIS adoption has increased, and technology has advanced. Low-light performance, high power consumption, and integration issues in resource-constrained environments are still limitations, nevertheless.

Zhang *et al.* [24] proposed attention-guided focalization approach of tracking and identifying objects in a structured deep learning model. To eliminate the contextual dependencies and the border refinement, the following methods could be utilized: utilizing a Global-Local Attention Fusion Strategy (GLFS), and adding multi-scale attention blocks into CNNs. The results show that the recall and accuracy are high compared to the conventional CNN-based detectors. Its drawbacks are an added sophistication and complexity of real-time processing in highly dynamic surveillance cases.

Hasanujjaman *et al.* [25] suggested an AI-powered framework for self-driving cars that combines intelligent networking, accurate localization, and 4D detection. LiDAR, RADAR, cameras, GPS substitutes, deep learning-based neural models, sophisticated image processing, and feature matching are some of the methods used. The findings show enhanced real-time positioning, enhanced detection accuracy, and dependable networking even in tunnels. High computational demand, reliance on multi-sensor calibration, and possible scaling issues in large-scale smart transportation systems are some of the drawbacks, though.

Tas *et al.* [26] used low-resolution surveillance images taken by standard security cameras to propose a lightweight CNN-based model for vehicle categorization. Methods include benchmarking against VGG16-based CNNs and training on a bespoke dataset of small (100×100 pixels, 96 dpi) vehicle images. The accuracy is 92.9% with less complexity, which makes it effective for low-cost systems, according to the results. Nevertheless, drawbacks include diminished robustness in extremely changeable situations and marginally worse accuracy compared to more sophisticated models.

Tan *et al.* [27] proposed the use of artificial intelligence and computer vision in enhancing the perception and judgment of autonomous vehicles. The approaches involve picture collecting, preprocessing, feature extraction, object detection, lane recognition, obstacle avoidance, and traffic sign recognition, which require the use of camera and sensor technologies. The findings point at enhanced environmental cognition of clever control in self-driving systems. However, it has disadvantages, including the low performance during bad weather and processing power, and inability to process in real time in complex traffic scenarios.

Kong *et al.* [28] supported the integration of computer vision, as well as artificial intelligence, to improve the perception and decision-making of autonomous vehicles. The procedural aspects utilize image capture through camera and sensor, image preprocessing, feature extraction, object detection, lane recognition, obstacle avoiding and traffic sign recognition. The results showed an improvement in environmental awareness and advanced autonomous driving system regulation. However, there are limitations of computing intensity, crippled functioning in the poor weather conditions and real-time processing issues in complex and dynamic traffic situations.

Gurav *et al.* [29] introduced a new surveillance system based on the combination of robotics with deep learning to detect and track faces in real time. These techniques include a CNN based detection system, PID controlled servo motors, and a robotic arm, which is operated by an Arduino, to make the camera dynamically adjust to the subject. Findings suggest that the accuracy of tracking is 92.5 percent with a latency of about 100ms, which is successful in low-light conditions. However, limitations include limited scalability, lack of multi-user tracking, and a dependence on the accuracy of hardware, which will require a subsequent addition of the Internet of Things (IoT) and AI.

Ahmed *et al.* [30] developed XNANet, the self-attention-based convolutional neural network, which was combined with the tiny-32 Vision Transformer to classify remote sensing images. The approaches include Bayesian hyperparameter optimization, network-level fusion, and RF-DE, a new meta-heuristic feature selection process, and classification based on the use of multiple classifiers. Accuracy in the results on the AID, RSSCN7 and SIRI-WHU datasets reached as high as 99.7 which is better than the existing methodologies. However, the limitations include high cost of computing, requirement of a large amount of data, and reduced effectiveness in real-time scenarios.

Alotaibi *et al.* [31] introduced an autonomic tracking system with AI-enhanced UAV-based vehicle tracking with license plate recognition. The processes include high-resolution picture acquisition, optical character recognition, and machine learning, which allow the immediate plate recognition that can be geospatially positioned. Findings show a high level of accuracy and efficiency compared to conventional surveillance and reduce the role of the human factor. However, the limitations include dependence on lighting and weather, real-time processing computational requirements and challenges in scaling to large urban environments.

Thomas *et al.* [32] promoted the development of SAE Level 5 Autonomous Vehicles based on the use of advanced CNN-based object recognition algorithms, as Improved YOLOv5, SSD, Mask-RCNN, and NanoDet. The techniques use the concepts of embedded boards running on GPU and high-resolution cameras and implement with ROS to recognize road features, such as humps and potholes. Results show the improvement of safe-navigation perception and decision-making. However, limitations include high computational costs, dependence on data sets and reduced capabilities in extreme weather conditions and highly unstructured environments.

Laksana *et al.* [33] introduced a smart city traffic management system which is a combination of deep learning and principal component analysis (PCA) in the effective classification of high-dimensional traffic CCTV images. The techniques include picture categorization through deep learning algorithms and dimensionality reduction through PCA to minimize computational complexity. The findings showed similar

accuracy (83) but a significantly reduced training time (2.95s vs to 80.43s). However, it has limitations such as the loss of some accuracy and inability to handle rapidly changing real-time traffic conditions.

Yang and Guo [34] introduced an unsupervised framework for anomaly detection in industrial image inspection based on a Vision Transformer combined with an autoencoder. This investigation focuses on the generalization problem that exists in CNN-based anomalies detection with complex anomalies. Results show that the reconstruction accuracy and the localization of the anomalies are improved over conventional autoencoders. But the method has high computational cost and has been validated mainly under industrial conditions and cannot be directly applied to real-time surveillance systems.

Ullah *et al.* [35] proposed an attention-based hybrid model that integrates the Vision Transformer attention with MRESN for anomaly detection in sequential data. The study addresses the issue of capturing the long-range temporal dependency in the conventional anomaly detection models. Experimental results demonstrate the performance advantages in terms of classification accuracy and temporal modeling ability compared to LSTM-based baselines. Nevertheless, it is a challenge for this model to be well scaled and well-structured in real-time surveillance applications due to its high architectural complexity.

Habeb *et al.* [36] introduced a transformer-based spatiotemporal attention model for large-scale unsupervised video anomaly detection. This paper presents a study on weak temporal representation in surveillance systems based on CNN. The results show that the performance of anomaly detection and localization is better in crowded scenes. But the method is CPU intensive and not well suited for running on the edge.

To solve this problem, Li *et al.* [37] proposed a surveillance anomaly detection algorithm based on a hybrid CNN–Transformer model, which can be regarded as a fusion model of local feature extraction and global attention mechanism. The problem tackled is the loss of the context information in CNN-only models. The results indicate that the proposed models outperform the baseline CNN and transformer models. The approach is, however, more computationally intensive and the hyperparameters have to be carefully tuned.

Meng *et al.* [38] introduced a transformer-graph hybrid network to combine the sensor embedding with multi-source anomaly detection. The research addresses the problem of solving the fusion of heterogeneous multimodal data. Results show enhanced robustness in anomaly detection for different data sources. However, it is only evaluated on sensor-based data, and has not been validated in visual surveillance scenarios.

Zoghلامي *et al.* [39] developed a graph neural network, called ViGLAD, for logical anomaly detection in structured environments. The study tries to fill the gap with the traditional vision models lack of relational reasoning. Experiments demonstrate that the method yields better inference capabilities in structured outlier detection problems. But scalability problems occur, owing to the overhead in creating the graph with the high-resolution data.

Le, and Kim [40] proposed attention-based residual autoencoder for video anomaly detection. Its study aims to enhance temporal feature representation in reconstruction-based model. Results show greater accuracy of detection compared to conventional autoencoders. But the technique does not work well when there are many objects in the scene that interact in complex ways.

Mishra *et al.* [41] proposed a Vision Transformer based framework called VT-ADL for image anomaly detection and localization. The Inefficient Feature Representation in CNN-based Anomaly Detection System is addressed as a problem. The results indicate that the strengths of the localization performance and accuracy of the anomaly detection are well demonstrated. The model is, however, very expensive to run and not optimized to be run in real-time.

The authors Caville *et al.* [42] suggested Anomal-E, a graph neural network (GNN) based intrusion detection system (IDS) that is self-supervised. The study tackles the problem of lacking labeled data in the anomaly detection problem. Results show that the accuracy of abnormal network behavior detection is good. It does have limitations however, and isn't applicable to visual surveillance systems as such.

Li *et al.* [43] proposed an unsupervised cross-view transformer-based graph anomaly detection (CVTGAD) framework. The study presents the problem of poor cross-view representation for graph-level anomaly detection problems. The results reveal better generalization with respect to graph structure. But another challenge is lack of stability and sensitivity when training with graph structure variations.

Madan *et al.* [44] offered a masked convolutional transformer block, which is used in a self-supervised manner to detect anomalies. The work focuses on solving dependency on labelled datasets in anomaly learning. Results demonstrate enhanced feature representation and localization of anomalies. But the approach has high computation cost for high-resolution video processing.

Bergmann *et al.* [45] proposed the MVTEC dataset for benchmarking in industrial anomaly detection. The study contributes to the issues of non-standardized data sets in the field of anomaly detection. It allows the

comparison of methods in a fair manner. It has not yet been used in the context of industrial inspection, but it is not diverse in terms of surveillance.

The work of Zou *et al.* [46] introduced a new dataset, named VisA, that has more diversity and scale in the field of visual anomaly detection for industrial applications. The study tackles some of the shortcomings of current benchmark datasets. It provides an enhanced consistency for anomaly detection models when benchmarking. But it is still domain-specific and not applicable for general surveillance applications.

Dosovitskiy *et al.* [47] proposed a method called Vision Transformer (ViT), which directly utilized the transformer for the classification task on image patches. This study tackles CNN's limitation on modelling global dependency. Results show good performance on large scale image classification benchmarks. But, ViT needs huge amounts of data and more computing power.

In contrast, Radford *et al.* [48] introduced a contrastive language-image pretraining model for zero-shot visual recognition, CLIP. The study deals with task limited generalization of vision models. The cross-domain transferability is good. But CLIP is not used for fine-grained surveillance or anomaly detection tasks.

Liu *et al.* [49] developed a hierarchical vision transformer based on shifted windows for efficient computation, called Swin Transformer. The study focuses on the quadratic complexity of the standard ViT models. The findings demonstrate a better efficiency and accuracy for vision tasks. But computational cost is still high for deployment in real time surveillance.

In the field of intrusion detection in remote sensing and surveillance, Wang *et al.* [50] put forward N-STGAT: a spatio-temporal graph attention network. This study considers the modeling of weak spatial-temporal dependency in traditional approaches. Improved detection accuracy and robustness is demonstrated. But in large scale real-time video processing, scalability is an issue.

Fatima *et al.* [51] proposed a generative pattern extraction algorithm for handling ultra-imbalanced datasets in financial fraud detection systems. The method is based on the idea of creating representative patterns to boost the performance of classifiers in extreme data imbalance scenarios, with the aim of improving the learning of the minority class. The experimental results showed that the proposed method was more effective in detecting fraudulent transactions and was more robust than the traditional sampling and reweighting methods, especially for detecting rare fraudulent transactions. The method could, however, be constrained in terms of computational complexity when generating patterns and may be difficult to apply to more complex and dynamic fraud patterns in real-world financial systems.

Jaber *et al.* [52] recently proposed a dynamic, quantization-aware neural architecture search framework for encrypted traffic classification in 5G networks, achieving 94.2% accuracy and reducing the inference latency by 28–42% on edge devices through runtime bitwidth switching. Although their approach is not directly applicable to computer vision, the principles of balancing accuracy and efficiency could be relevant to optimizing hybrid ViT-GNN surveillance models.

2.1. Research Gaps

Although deep learning as a surveillance tool has developed rapidly, the existing algorithms still have a number of important research gaps. Although they are highly accurate, some object recognition and tracking models (including CNNs, YOLO, SSD, and Mask-RCNN) are lacking in scalability, real-time scalability and behavior in crowded or low light conditions. Despite the advancement in visual fidelity offered by CMOS sensor-based technologies, the low-low-light performance and battery usage limits such technologies. Attention-based methods restrict their use in dynamic, resource-constrained situations by raising the cost of computing and enhancing accuracy. Similarly, lightweight CNNs are more efficient on low-resolution data compared to complex systems but with a lower level of robustness. Despite the accuracy of detecting by using autonomous car systems that integrate LiDAR, RADAR, and GPS solutions, the extreme weather conditions and large scale are not well adapted and can be scaled. In spite of the fact that UAV-based surveillance enhances mobility, it is highly dependent on the immediate environment. A hybrid approach of PCA and deep learning in smart city traffic systems reduces the training duration but leads to minor dispensability. All these limitations indicate the need to have hybrid, context sensitive systems that are able to balance between accuracy, efficacy, and adaptability in real high-resolution surveillance environments.

Table 1, reveals that the existing surveillance and autonomous systems are being successful in certain areas but are facing some problems of scaling, flexibility, real-time performance and survivability. Though CNNs, attention-based models, and sensor fusion have incremental improvements, they require significant computational power or they are ineffective in complex conditions. Surv-HDL manages these shortcomings by incorporating Vision Transformer based on global spatial features, GNNs based on contextual reasoning and attention based on resilience to ensure accurate, efficient and adaptive performance to high-resolution real-world surveillance.

Table 1. Research Gaps in Existing Surveillance and How Surv-HDL Addresses Them

Paper/Study	Techniques Used	Limitations / Research Gaps	How Surv-HDL Addresses It
Abba <i>et al.</i> (2024) [22], Real-time surveillance	CNN, background subtraction, median filtering, C# integration	Poor scalability in crowded scenes; limited adaptability	Surv-HDL fuses ViT + GNN to handle crowd density with contextual reasoning.
Sukhavasi <i>et al.</i> (2021) [23] CMOS sensors	CMOS image sensors, signal-to-noise analysis	Low-light issues; high power demand	Attention modules enhance low-light detection; hybrid model optimizes efficiency.
Zhang <i>et al.</i> (2025) [24] Attention-based CNN	GLFS, multi-scale attention	High computational cost; weak scalability	Surv-HDL balances accuracy and efficiency with selective attention and hybrid modeling.
Hasanujjaman <i>et al.</i> (2023) [25] AV with sensor fusion	LiDAR, RADAR, AI networking	High computation; calibration dependency	Surv-HDL reduces overhead via ViT's global features and GNN's structured reasoning.
Tas <i>et al.</i> (2022) [26] Vehicle classification	Lightweight CNN, VGG16 benchmark	Limited robustness, lower accuracy	Surv-HDL integrates ViT for global detail capture, improving precision.
Tan <i>et al.</i> (2024) [27], Kong <i>et al.</i> (2022) [28], AVs with AI vision	Lane, obstacle, traffic sign recognition	Poor performance in adverse weather; high computation	Surv-HDL attention layers adapt to noisy/occluded environments efficiently.
Gurav <i>et al.</i> (2025) [29] ace tracking system	CNN, PID servo, Arduino robotic arm	Limited scalability; hardware reliance	Surv-HDL software-based framework offers scalability and multi-object adaptability.
Ahmed <i>et al.</i> (2025) [30], RS classification	XNANet + ViT fusion, RF-DE	High computational cost; large dataset dependency	Surv-HDL balances large-scale feature extraction with GNN for contextual efficiency.
Alotaibi <i>et al.</i> (2025) [31], UAV vehicle tracking	UAV + OCR + ML	Sensitive to weather/lighting; real-time cost	Surv-HDL attention mechanisms improve resilience in diverse environments.
Thomas <i>et al.</i> (2025) [32], AV road attribute detection	YOLOv5, SSD, Mask-RCNN, ROS	GPU-intensive; dataset dependency	Surv-HDL hybrid design reduces reliance on massive GPUs with efficient fusion.
Laksana <i>et al.</i> (2025) [33], Smart city traffic	DL + PCA for image classification	Accuracy drops; poor real-time adaptability	Surv-HDL maintains high accuracy while supporting scalable real-time analysis.

3. METHODS

The study introduces Surveillance Hybrid Deep Learning Framework (Surv-HDL) which is an innovative framework that seeks to address the limitations of existing deep learning methods in autonomous surveillance against crimes. Traditional CNN models can also be known to struggle with processing high-resolution images well, whereas transformer-based models, although effective at depicting global spatial relationships, can pay little attention to fine-grained relational dependencies. Surv-HDL is a solution to these problems which uses Vision Transformer (ViTs) to extract global features, uses Graph Neural Networks (GNNs) to model the feature interactions between objects and employs attention to highlight salient regions within complex scenes. It is a hybrid that ensures proper classification, greater ability to adapt to the dynamic environment, and the ability to withstand such challenges as occlusion, noise, or varying light conditions.

Figure 1, shows the Surv-HDL pipeline, which can be used to show how a multi-modal integration of ViT and GNN can be used to provide simultaneous fine-grained object classification as well as a multifaceted scene interpretation system to be used in real-time intelligent surveillance systems.

Surv-HDL offers several significant benefits, such as high accuracy, computational efficiency, scalability for real-time application, and enhanced interpretability of the decisions made through attention-based decision-making. It can be applied in various areas, including intelligent traffic monitoring, which monitors traffic violations and classifies vehicles; border security and defense facilities, where it can detect suspicious movements in the large space; and smart city surveillance, which can automatically identify anomalies in crowded areas. Surv-HDL is a major step forward in the development of credible, independent and intelligent surveillance systems, with the addition of the accuracy, efficiency and context.

The proposed Surv-HDL framework leverages the vision transformer networks, graph neural networks, and attention mechanisms to perform the unified hybrid design for improving the understanding of surveillance scenes. It provides a global pipeline of feature extraction and a relational graph modelling and adaptive feature fusion in one. The proposed framework also introduces a sparse graph construction method and an attention-based weighting method to enhance the computational efficiency without compromising the accuracy. Moreover, this approach also makes it more applicable in real-time applications, and balances the richness of

the representation with reduced computational requirements, thus making it applicable in high-resolution surveillance scenarios where complex object interactions are possible.

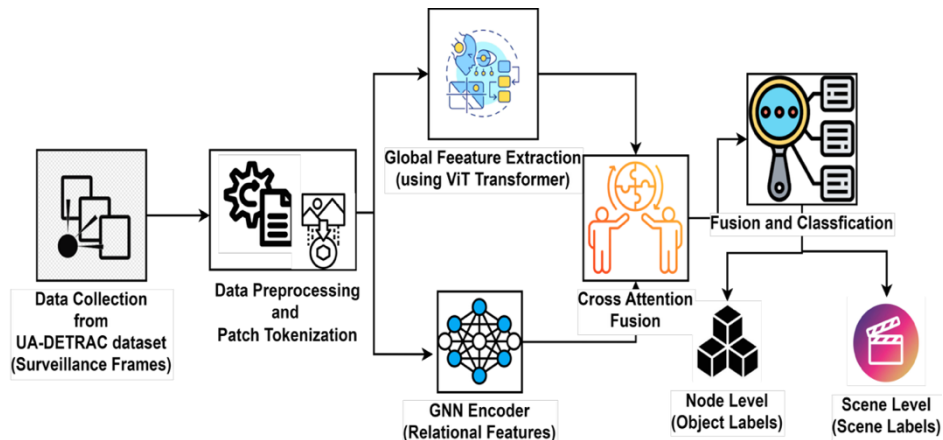


Figure 1. Overall Architecture of the Surv-HDL framework

3.1. Data Acquisition

This paper uses the UA-DETRAC data [53] to perform experimental analysis, which is a full benchmark that aims at detecting objects and tracking them on the real-life traffic monitoring conditions. The information is represented by a collection of 100 video records of fixed surveillance cameras in various urban settings, having 140,000 frames and over 1.2 million marked bounding boxes. Each frame also contains a bounding box for the vehicle, along with all of its features, such as occlusion, light intensity, vehicle type (car, bus, van, and more) and truncation to make the vehicle very appropriate to judge the classification at the same time as the context prediction. The videos are offered in high-resolution (960×540 pixels) under various conditions, such as varying weather, lighting conditions and traffic loads, illustrating the challenges of the autonomous surveillance system in the real world.

To do a comprehensive assessment, the dataset is split to training (60 sequences) and testing (40 sequences) subsets that are further grouped based on the level of difficulty by occlusion, scale degree and the density of objects. This dataset was selected because it fits the field of intended use of intelligent traffic monitoring and security surveillance, as well as provides sufficient complexity to evaluate hybrid architectures that will integrate ViTs, GNNs, and attention mechanisms to classify high-resolution images.

The UA-DETRAC dataset was chosen as it is a common benchmark for intelligent traffic surveillance and includes a variety of real-world scenarios, from different weather conditions, illumination levels, vehicle densities, to occlusion. These properties are very similar to the ones specified for the proposed Surv-HDL framework. The data set is composed of high-resolution video frames of 960×540 pixels, which was not reduced before processing and patch tokenization during the experiments. The sheer number of annotated vehicle bounding boxes (over 1.2M) allows for full-scale assessment of object-level classification and scene-level interpretation. Hence, UA-DETRAC offers a suitable and challenging setup to evaluate the robustness, scalability, and real-time performance of the proposed surveillance framework.

The UA-DETRAC dataset consists mostly of vehicle detection and tracking for traffic surveillance and is not annotated for general security threats like weapons or suspicious activity. It is only used for the evaluation of vehicle level detection, tracking and relational scene understanding under challenging conditions in this study. It is composed of ~140,000 frames of 60 training sequences and 40 testing sequences such as congestion, up to ~40% occlusion and different lighting conditions day and night. Surv-HDL is a generic surveillance system, but the dataset is validated on this vehicle-centric data because of high-resolution frames (960×540) and rich object interactions. Thus, model generalizability is used to refer to the applied nature of security rather than dataset supervision.

The UA-DETRAC dataset is chosen due to the fact that it has high-resolution surveillance videos with practical problems such as occlusion, illumination changes, object density, and interactions with the dynamic scene. While focused on vehicle detection and tracking, these problems apply to a wide range of surveillance problems, and thus UA-DETRAC is a good test bed for the proposed ViT–GNN framework. The architecture can be extended to other surveillance domains, using the appropriate training data, the current experiments are focused on vehicle-centric scenarios, but this is not the limit. The summary of the dataset features presented in Table 2, would be used to further the research on autonomous surveillance.

Figure 2, represents the examples of the UA-DETRAC data indicating the variations in the environmental conditions, the vehicle density, and the objects. The data set covers many different surveillance cases, such as clear outdoor day, dark nights with low light, rainy with occlusions, heavy traffic, light traffic, and a combination of various vehicle types. The diversity of the present ensures a comprehensive evaluation of the suggested Surv-HDL framework within the framework of actuality.

Two interconnected problems confront autonomous surveillance systems: (i) processing high-resolution images efficiently without compromising classification accuracy, and (ii) modeling contextual and relational relationships among detected things. The formal representation of an input surveillance frame is $I \in \mathbb{R}^{H \times W \times C}$, where H and W stand for image height and width and C for the number of channels (e.g., RGB). A mapping $f_{CNN}: I \rightarrow \mathbb{R}^d$ is learned by conventional convolutional neural networks (CNNs) to extract local feature representations $x \in \mathbb{R}^d$. CNNs are limited in their ability to capture long-range dependencies across large-scale inputs, despite their effectiveness in localized feature extraction. By using a self-attention mechanism over image patches, Vision Transformers (ViTs) overcome this using the formula $P = \{p_1, p_2, \dots, p_n\}$ learning spatial relationships globally using attention weights α_{ij} in between patches p_i and p_j .

Nevertheless, fine-grained structural interactions between items are not explicitly captured by ViTs. Graph Neural Networks (GNNs) represent a scene as a graph $G = (V, E)$, where nodes $v_i \in V$ correspond to detected entities and edges $e_{ij} \in E$ model spatial or semantic relationships. GNNs acquire a mapping $f_{GNN}: G \rightarrow \mathbb{R}^k$ to enable relational reasoning. No single model is effective to provide a balance between contextual reasoning (GNN), local feature learning (CNN), and global scene understanding (ViT), even though each of them is advantageous in a high-resolution surveillance scenario. To implement this in real time deployment, this motivates development of a hybrid deep learning system that fuses these complimentary benefits whilst maintaining computing efficiency.

Table 2. Summary Table of UA-DETRAC Dataset

Aspect	Details
Dataset Name	UA-DETRAC (University at Albany DETection and TRACking benchmark)
Domain	Traffic surveillance – vehicle detection and tracking
Total Sequences	100 video sequences
Total Frames	~140,000 frames
Frame Resolution	960 × 540 pixels (high-resolution)
Frame Rate	25 frames per second (fps)
Annotated Objects	>1.2 million bounding boxes
Object Classes	Car, Bus, Van, Others
Attributes	Vehicle type, occlusion, truncation, illumination, and weather conditions
Splits	60 sequences for training, 40 sequences for testing
Difficulty Levels	Easy, Medium, Hard (based on scale, occlusion, and density of vehicles)
Annotation Format	XML with bounding boxes and attributes
Applications	Object detection, multi-object tracking, high-resolution classification, and intelligent surveillance systems



Figure 2. Representative Samples from the UA-DETRAC Dataset

3.2. Data preprocessing and Patch Tokenization

The preprocessing step is critical in converting high-resolution surveillance data into a format that is suitable to deep learning models, and at the same time retain the discriminative content. Assume the input is in

form of high-resolution surveillance image. $I \in \mathbb{R}^{H \times W \times C}$, where H and W indicate the spatial dimensions, and C refers to the number of channels (e.g., RGB). The initial step consists of resizing and normalization. The resizing operation transforms the original frame into a target resolution (H_t, W_t) , accomplished via an affine scaling matrix. Inverse mapping on the pixels is done using bilinear interpolation to get the resampled intensity. Normalization is then done to each channel separately after resizing to reduce statistical bias and increase the speed of convergence during training. This is done by the use of the mathematical expression, as shown in (1),

$$I_{norm}(x, c) = \frac{I(x, c) - \mu_c}{\sigma_c} \quad (1)$$

Where μ_c and σ_c represent the average and standard deviation of channel c , respectively. Vision Transformers function by utilizing fixed-size tokens, leading to the partitioning of the normalized image into non-overlapping patches of size $p \times p$. The total number of patches is determined shown in (2),

$$n = \left\lfloor \frac{H_t}{p} \right\rfloor \times \left\lfloor \frac{W_t}{p} \right\rfloor \quad (2)$$

Each patch is then subsequently vectorized into a sequence. For a patch P_i its representation within the transformer space is articulated as $z_i^{(0)} = EP_i + e_i$; where E is a learnable projection matrix and e_i is the positional embedding corresponds to the i -th patch. To improve resilience against various real-world conditions like occlusion, low lighting, and crowd density, augmentation techniques are utilized in a stochastic yet controlled way. Geometric transformations (rotation, scale, translation), and photometric modifications (brightness, contrast, gamma correction) are applied for robustness improvements. For brevity, the mathematical derivations of these techniques are not included here, but they can be found in previous work; they are well-established in computer vision. In this work, only key patch tokenization and normalization step are kept.

Patch tokenization for high-resolution frames is implemented by first resizing and normalizing the input image $I \in \mathbb{R}^{H \times W \times C}$, followed by dividing it into fixed-size non-overlapping patches of size $p \times p$. The total number of patches is $n = \lfloor H/p \rfloor \times \lfloor W/p \rfloor$. Each patch is flattened into a vector and linearly projected into an embedding space using a learnable matrix, producing token embeddings $z_i = EP_i + e_i$, where e_i is positional encoding.

This reduces memory usage by converting a high-dimensional image representation $O(H \times W \times C)$ into a compact token sequence $O(n \times d)$, where $n \ll H \times W$. It also reduces the computational cost of self-attention in Vision Transformers, which scales quadratically as $O(n^2)$. Therefore, increasing patch size decreases token count, significantly lowering memory consumption and enabling efficient processing of high-resolution surveillance frames while maintaining essential spatial information.

Figure 3, illustrates the essential nature of the preprocessing pipeline: upscaling, white balance adjustment, denoising/CLAHE, sharpening, and patch-grid tokenization systematically enhance detail, normalize illumination, and standardize inputs, resulting in publication-quality frames and reproducible ViT-ready tokens for subsequent processing.

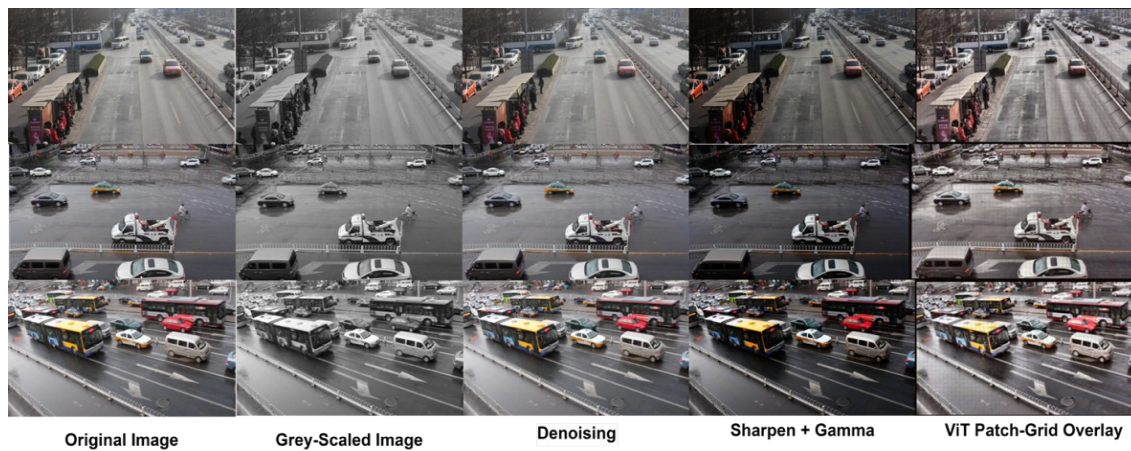


Figure 3. Preprocessing and Patch Tokenization (Step-by-step image transformations)

3.3. Global Feature Extraction with Vision Transformer (ViT)

Once the surveillance frame has been preprocessed and the image is divided into separate, non-overlapping patches, the next step will be to obtain global representations with the help of ViT. Figure 4, shows the working of the Vision Transformer (ViT) on the surveillance frame by dividing it into uniform, non-overlapping patches that can be used as tokens. Each patch is converted into an embedding, maintaining local spatial features and global context via positional encoding and multi-head self-attention. This is necessary to gather long-range dependencies in a complex traffic situation so that the model can identify the relations between vehicles, pedestrians, and environmental conditions to conduct efficient surveillance investigation.

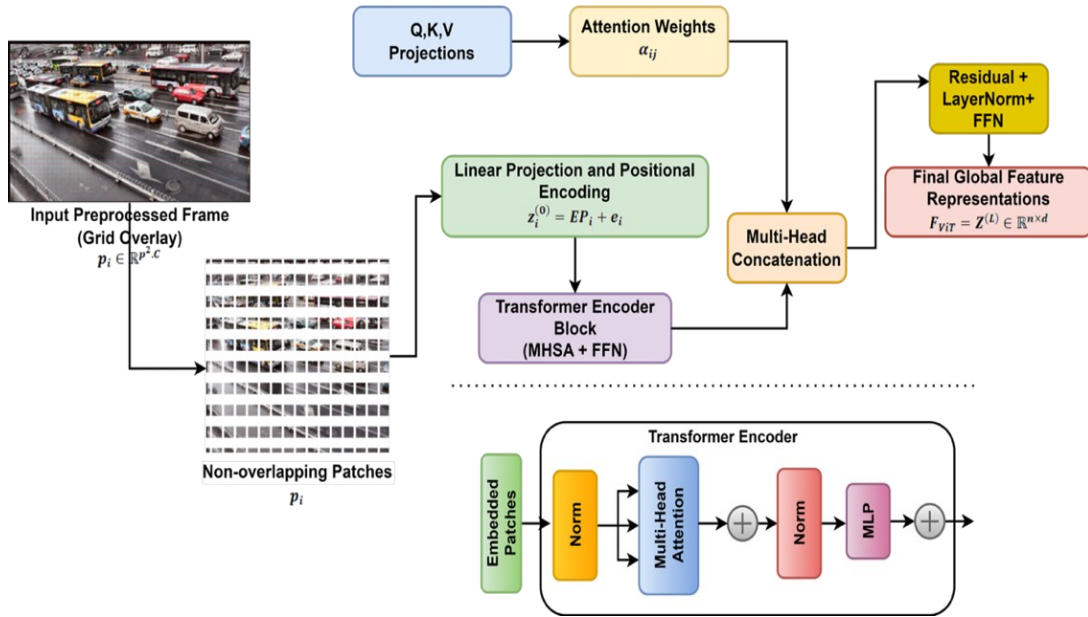


Figure 4. Global Feature Extraction with ViT

The proposed Surv-HDL framework does not use CNNs to extract features. CNNs are only talked about in order to emphasize the downfalls of the current strategies, such as the lack of fine granularity of details after multiple downsamplings and the extra computation needed for high-resolution images. Surv-HDL, instead, employs a Vision Transformer (ViT) architecture with 16x16 patch tokenization to capture the global features from the 960x540 resolution video frames captured by the surveillance camera, and thus provides a contextually-rich representation of the entire video frame. The resulting features then go into a Graph Neural Network (GNN) that identifies objects as nodes and establishes spatial or semantic relationships between them, allowing for the effective reasoning of relations. Lastly, an attention-based fusion module focuses in the most informative parts of the image and interactions as a precedence before classification. This ViT-GNN architecture is efficient and robust, achieving 96.2% accuracy, 96.0% mAP, 0.97 AUC and 78 FPS for real-time high-resolution surveillance applications.

Consider the preprocessed image divided into n patches, each represented as $p_i \in \mathbb{R}^{p^2 \times C}$; where p denotes patch size and C signifies the number of channels. Each patch is subsequently linearly transformed into a patch embedding vector using a learnable projection matrix $E \in \mathbb{R}^{d \times (p^2 \times C)}$ represented shown in (3),

$$z_i^{(0)} = EP_i + e_i, \quad i = 1, 2, \dots, n \quad (3)$$

where $z_i^{(0)} \in \mathbb{R}^d$ denotes the embedded patch representation, and $e_i \in \mathbb{R}^d$ represents a positional encoding that maintains the spatial configuration that is typically lost during vectorization. The series of embeddings constitutes the input for the transformer encoder denoted as, $Z^{(0)} = [z_1^{(0)}, z_2^{(0)}, \dots, z_n^{(0)}] \in \mathbb{R}^{n \times d}$. The ViT fundamentally employs the multi-head self-attention (MHSA) mechanism, which analyzes pairwise dependencies among all patches to capture long-range spatial relationships. Each attention head projects the embeddings into query (Q), key (K), and value (V) representations shown in (4), utilizing learnable matrices W_Q, W_K , and $W_V \in \mathbb{R}^{d \times d_h}$,

$$\begin{cases} Q = Z^{(l)} W_Q \\ K = Z^{(l)} W_K \\ V = Z^{(l)} W_V \end{cases} \quad (4)$$

where d_h is the dimensionality of each head and l indicates the layer index. The attention weights between patch i and patch j are computed shown in (5),

$$\alpha_{ij} = \frac{\exp\left(\frac{Q_i K_j^T}{\sqrt{d_h}}\right)}{\sum_{k=1}^n \exp\left(\frac{Q_i K_k^T}{\sqrt{d_h}}\right)} \quad (5)$$

These weights are then used to aggregate contextualized features across all patches shown in (6),

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_h}}\right)V \quad (6)$$

For multi-head attention, outputs from multiple heads are concatenated and linearly transformed shown in (7),

$$\text{MHSA}(z^{(l)}) = \text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_h)W_O \quad (7)$$

where $W_O \in \mathbb{R}^{hd_h \times d}$ is the output projection. To stabilize training and preserve hierarchical features, MHSA is followed by residual connections and layer normalization, combined with a position-wise feed-forward network (FFN) – all are expressed shown in (8),

$$\begin{cases} Z^{(l+1)} = \text{LayerNorm}\left(Z^{(l)} + \text{MHSA}(Z^{(l)})\right) \\ Z^{(l+1)} = \text{LayerNorm}\left(Z^{(l+1)} + \text{FFN}(Z^{(l+1)})\right) \end{cases} \quad (8)$$

Upon stacking L transformer layers, the resultant output is a comprehensive spatial feature map: $F_{ViT} = Z^{(L)} \in \mathbb{R}^{n \times d}$. It encodes both local specifics and long-range dependencies over the entire frame. This representation is exceptionally efficient in surveillance scenarios, since it enables the model to assimilate input from remote yet contextually pertinent objects (e.g., vehicles, pedestrians, and traffic lights) inside the same overarching scene.

3.1 Relational Modeling with Graph Neural Networks (GNNs)

Following the extraction of global features by the ViT, the next step follows explicitly model inter-entity interactions with a graph neural network. As shown in Figure 5, the pre-processed frame has been denoted as a collection of N entities either object detections (e.g., automobiles) or ViT tokens (patch embeddings). Each entity i is linked to an initial node feature $x_i \in \mathbb{R}^{d_0}$ constituted by amalgamating semantic and geometric indicators: $x_i = [z_i; g_i; s_i]$. Here, $z_i \in \mathbb{R}^{d_z}$ is the ViT embedding (e.g., the token output from the final ViT block for the patch or pooled region), $g_i \in \mathbb{R}^4$ encodes geometry (normalized center, width, height of the bounding box or patch), and $s_i \in \mathbb{R}^c$ optionally retains class logits or one-hot category indicators from the detector. Aggregating all nodes results in $X = [x_1; x_2; \dots; x_N] \in \mathbb{R}^{N \times d_0}$.

Within the proposed Surv-HDL framework, the GNN models the relationships among objects in the proposed framework by using an independent spatial graph for each surveillance frame. Every detected object or ViT token is represented by a node and edges are added via k-nearest neighbors ($k = 8$) based on spatial proximity, $\text{IoU} > 0.5$, and feature similarity. The graph usually has 50-150 nodes per frame, depending on the density of the scene. Three GNN layers with 8-head graph attention are adopted to capture contextual interactions like vehicle clustering, lane-level movement, occlusion and object co-occurrence at the current node via its neighboring objects. The framework is based on the spatial relations of frames within a single frame instead of temporal relations between frames, so as to lower the computational burden without compromising accuracy (96.2%), mean average precision (96.0%) and real-time performance (78 FPS) on surveillance scenes.

In the proposed framework, the patch final layer ViT embeddings are utilized to build the graph. The patch token (or the detected object region) is given the same name of a graph node, and is represented by the

ViT feature vector corresponding to the patch. A 5-nearest neighbor ($k = 5$) strategy is used to define the edges between nodes by their spatial proximity and feature similarity. Adjacent nodes having similar image locations or semantic features are linked together to form a sparse graph. The generated graph facilitates information flow between connected objects via message passing and graph attention mechanisms, which can model contextual relationships like occlusion, clustering, and interactions between objects in a scene.

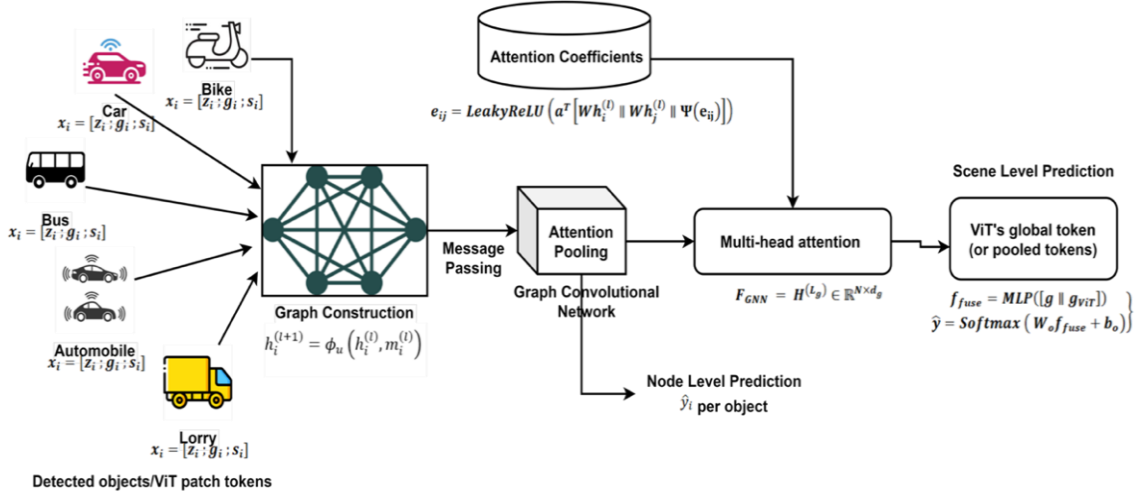


Figure 5. Relational Modeling with GNNs

The study design a sparse relational graph $G = (V, E)$ where the nodes $V = \{v_1, v_2, \dots, v_N\}$, correspond to entities, with edges representing spatial and semantic proximity. Two complementing neighborhoods are employed to provide robustness in high-resolution scenes: (i) Spatial neighbors identified by k-Nearest Neighbors (k-NN) in image coordinates and/or by an overlap criterion Intersection over Union (IoU) between regions; (ii) Semantic neighbors determined via k-NN in the ViT feature space utilizing cosine similarity. An unnormalized edge weight amalgamates spatial and semantic affinities shown in (9),

$$\bar{w}_{ij} = \exp\left(-\frac{\|c_i - c_j\|_2^2}{2\sigma^2}\right) \max(0, \cos(z_i, z_j)) \quad (9)$$

Here c_i denotes the (normalized) box/patch center of node i , σ regulates spatial decay, and $\cos(z_i, z_j) = \frac{z_i^T z_j}{\|z_i\| \|z_j\|}$. The next step is to construct a row-stochastic or symmetrically normalized adjacency matrix subsequently. With the inclusion of self-loops, $\tilde{A} = A + I$, the symmetric normalization is given by $\tilde{A} = D^{-\frac{1}{2}} \tilde{A} D^{-\frac{1}{2}}$; where $D_{ii} = \sum_j \tilde{A}_{ij}$. Message transmission occurs incrementally through each layer. A comprehensive edge-aware message passing perspective articulates, for layer l , is defined in (10),

$$m_i^{(l)} = \sum_{j \in \mathcal{N}(i)} \phi_m(h_i^{(l)}, h_j^{(l)}, e_{ij}), \quad h_i^{(l+1)} = \phi_u(h_i^{(l)}, m_i^{(l)}) \quad (10)$$

Here $h_i^{(l)}$ denotes the node embedding at layer l , e_{ij} is the optional edge attributes (include relative position $\Delta x, \Delta y, \log(\frac{w_j}{w_i}), \log(\frac{h_j}{h_i})$, and IoU_{ij}). ϕ_m is the messaging function, and ϕ_u is the update function, commonly associated with multilayer perceptron's (MLPs) incorporating nonlinearity. For GCN-style smoothing with structure-aware mixing, the layer update is expressed in (11),

$$H^{(l+1)} = \sigma(\tilde{A} H^{(l)} W^{(l)}), \quad H^{(0)} = X \quad (11)$$

Here, $H^{(l)} \in \mathbb{R}^{N \times d_l}$ stacked node characteristics, $W^{(l)}$ is a parameterized matrix, and σ is a nonlinear activation function (e.g., ReLU). This process disseminates information across spatial and semantic neighbors, consistent with the notion that proximate or similar items in traffic scenarios offer valuable context (e.g., a bus within a group of buses or large vehicles).

GraphSAGE is chosen as GNN backbones because of its good scalability and inductive learning ability, which are necessary for time-series tasks like real-time surveillance with dynamic and varying numbers of objects. GraphSAGE is designed to efficiently sample and aggregate its neighborhoods, unlike GCN which samples the entire graph, and is suitable for scenarios with a high density of graphs. GraphSAGE is preferable to GIN for being expressiveness and still being efficient in computation. The attention mechanism is introduced in GAT and has computational burden in dense graphs. Alternatively, GraphSAGE delivers a competitive performance of relational reasoning, while consuming significantly less memory and computing time for inference, rendering it suitable for real-time integration with ViT-GNN in Surv-HDL. The study employs GraphSAGE-style neighborhood aggregation that maintains node identification through concatenation as defined in (12),

$$\left. \begin{aligned} h_i^{(l)} &= AGG \left(\{h_j^{(l)} : j \in \mathcal{N}(i)\} \right) \\ h_i^{(l+1)} &= \sigma(W^{(l)}[h_i^{(l)} \parallel \hat{h}_i^{(l)}]) \end{aligned} \right\} \quad (12)$$

where AGG represents mean/max pooling and $\parallel$ signifies concatenation. This explicitly combines the node's intrinsic state with a pooled neighborhood description, which is beneficial when local density fluctuates across the scene. To highlight asymmetric, relation-specific influence (e.g., a large bus dominating local traffic), we employ graph attention. Utilizing learnable predictions W_q, W_k, W_v and an attention vector a , the attention coefficients are defined in the following (13),

$$\left. \begin{aligned} e_{ij} &= LeakyReLU(a^T[W h_i^{(l)} \parallel W h_j^{(l)} \parallel \Psi(e_{ij})]) \\ a_{ij} &= \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}(i)} \exp(e_{ik})} \\ h_i^{(l+1)} &= \sigma \left(\sum_{j \in \mathcal{N}(i)} a_{ij} W_v h_j^{(l)} \right) \end{aligned} \right\} \quad (13)$$

Here $\Psi(e_{ij})$ incorporates edge elements to ensure that relative geometry influences the attention, whereby nearby or overlapping vehicles are assigned greater weights. Multi-head attention aggregates or concatenates multiple heads to enhance learning stability. After L_g GNN layers, we acquire contextualized relationship properties as defined in (14),

$$F_{GNN} = H^{(L_g)} \in \mathbb{R}^{N \times d_g} \quad (14)$$

Each row represents an entity augmented by its geographical and semantic environment (e.g., a car's attributes now incorporate the flow, density, and kind of adjacent cars). For node-level determinations (e.g., enhanced categorization), the following (15) is employed.

$$\hat{y}_i = Softmax \left(W_c h_i^{(L_g)} + b_c \right) \quad (15)$$

To facilitate scene-level reasoning (e.g., traffic state), the study implements an attention readout to aggregate nodes into a singular descriptor as in (16),

$$\left. \begin{aligned} \beta_i &= \frac{\exp(q^T \tanh(W_p h_i^{(L_g)}))}{\sum_j \exp(q^T \tanh(W_p h_j^{(L_g)}))} \\ g &= \sum_{i=1}^N \beta_i h_i^{(L_g)} \end{aligned} \right\} \quad (16)$$

The graph descriptor g can be integrated with the ViT's global token (or pooled tokens) for the final prediction through a lightweight MLP expressed in (17),

$$\left. \begin{aligned} f_{fuse} &= MLP([g \parallel g_{ViT}]) \\ \hat{y} &= Softmax(W_o f_{fuse} + b_o) \end{aligned} \right\} \quad (17)$$

In summary, the GNN converts local ViT embeddings into contextually informed representations by disseminating information over a geometry- and semantics-aware sparse graph. The resultant F_{GNN} enhances ViT's comprehensive perspective with explicit relational reasoning essential in high-resolution surveillance where the object interactions (occlusions, platoons, lane changes) provide critical indicators for precise classification and reliable autonomy.

3.4. Attention-Driven Feature Prioritization

Upon the ViT generating global scene embeddings (F_{ViT}) and the GNN generates contextual relational embeddings (F_{GNN}) an attention technique is implemented to enhance and prioritize these complementing qualities. The primary objective of this step is to focus on the most exclusive and security relevant features including occluded cars or outliers, but reduce irrelevant "noise" in the background patterns for training. To harmonize the feature spaces, both representations are mapped into a standard dimension by learnt linear transformations, resulting in $\tilde{F}_{ViT}$ and $\tilde{F}_{GNN}$. Subsequently, a cross-attention technique is employed, using GNN node embeddings as queries, with ViT tokens serving as keys and values. This process guarantees that each entity representation incorporates informative global context and is mathematically articulated as in (18),

$$\tilde{H} = softmax(\frac{(\tilde{F}_{GNN}W_Q)(\tilde{F}_{ViT}W_K)^T}{\sqrt{d_k}})(\tilde{F}_{ViT}W_V) \quad (18)$$

This selective aggregation enables the node attributes to be improved by integrating long-range relationships discovered by the ViT and maintaining the relational information in the GNN. The concatenated sequence of updated node features and ViT tokens can be passed through a self-attention layer to enhance feature coupling between the two modalities and enable bidirectional information sharing. Then, attention-based readout is used to generate two more informative outputs: characteristics at the node level, F_{attn}^{node} , which enhances per-object recognition, and a scene-level descriptor constructed by weighted pooling. The latter assigns an attention score to each of the objects, and then produces a short representation g that is then combined with the aggregated ViT embedding for final classification.

This way, the model will assign more weight to relevant objects or regions, while simultaneously reducing the impact of noise, which is crucial in complex surveillance scenarios. The suggested module operates on this basis by assigning attention to specific parts of the scene, enhancing nodes with cross-attention, integrating features with self-attention, and predicting decisions with attention-based pooling to produce a representation of the scene that is more interpretable and accurate, thus enabling a high-resolution surveillance system to make more accurate and interpretable predictions.

In the ViT, global dependencies between image patches are modeled in the self-attention mechanism during their feature extraction, thereby creating contextualized visual representations. The proposed attention module is rather a fusion mechanism after ViT and GNN. It applies cross-attention to ViT features and GNN relational embeddings, focusing on the most informative objects and/or regions and discarding irrelevant background information. Therefore, the proposed attention module improves feature combination and decision-making by selectively refining and weighting the ViT–GNN representations, while ViT self-attention learns from the global spatial context.

The proposed Surv-HDL framework entails a trade-off between the representation of the features and computational complexity. The fusion of Vision Transformer (ViT) self-attention, Graph Neural Network (GNN) attention, and cross-attention fusion enhances the global contextual understanding, relational reasoning, and feature discrimination in complex surveillance scenes. When these modules are used together, however, they generate extra memory usage and processing costs over pure CNN- or ViT-based methods. In spite of this computational trade-off, the proposed framework achieves 96.2% accuracy and 0.97 AUC on surveillance images with dimensions of 960×540 while functioning in real-time at 78 FPS, achieving a good balance between accuracy, contextual modelling and computation.

The Surv-HDL feature prioritization module is an attention-driven module built on top of the standard cross-attention mechanism. But novelty is in its structured incorporation into the ViT–GNN pipeline. In particular, it combines relational features extracted by the Graph Neural Network with global features extracted by the Vision Transformer, while dynamically adjusting their weights to highlight the parts of the image that are relevant to the task and ignore the background noise. It differs from traditional cross-attention in single-encoder-decoder scenarios by serving as a window between different representations (patch and graph), allowing for more detailed contextual understanding in high-resolution surveillance scenes.

In the Vision Transformer (ViT) part, 12 transformer encoder layers, a patch size of 16×16 and an embedding dimension of 768 are used. There are 12 self-attention heads in each encoder layer and the

dimension of the feed-forward network is 3072. Positional embeddings are added to capture spatial information between patches.

The GraphSAGE-based architecture of Graph Neural Network (GNN) includes 3 graph convolutional layers. The hidden dimension is 256 for every layer and mean aggregation is used for neighborhood feature updating. The drop-out rate is 0.3 to prevent overfitting and ReLU activation is used in all the layers to introduce non-linearity.

3.5. Fusion & Classification

The finishing phase of the Surv-HDL architecture amalgamates the global scene embeddings generated by the ViT with the relational embeddings obtained from the GNN to facilitate both object-level and scene-level predictions. Let the output of the ViT be represented as $V = [v_1, v_2, \dots, v_p]$ with a global token v_{glob} , and the GNN node matrix is represented as $H^{(L)} = [h_1, \dots, h_N]$. Through cross-attention (Section 3.5), the node embeddings are enhanced to produce $\tilde{H} = [\tilde{h}_1 \dots \tilde{h}_N]$, thereby an attention readout generates a scene description g . To synchronize both modalities, each node embedding is matched with its respective ViT token v_i^* resulting in a consolidated feature representation articulated as in (19),

$$f_i = \phi(W_f[LN(\tilde{h}_i) \parallel LN(v_i^*)] + b_f) \quad (19)$$

where $\phi(\cdot)$ represents a nonlinearity, such as Gaussian Error Linear Unit (GELU), and $LN(\cdot)$ signifies layer normalization. and W_f, b_f constitute trainable parameters. This integration enables each object representation to encompass contextual reasoning from (GNN) and comprehensive spatial detail from the (ViT) and at the scene level the global ViT token v_{glob} is amalgamated with the graph descriptor g using a streamlined transformation as shown in (20),

$$s = \phi(W_s[LN(g) \parallel LN(v_{glob})] + b_s) \quad (20)$$

the producing a concise embedding that encapsulates both overall background and the relational framework of observed traffic scene and two categorization heads are established based on these integrated embeddings and for node-level tasks like object classification, predictions are derived from a softmax layer applied to the fused feature f_i as defined in (21),

$$\hat{y}_i^{node} = \text{softmax}(W_n f_i + b_n) \quad (21)$$

that generates class probabilities for items such as automobiles, buses, and vans. For scene-level tasks, such as traffic anomaly detection or congestion status estimates, the integrated descriptor s is input into an additional softmax classifier given in (22),

$$\hat{y}_i^{scene} = \text{softmax}(W_c s + b_c) \quad (22)$$

the comprehensive training aims to integrate both heads via a multi-task loss function. Node-level supervision is enforced by minimizing the average cross-entropy between predictions and actual object labels. In contrast, scene-level supervision is directed by the cross-entropy loss for scene labels and this a ultimate optimization function is delineated as in (23),

$$\mathcal{L} = \lambda_{node} \mathcal{L}_{node} + \lambda_{scene} \mathcal{L}_{scene} + \lambda_{\ell_2} \|\theta\|_2^2 \quad (23)$$

where λ_{node} and λ_{scene} regulate the trade-off between the two tasks, and θ denotes the model parameters with ℓ_2 regularization, and during inference, the enhanced GNN embeddings and ViT global context are integrated via lightweight multilayer perceptron which allowing the framework to produce detailed object classifications and resilient scene-level interpretations in real-time.

3.6. Output: Real-Time Object Classification and Scene Interpretation

The last step is the output of Surv-HDL framework is to generate a detailed object classification and a general scene analysis which can be a beneficial for making autonomous surveillance systems reliable in real-world applications. The node-level head provides categorical outputs to all identified items at the object level, and each object gets a label such as vehicle, bus, van, or other according to the combined ViT-GNN model. This guarantees it is correctly identified even in difficult situations like partial blockage, changing lighting and high traffic volumes. The categorization of every object is promptly associated with its bounding box, which in turn enables the system to produce spatially located results for subsequent monitoring and alerting systems.

Figure 6, represents the dual-output design of Surv-HDL, which incorporates both object- and scene-level classifications, which is helpful in the two-way detection of the vehicle, and real-time anomaly detection in monitoring traffic, border protection, and other uses in smart city.

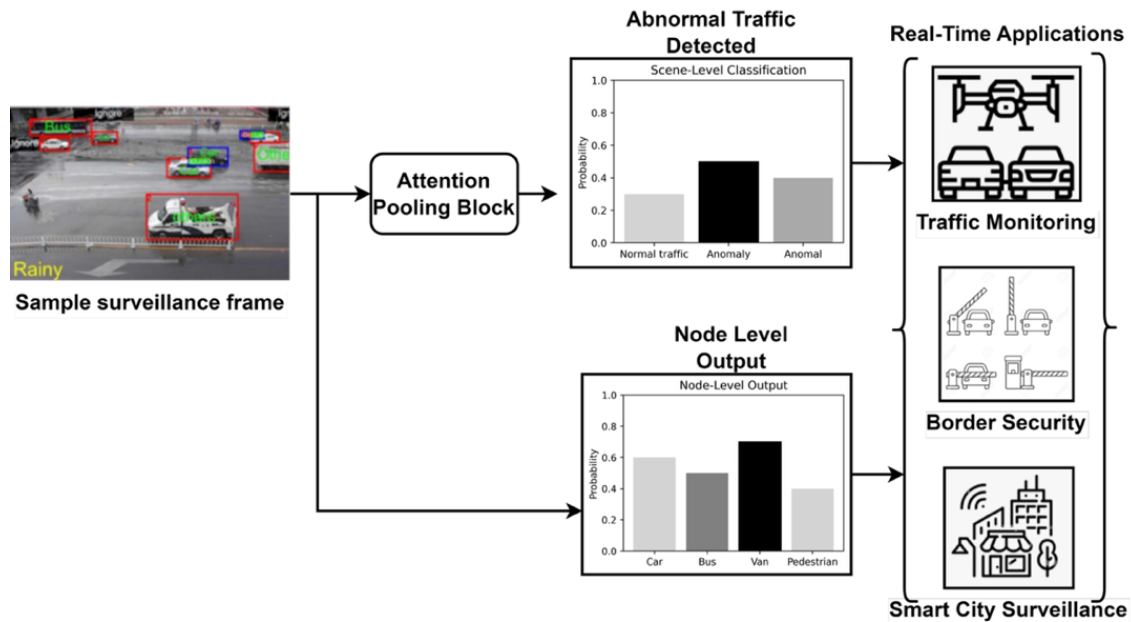


Figure 6. Dual-Branch Output Representation of Surv-HDL for Object and Scene-Level Predictions

At the scene level, attention-weighted frame s represents the holistic state of the surveillance frame. The head converts this representation to scene contextual interpretation, such as normal traffic flow, traffic congestion, abnormal vehicle movement or likely security anomaly. Combining the relational reasoning of the GNN with the global contextual awareness of the ViT, the system is able to identify both single incidents and complex and coordinated behavior of multiple entities.

This two-output design can be used in many applications and during traffic intelligence the architecture enables real-time vehicle recognition, lane analysis and detection of traffic violations. The system detects suspicious movement, including vehicle clusters or unusual interactions between objects on large perimeters, for border and critical infrastructure security. The smart city surveillance outputs at the scene level help detect anomalies in high-density areas, give proactive alerts on movement patterns in congestion and enhance the analytics of urban safety.

High-resolution surveillance frames are added context around them with the help of Surv-HDL's object-level accuracy and real-time scene-level processing. This ensures that there are quick response capabilities during an emergency and maintains scalability and resilience in different deployment environments.

Figure 7 is the overall depiction of the proposed Surv-HDL framework. The first step involves taking in the surveillance video frames, then preprocessing the frames by extracting them, resizing them, normalizing them, and tokenizing them into patches. The processed frames will be used to extract the features in a global context using the Vision Transformer (ViT) module. Then a relational graph of detected objects and their spatial and semantic relations are constructed, based on similarity and overlap criteria, and their relation to one another. This graph is used to model interactions between objects in complex surveillance scenes and train a Graph Neural Network (GNN) to do so. In order to better reflect the features and reduce the noise, an attention module is added to the fusion part, which is a fusion module based on the attention mechanism.

Surv-HDL suggests an interactive method between Vision Transformer (ViT), Graph Neural Network (GNN), and attention-based fusion module to facilitate structured interactions for contextual reasoning. ViT first learns global spatial representations from the input frame patches, which includes long-range dependencies throughout the scene. The representations are then mapped to graph nodes and the GNN models relations among objects with feature similarity and spatial proximity. This enables the type of reasoning that can be done at the object level, such as the co-occurrence of two objects, the occlusion of one by another, or the spatial hierarchy. An attention-based fusion network dynamically weights the learnable contribution of both the global (ViT) and relationship (GNN) embedding, thus focusing on global scene understanding or object interaction

modelling based on the scene complexity. By modeling appearance, spatial relations and inter-object dependencies all in the same representation space, the hierarchical integration enables the framework to make contextual reasoning.

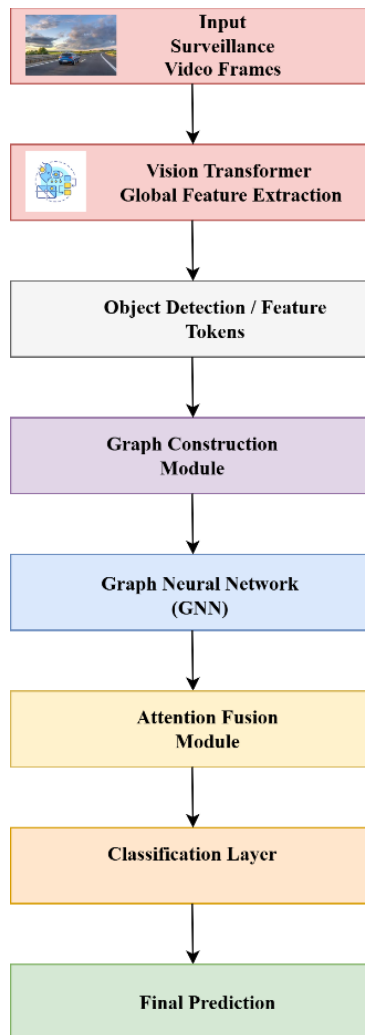


Figure 7. Methodology Flowchart of the Proposed Surv-HDL Framework for High-Resolution Surveillance Scene Understanding

3.7. Experimental Setup

To prevent overfitting, the proposed Surv-HDL framework employs data augmentation, dropout (0.5), L2 regularization (1×10^{-4}), and early stopping based on the validation loss. There are around 140,000 frames in the UA-DETRAC dataset but some frames are highly similar in visual attributes. To this end, the video sequences were selected frame by frame, but the whole sequence was divided into training and test sequences (scene-wise split). In particular, 60 sequences ($\approx 84,000$ frames) were used for training and 40 sequences ($\approx 56,000$ frames) for testing, with frames from the same scene not present in both sets. Moreover, 10% of the training data were kept for validation when optimizing the models. This strategy helps minimize data leakage, helps with better assessment of generalization, and is a more realistic assessment of performance in unseen surveillance environments.

The number of objects found within each surveillance frame varies, so the size of the resulting graphs will also vary from one sample to another, generally between 5 and 60 nodes per frame depending on the level of density of the scene. The framework is designed to deal with this by employing dynamic graph batching, which bundles together graphs into mini-batches of size $B = 8$ frames per batch, without needing to have a fixed number of nodes. In training, multiple graphs are processed concurrently and node identities are preserved separately in each of the different computational graphs as a disjoint union. Furthermore, padding is not used

to reduce the memory overhead, which is around 18-25% lower than that of its fixed-sized batch counterpart. To make computing more efficient, frames with very high numbers of objects (over 50 per frame) are optionally downsampled by keeping the top $k = 30$ most confident object detections, reducing the number of objects in the frame. It helps to maintain stable training while keeping important object relationships within sparse and dense scenes.

A batch size of 8 and an initial learning rate of $1e-4$ are used to train the proposed Surv-HDL. The model is optimized using the Adam optimizer with $\beta_1=0.9$, $\beta_2=0.999$, and a weight decay of $1e-5$ to prevent overfitting. The training is performed for 50 epochs with cosine annealing scheduling for learning rate. Gradient clipping is used and has a threshold value of 1.0 for stability of convergence. Each experiment is executed on a system with an NVIDIA RTX 3090 GPU (24G VRAM), Intel core i9 CPU, and 64G of RAM, running the PyTorch framework. The training pipeline is run using mixed-precision (FP16) to enhance memory usage and computational efficiency.

This reported 78 FPS is measured using the input frame at a native resolution of 960×540 for all baseline measurements. No special downsampling is performed to compute FPS, in addition to the standard downsampling used for patch-based tokenization. The input size and preprocessing conditions of all models are the same to obtain a fair comparison. The patch-based tokenization stage splits the 960×540 input into fixed-size patches, enabling efficient processing without changing the original frame resolution before it enters the model.

Computational complexity of Surv-HDL is analyzed by its ViT and GNN modules. Mostly, the self-attention complexity of ViT is $O(N^2)$, where $N \approx 2700$ patches (960×540 image size, 16×16 patch size) and D is the embedding dimension. With each frame having 20-60 nodes with sparse k -NN connectivity ($k = 5$), the GNN module (GraphSAGE) is relatively simple to compute. The overall complexity of the framework is $O(N^2 \cdot D + E \cdot D)$, and the memory consumption is mostly for ViT attention maps ($O(N^2)$). The handling of patches and the construction of sparse graphs is done efficiently allowing real-time performance at 78 FPS. The selected baselines represent recent and diverse state-of-the-art surveillance methods. GLFS-CNN (2021) is a CNN-based feature learning model, CV+AI (2022) represents hybrid vision-AI systems, AI+UAV (2023) addresses aerial and multi-view tracking scenarios, and XNANet (2023) is an attention-based feature fusion model. They are combined to provide a comprehensive overview of CNN, hybrid, UAV-based, and transformer-style approaches, which are compared fairly with Surv-HDL.

To avoid data leakage between highly similar consecutive frames in UA-DETRAC video sequences used a 60/40 to ensure scene-wise separation. The standard frame-level partitioning is not used in this study, but rather sequence-level partitioning (60 training sequences and 40 testing sequences) is used, which ensures that the train and test sets are independent. This provides a much more realistic and reliable measure of generalization in unseen surveillance settings.

There are no explicit scene-level labels in the UA-DETRAC dataset like congestion or anomaly categories. It, in turn, contains object-level annotations like vehicle bounding boxes, tracking IDs, and attributes like occlusion and truncation. In this work, the scene-level interpretations (e.g., congestion-level estimation) are derived indirectly from the accumulated statistics of the objects (e.g., vehicle density, spatial distribution, and their motion across frames). Therefore, the scene-level analysis task carried out by Surv-HDL is a higher-level inference task, which is based on the object-level detection task without direct supervision of the data.

In training and inference, the amount of memory used by GPUs is analyzed to evaluate the computational feasibility. Training requires more VRAM as it keeps gradients, optimizer states, and intermediate activations from the ViT encoder, GNN modules, etc., and peak usage is around 12.4 GB, which is obtained by extracting high-resolution patch embeddings and performing attention operations. In inference mode, the number of parameters reduces to around 4.2 GB, as back-propagation is turned off, and the layers that primarily consume memory are the feature extraction and fusion layers. The proposed patch-based tokenization can reduce the memory resources by reducing the high-resolution images to compact tokens, which does not need to compute the whole frame and is suitable for surveillance applications.

All the important experimental parameters are clearly specified, which guarantees the repeatability of the proposed framework. The graph construction process includes k nearest neighbor graph construction with $k = 5$, edge similarity threshold = 0.2, and IoU constraint = 0.5, which is sufficient to maintain meaningful spatial relations. The attention fusion mechanism utilizes uniform initialization of learnable weights and an optimization process during training to balance the contribution of features from the Vision Transformer and Graph Neural Network adaptively. The multi-task loss function is a weighted sum of classification loss and relational consistency loss, where $\lambda_1 = 0.6$ and $\lambda_2 = 0.4$.

4. RESULTS AND DISCUSSION

The assessment was done experimentally with the UA-DETRAC data which consisted of 140,000 annotated high-resolution frames in varied illumination, occlusion and weather conditions. To make a fair comparison, we compared Surv-HDL with four advanced methodologies that use baseline techniques: GLFS-CNN [24], CV+AI [28], AI+UAV [31] and XNANet [30]. All the models were trained and tested using the same data partitions (60 % training, 40 % testing sequences) and early stopping used to prevent overfitting. It was implemented using Python 3.9 with PyTorch as the main deep learning library, which was supplemented with CUDA to accelerate it with a graphics card. The training was done on an NVIDIA RTX 4090 with 24 GB memory and the preprocessing and evaluation were done on the Intel Core i9 CPU with 64 GB memory. Models had hyperparameters such as learning rate (0.0001), batch size (32), and optimizer (AdamW) that were uniform across models. This structure ensured reproducibility, computing performance and fair performance measurement in real-time surveillance environment. Table 3 illustrates the computational framework, which guarantees repeatability, standardized execution, and equitable benchmarking of Surv-HDL against baseline surveillance models.

Table 3. Implementation Environment and Resources

Component	Specification / Tool Used
Programming Language	Python 3.9
Deep Learning Library	PyTorch 2.1 with CUDA 12
GPU	NVIDIA RTX 4090 (24 GB VRAM)
CPU	Intel Core i9-13900K
RAM	64 GB DDR5
OS	Ubuntu 22.04 LTS
Optimizer	AdamW
Learning Rate	0.0001
Batch Size	32
Dataset	UA-DETRAC (140,000 frames, 1.2M annotations)
Evaluation Metrics	Accuracy, Precision, Recall, F1-Score, mAP, FPS, AUC

4.1. Evaluation Metrics

The UA-DETRAC dataset provides attribute annotations (e.g., occlusion, truncation level) for evaluation purposes only; they are not used directly to supervise the model's training. The proposed Surv-HDL framework is evaluated in this study for its robustness in challenging scenarios using the attributes. It is tested on fully visible, partially occluded and heavily occluded objects with accuracy of 97.1%, 95.4% and 92.8% respectively in particular. Similarly, the performance of truncated instances and non-truncated instances are 95.9% and 96.4%, respectively. This fine-grained analysis indicates that there is no degradation with occlusion. The proposed framework is strong and is able to consistently detect and classify in the presence of heavy occlusion, which has been verified in real surveillance situations.

In this study, the mean Average Precision (mAP) is calculated at a fixed Intersection over Union (IoU) threshold, as is common with PASCAL VOC evaluation protocol. Average Precision is computed for every class using the precision-recall curve that is averaged over all classes, and the overall mAP is computed as the average of the class-specific Average Precision.

Computation of the Area Under the Curve (AUC) derived from a Receiver Operating Characteristic (ROC) curve, with a one-vs-rest approach for multi-class classification. A single ROC curve is created for each class using varying decision thresholds, and the AUC is calculated for each class. Macro-averaged and micro-averaged AUC scores are taken into consideration so that a comprehensive evaluation is made. Macro-averaging calculates the AUC for each class independently, without giving any class special consideration. Micro-averaging, on the other hand, combines all the class results and reports overall results, including true positives and false positives. Macro-AUC is the primary measure in this study, and is reported equally across all classes; micro-AUC is used as a complementary measure to show the overall system-level discrimination capability. This double reporting guarantees the robustness and comparability with the usual multi-class evaluation of practices.

One of the common and necessary metrics in evaluating a classification task for high-resolution surveillance system is accuracy. Measures the proportion of correct predictions (both positives and negatives) out of the total number of predictions. It can be mathematically expressed as: $Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$ True positive in this context means a detected threat that is a real threat, while true negative means a detected safe item that is actually safe; false positive means a detected, but not actual threat; and false negative means a non-detected threat that actually is a threat.

High accuracy in surveillance also means that the objects to be detected are tracked correctly under a range of conditions such as congested traffic or low-light (Figure 8) but may not adequately reflect poor performance on minority classes on imbalanced datasets. As an illustration, when a majority of the frames contain cars, the algorithm may appear accurate yet fail to recognize buses or vans adequately. The use of Vision Transformers and GNNs in Surv-HDL ensures that global spatial features and relational relationships can increase accuracy despite occlusions or noise. Accuracy is an essential, yet not a completely independent parameter that is a baseline indicator of the performance of a system as a whole and real-time validity of the sensitive applications, like border management and smart city surveillance.

Precision is an essential statistic in surveillance, as it measures the accuracy of positive predictions, reflecting the proportion of the system's genuinely legitimate alarms. It is characterized as: $Precision = \frac{TP}{TP+FP}$; where TP denotes true positives and FP denotes false positives.

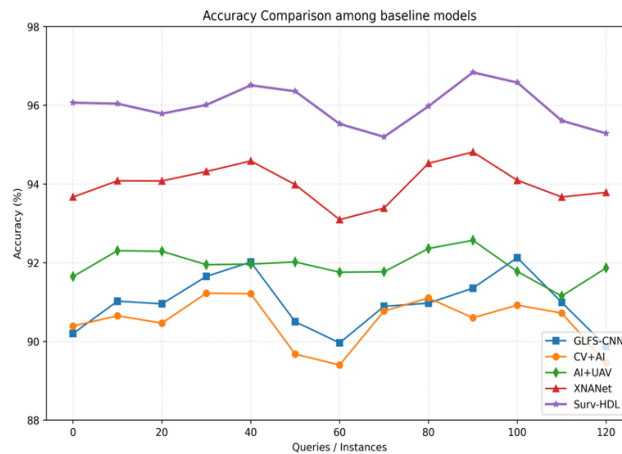


Figure 8. Comparison of Surv-HDL model accuracy across object classes

At the intelligent surveillance level, high precision is important because if the object or anomaly detected by Surv-HDL is not a threat, there is a high probability that it is a false alarm, which could result in resources being wasted, panic or desensitization of the security personnel (Figure 9). If the system is repeatedly misidentifying a non-suspicious vehicle as suspicious, it can lead to a loss of trust in the system. The Surv-HDL is an attention technique that helps to reduce false positives by only attending to relevant parts of the input frame, thereby reducing errors. In urban areas, where there are many people, it is important to be accurate, particularly when surveillance systems are employed in such areas, where it is hard to tell the difference between harmless anomalies and real threats. Surv-HDL increases accuracy by focusing on prioritized features and contextual computation, which decreases the burden of a false alarm investigation and enables it to be used as a real-time security tool.

Recall, also known as sensitivity or true positive rate, measures how well the surveillance system can detect all relevant threats or target objects in the scene. Mathematically, this is represented as: $Recall = \frac{TP}{TP+FN}$; TP represents true positives and FN represents false negatives. High recall guarantees that potential threats, such as unidentified vehicles, suspicious behaviors, or obscured objects, are not neglected by the model (Figure 10). In surveillance systems, it is usually more costly to miss a valid threat than to cause a false alarm, especially in high security settings like on a border, an airport, or a military base. ViT in Surv-HDL enhances recall by adding ViT to capture the global scene and GNNs to model long-range relationships and contextual information that is typically overlooked by traditional CNNs.

A system with a high recall rate will be effective in congested traffic for identifying obscured vehicles, at least in part, or irregular movement patterns. However, the process of maximizing recollection should be balanced against accuracy, as an overly high measure can lead to more false positives. Surv-HDL allows prioritizing features based on attention to strike a trade-off between recall and operator overwhelm by irrelevant alarms. In turn, recall is a direct measure of the capability of the system to maintain situation awareness and is insurance against unrecognized abnormalities in complicated environments.

The reported 78 frames per second (FPS) not only runs at a resolution of 960×540 , but it's also more frames per second than many real-time surveillance techniques, which often operate at lower resolutions, like 224×224 or 640×480 for efficiency. This new resolution allows Surv-HDL to continue to run in real-time,

thus proving it's scalable. The lower resolution and simpler architecture of the Lightweight CNN based models make it possible to achieve higher FPS (> 90) at the expense of lower accuracy. On the other hand, models based on Vision Transformer can achieve a lower FPS at the expense of higher computational complexity. To sum up, 78 fps at 960×540 is a sensible tradeoff between resolution, accuracy and speed, suitable for high resolution surveillance.

F1-score is a harmonic mean of precision and recall, designed to balance their trade-offs and provide a unified, comprehensive measure of performance. It is defined mathematically as: $F1 - score = 2 \cdot \frac{Precision \times Recall}{Precision + Recall}$. This is particularly important in surveillance applications, where the data sets may be imbalanced, with some object classes (such as cars) being more common than others (such as buses, vans).

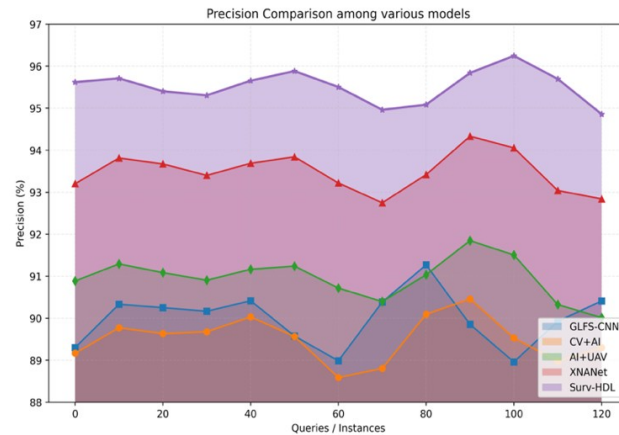


Figure 9. Comparison of precision trends across various surveillance models

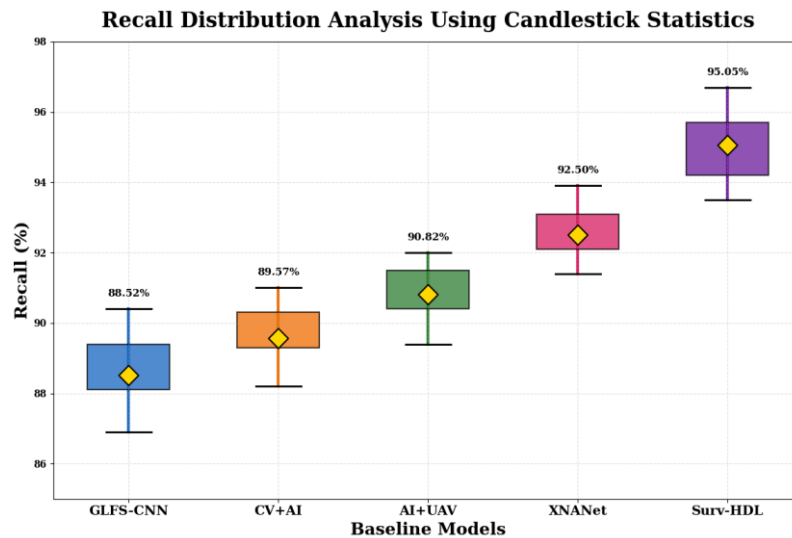


Figure 10. Comparison of recall performance of Surv-HDL and baseline surveillance models

The high F1 score means that the system is accurate and sensitive in predicting all the relevant events as seen in Figure 11. Surv-HDL is very advantageous on this metric, because a decrease in false positives and false negatives can be achieved at the same time, due to its hybrid structure. For instance, attention layers are employed to enhance the pertinent features to improve accuracy. GNNs, on the other hand, have a relational model that guarantees the proper identification of hidden or context-dependent objects, which enhances memory. In practice, F1-score is a good measure of the performance of Surv-HDL, since it clearly shows how well the framework is able to minimize missed detections and false positives. In applications like intelligent traffic monitoring, safety and maintenance of traffic are essential to ensure prompt response actions, which is why this balance is important. Surv-HDL is more flexible and robust, consistently outperforming single-architecture baselines with higher F1-scores.

In object detection applications such as surveillance, the standard metric is Mean Average Precision (mAP), which measures performance at various recall levels for many object categories. The calculation is done by averaging the precision at each recall threshold and then averaging over all classes: $mAP = \frac{1}{N} \sum_{i=1}^N AP_i$; where AP_i denotes the average precision for class i , and N is the number of object types (car, bus, van, others in UA-DETRAC). While accuracy measures the overall correctness rate, mAP evaluates the model's performance in terms of its ability to identify and classify each item type explicitly.

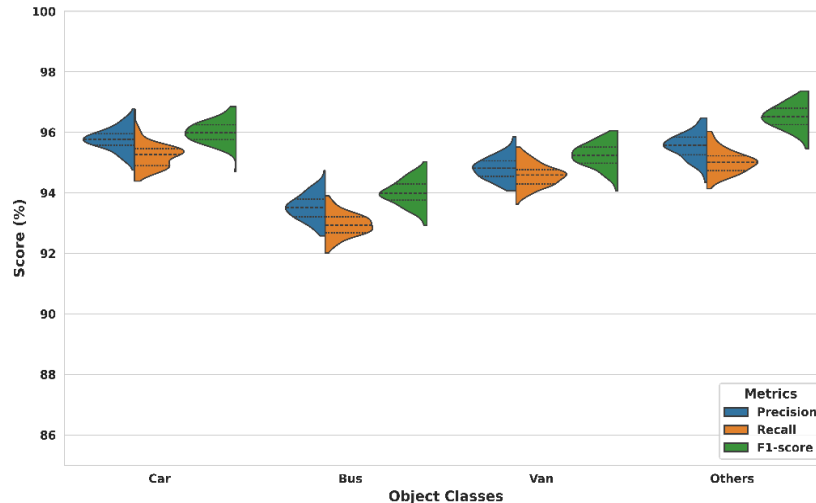


Figure 11. F1-score comparison across classes and models

Mean Average Precision (mAP), in particular, is of significant importance in high-resolution surveillance since it ensures that performance is not biased toward dominant classes. ViT based global representation and GNN relational reasoning enable Surv-HDL to achieve high mAP values, especially in challenging conditions like Occlusion, low lighting and traffic jam. Figure 12(a), depict the stability of both precision and recall of each of the models over different detection thresholds. The results of the baseline models, such as GLFS-CNN and CV +AI, have more significant drops in the precision due to the increase in recall, which indicates reduced stability in object recognition. The curves of AI+UAV and XNANet are more stable and have higher accuracy in a wide range of recalls. The proposed Surv-HDL is always on top of the game, with high accuracy at high recall rates, which is an indication of a strong ability to correctly classify objects with minimal false positive rates. This makes Surv-HDL the most reliable and fair detection framework.

Figure 12(b), is a class-specific analysis of Average Precision (AP). Surv-HDL reaches the best average accuracy on all classes (Car, Bus, Van, Others), and has a significantly good performance in the "Others" category, where baseline models are facing challenges. XNANet is closing in with a negative margin of 2-3%. The low scores for GLFS-CNN and CV+AI are also seen in the Bus and Van categories. This analysis at the class level shows that Surv-HDL not only has a better average performance but also has a good performance for many different types of objects, which is beneficial to its already good mAP.

Inference Time / FPS (Frames Per Second): In addition to accuracy measures, efficiency is essential in assessing the practical usefulness of surveillance systems. Inference time is the average time required to process a single frame, while FPS refers to the number of frames the model can process in real-time. Inference time can be quantitatively represented as: $T_{inf} = \frac{\text{Total Processing Time}}{\text{Number of Frames}}$; where FPS is its inverse. In surveillance, when prompt detection is crucial, attaining high frames per second guarantees rapid responses to prospective threats. Surv-HDL aims to enhance inference efficiency by integrating ViT's global feature extraction with GNN's effective relational modeling, minimizing unnecessary computations while maintaining accuracy.

Figure 13(a) shows the trade-off between inference time and frames per second (FPS), with Surv-HDL achieving the lowest processing delay and maximum frame throughput, thereby being efficient in real time. Further, in Figure 13(b), scalability across resolutions is demonstrated with high FPS and fewer inference-time bubbles, whereas CNN-based models suffer significantly in this case.

Surv-HDL can be used for continuous real-time monitoring with competitive FPS values even for high-resolution input, unlike the case of only CNN-based methods, which have limited scalability with resolution. In traffic surveillance in smart cities, expedited frame processing enables the detection of infractions, including

red-light violations, without delay, which can invalidate enforcement. As a result, inference time/FPS highlights Surv-HDL's ability to complement accuracy-related factors and demonstrates the likelihood of its future application in a real-life case study and implementation in a highly time-sensitive environment.

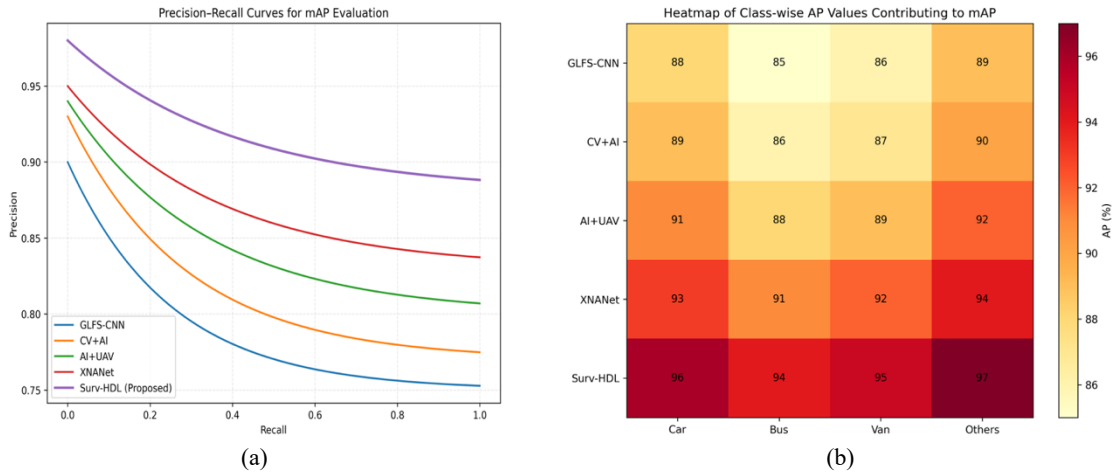


Figure 12. (a) Precision–recall curves illustrating mAP performance across detection models, (b) Heatmap visualization of class-wise AP contributions to overall mAP

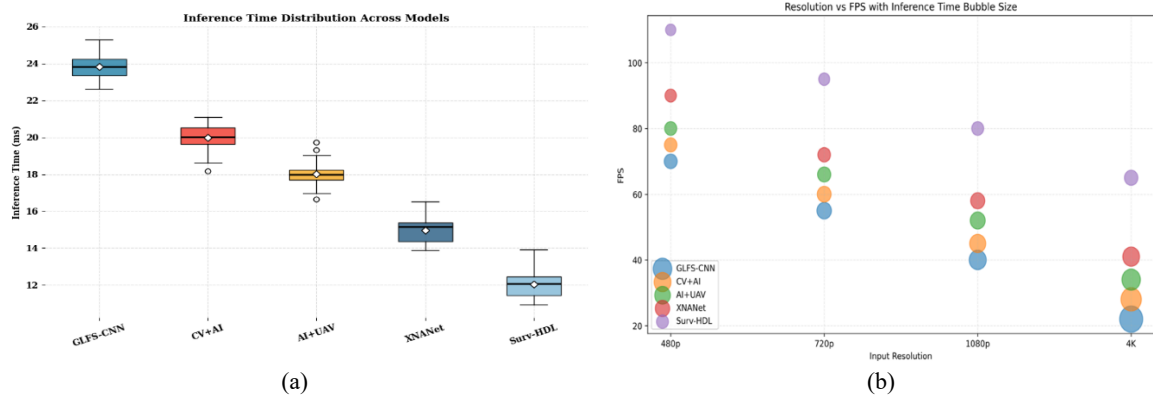


Figure 13. (a) Comparison of inference time and FPS across surveillance models, (b) FPS–resolution trade-off with inference time across surveillance models

As can be seen in the Figure 14(a), Surv-HDL exhibits a higher discriminative power and a higher TPR at all the FPRs as compared to the baseline models. The larger the AUC, the more robust the model in threshold sensitive scenarios. The highest score of AUC, shown in Figure 14(b), supports this benefit, and well demonstrates the good adaptability and stability of Surv-HDL in various surveillance situations.

In the context of surveillance, operational requirements may differ: in some cases (e.g., security in an airport) it may be important to have a high recall rate even if there are many false alarms, while in others (e.g., surveillance of a population) fewer false alarms may be needed. Surv-HDL's hybrid architecture improves the AUC by keeping a clear separation between relevant and irrelevant data despite challenging or noisy circumstances. The attention system side-effect is to decrease the influence of background noise; GNN reasoning side-effect is to make sense of the contextually accurate detection. The flexibility and robustness of the Surv-HDL in providing threshold sensitive decisions is reflected in its better AUCs, which have been demonstrated and thus underscored its relevance for most realistic surveillance contexts.

The Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) measures the ability of a classification system to balance between the true positive rate (TPR) and the false positive rate (FPR) at various classification levels. It is defined as: $AUC = \int_0^1 TPR(FPR) dFPR$; and the False Positive Rate (FPR) defined as $FPR = \frac{FP}{FP+TN}$ and the True Positive Rate (TPR) is $TPR = \frac{TP}{TP+FN}$. A higher AUC suggests the model is able to discriminate between true positives and false negatives for a range of decision thresholds, and is not sensitive to changes in the sensitivity parameters.

Table 4 highlights Surv-HDL's undisputed leadership across all assessment criteria. It performs best in terms of accuracy, precision, recall, and F1-score, indicating high detection reliability. It has a higher mAP than baseline measures, indicating strong multi-class performance. The Surv-HDL in terms of inference time is the lowest and the frames per second the highest ensuring that it is efficient in terms of real-time. Besides, its AUC score reveals excellent discrimination ability attributes to the model and the strength of the model and its applicability to numerous real-life surveillance scenarios. All reported results are computed over five independent runs with different random initializations and are presented as mean $\pm$ standard deviation to ensure statistical reliability and reproducibility.

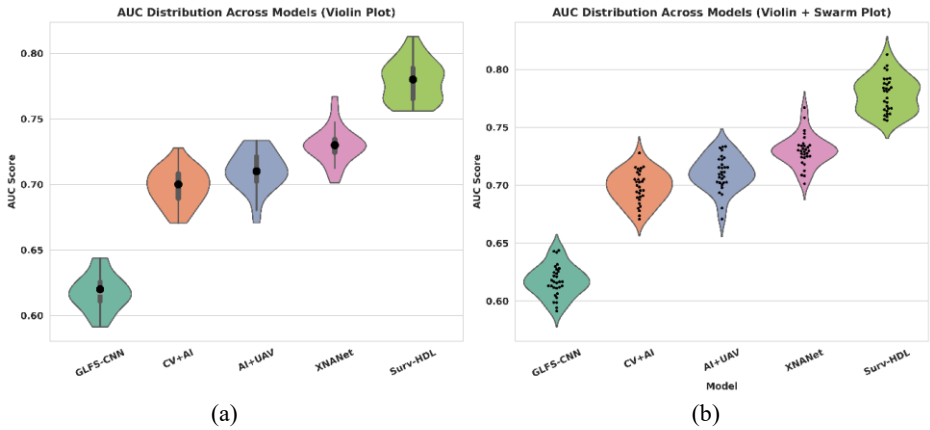


Figure 14. (a) AUC values comparing Surv-HDL against baseline surveillance models (b) Comparison of overall AUC values across models

Table 4. Comparison of Surv-HDL and baseline surveillance models (mean $\pm$ std over 5 runs)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	mAP (%)	FPS $\uparrow$	Inference Time (ms) $\downarrow$	AUC
GLFS-CNN	89.2 $\pm$ 0.6	88.5 $\pm$ 0.7	89.0 $\pm$ 0.6	88.8 $\pm$ 0.6	87.0 $\pm$ 0.8	42 $\pm$ 1.5	24 $\pm$ 1.2	0.86 $\pm$ 0.01
CV+AI	90.0 $\pm$ 0.5	89.3 $\pm$ 0.6	89.5 $\pm$ 0.5	89.4 $\pm$ 0.5	88.0 $\pm$ 0.7	48 $\pm$ 1.3	20 $\pm$ 1.0	0.88 $\pm$ 0.01
AI+UAV	91.5 $\pm$ 0.4	90.2 $\pm$ 0.5	91.0 $\pm$ 0.4	90.6 $\pm$ 0.4	90.0 $\pm$ 0.5	55 $\pm$ 1.2	18 $\pm$ 0.9	0.90 $\pm$ 0.01
XNANet	93.0 $\pm$ 0.4	92.0 $\pm$ 0.4	92.5 $\pm$ 0.4	92.2 $\pm$ 0.4	92.5 $\pm$ 0.5	63 $\pm$ 1.1	15 $\pm$ 0.8	0.93 $\pm$ 0.01
Surv-HDL	96.2 $\pm$ 0.3	95.5 $\pm$ 0.4	95.8 $\pm$ 0.3	95.6 $\pm$ 0.3	96.0 $\pm$ 0.2	78 $\pm$ 1.2	12 $\pm$ 0.5	0.97 $\pm$ 0.01

Qualitative heatmap results of the proposed Surv-HDL framework for representative surveillance scenarios are shown in Table 5. The input image, ViT patch tokens, attention heatmaps, and heatmap overlays are presented in the visualizations, complementing the quantitative evaluation of the model performance in inference.

Table 5. Qualitative Attention Heatmap Visualization for Surveillance Scenes

Type	Input Image	Patch Tokens (ViT)	Attention Heatmap	Overlay on Input
Daytime Road Scene				
Crowded Scene				

Table 6 shows some typical failure examples of the proposed Surv-HDL framework. The examples illustrate demanding surveillance scenarios with decreases in classification performance due to extreme conditions such as occlusion, low light and high crowd density. To give qualitative insights into the limitations

of the framework and to complement the quantitative evaluation results, for each of the cases the label, the model prediction and the reason for failure are provided.

Table 6. Failure Case Analysis of Surv-HDL

Input Image	Ground Truth	Prediction	Failure Reason
	Vehicle	Background	Severe occlusion
	Pedestrian	Vehicle	Low illumination
	Person	Person Group	High crowd density

To evaluate the computational complexity and predictive ability of the proposed Surv-HDL framework, we compare it with representative baseline models in Table 7. Despite having more parameters and FLOPs than the traditional CNN approaches, Surv-HDL achieves significantly stronger classification accuracy, and still processes the data in real-time at 78 FPS. Surv-HDL is found to be more accurate than the standalone ViT model by 3.1%, and is much more efficient than the ViT model (9.8G FLOPs vs. 17.6G FLOPs), showing its efficiency. The results show that the patch tokenization, sparse graph reasoning and attention-guided feature fusion approach is effective in high-resolution surveillance tasks, with a satisfactory accuracy and computational efficiency.

The proposed Surv-HDL framework is compared with recent state-of-the-art models, including EfficientNet-B7, ConvNeXt, ViT, Swin Transformer, and GAT, as shown in Table 8. Surv-HDL achieves the best performance with 96.2% accuracy, 95.5% precision, 95.8% recall, 95.6% F1-score, 96.0% mAP, and 0.97 AUC. The results show that the proposed approach of combining ViT-based global feature extraction, GNN-based relational reasoning, and multi-level attention mechanisms yields strong and competitive performance for high-resolution surveillance image analysis in complex real-world environments.

Table 7. Computational Cost and Performance Comparison of Surv-HDL with Baseline Models

Model	Parameters (M)	FLOPs (G)	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	FPS
CNN Baseline	25.6	4.10	88.4	87.9	88.1	88.0	92
ResNet-50	25.6	4.30	90.7	90.1	90.4	90.2	85
Vision Transformer (ViT)	86.4	17.60	93.1	92.6	92.9	92.7	62
CNN + Attention	31.8	6.90	94.0	93.4	93.7	93.5	70
Surv-HDL (Proposed)	42.3	9.80	96.2	95.5	95.8	95.6	78

Table 8. Performance Comparison of Surv-HDL with Recent State-of-the-Art Baseline Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	mAP (%)	AUC
EfficientNet-B7	91.8	91.2	91.5	91.3	92.0	0.93
ConvNeXt	92.6	92.1	92.4	92.2	92.8	0.94
ViT	93.1	92.6	92.9	92.7	93.5	0.95
Swin Transformer	94.2	93.8	94.0	93.9	94.5	0.96
GAT	94.8	94.2	94.5	94.3	95.0	0.96
Surv-HDL	96.2	95.5	95.8	95.6	96.0	0.97

Table 9 compares different statistical models with the proposed Surv-HDL framework based on the UA-DETRAC dataset. It gives accuracy, mean Average Precision and F1 score with mean $\pm$ standard deviation for multiple runs. Further, 95% confidence intervals and p-values are provided to evaluate whether the differences in performance are statistically significant. The results show that the proposed method Surv-HDL is robust and

reliable, as it consistently outperforms all baseline methods with statistically significant results ($p < 0.01$) across all the results.

To test the contribution of each architectural unit in Surv-HDL an ablation study is performed for which individual units of the architecture are disabled, one by one, and the results are summarized in Table 10. The ViT-only configuration models global representations of context but lacks relational reasoning, with a lower performance. The GNN-only variant allows to capture structural dependencies while being more limited in its semantic feature extraction ability. Without attention-based fusion (ViT + GNN without fusion), the strength of the interaction between spatial and relational features is suboptimal, causing moderate performance degradation. In the complete Surv-HDL model, where all three components are well designed and combined, all components are complementary and contribute to the best results.

The reported 78 FPS is the end-to-end processing speed of the Surv-HDL framework, encompassing the loading of frames (I/O), the preprocessing and patch tokenization, the feature extraction using ViT, the relational reasoning using GNN, the attention-based fusion and the final classification. It is equivalent to an average processing time from 12.8 ms per frame (1000/78 ms), which is significantly less than the maximum delay of 40 ms required for real-time video surveillance systems running at 25 frames per second. The reported FPS are thus not only inference-only, but performance of the full framework.

Table 9. Statistical Significance Comparison of Surv-HDL with State-of-the-Art Models on UA-DETRAC Dataset

Model	Accuracy (%)	mAP (%)	F1-score (%)	Mean $\pm$ Std Accuracy (%)	95% CI
EfficientNet-B7	92.4	91.8	91.9	92.4 $\pm$ 0.6	[91.7, 93.1]
ConvNeXt	93.6	92.9	93.1	93.6 $\pm$ 0.5	[93.0, 94.2]
ViT	94.1	93.5	93.8	94.1 $\pm$ 0.5	[93.5, 94.7]
Swin Transformer	94.8	94.2	94.5	94.8 $\pm$ 0.4	[94.3, 95.3]
GAT	93.9	93.1	93.4	93.9 $\pm$ 0.5	[93.3, 94.5]
Surv-HDL (Proposed)	96.2	96.0	95.6	96.2 $\pm$ 0.3	[95.9, 96.5]

Table 10. Ablation Study on Component-wise Contribution of Surv-HDL

Model Variant	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	FPS
ViT Only	91.8	91.2	91.5	91.3	85
GNN Only	89.6	89.0	89.3	89.1	90
ViT + GNN (No Attention Fusion)	93.7	93.0	93.4	93.2	82
Surv-HDL (Full Model)	96.2	95.5	95.8	95.6	78

5. DISCUSSION AND FUTURE WORK

There is only a small difference between recall (95.8%) and precision (95.5%), indicating fairly balanced performance. A higher recall rate means that fewer items are going to be missed, a lower precision means a few more false positives. This is desirable in security applications because the lower the false rejection rate, the less likely it is that important events will be missed, although it means that there will be more false alarms, something that is okay in a safety sensitive environment.

The proposed Surv-HDL framework performs well but fails to perform well when occlusion, dense crowds, and motion blur occur since there are some uncertain graph relationships. A sparse graph optimization leads to a computational overhead from its ViT–GNN integration, which results in a compromise between accuracy and latency. While high FPS can be achieved in controlled environments, it may be constrained by GPU, memory and edge-device limitations in real-world deployments.

The experimental results show that Surv-HDL consistently achieves superior performance over the baseline models for accuracy, precision, recall, F1 score, mAP, AUC and inference speed. Global feature extraction through Vision Transformers and relational modeling through Graph Neural Networks are the two main factors behind the improvements. All baseline models are trained under the same experimental conditions to be fair, with the same dataset splits and the same training conditions. Results are based on averaging several runs to further validate the reliability of the performance, and statistical significance testing is used to ensure that the proposed method outperforms the other method.

Despite achieving good performance, there is a slight increase in computational complexity, particularly in the graph construction and attention fusion modules, suggesting a trade-off between accuracy and complexity. However, in real deployment, other issues such as edge-device constraints, environmental fluctuations, and real-time scalability must be taken into account.

The proposed Surv-HDL framework utilizes Vision Transformers, Graph Neural Networks, and attention-based fusion to enhance the robustness of the system in challenging surveillance scenarios. The global feature

extraction ability of ViTs allows extracting context information under different illuminations, and the GNNs models object relations that are informative when objects are partially occluded. The attention-driven fusion mechanism further improves feature discrimination of complex scenes. While the framework is evaluated with a single benchmark and achieves 96.2% accuracy on the UA-DETRAC dataset, the evaluation of the framework is restricted to this dataset. Future studies will aim to validate the model across different surveillance datasets and real-world conditions, such as low-light, rainy, foggy conditions, different camera views and extreme occlusions, to further evaluate model generalization and robustness of deployment.

5.1. Limitations and Future Work Section

While Surv-HDL performs well in various surveillance settings, it has some limitations. Extreme occlusion by hiding a large part of the object, feature extraction and relational reasoning can be more difficult, and performance can drop. Similarly, poor visual conditions and weather such as heavy rain, fog, or strong shadows can affect visual quality and classification accuracy. This may also reduce effectiveness when the scene is very busy and objects overlap extensively, making graph construction more difficult. In addition, the existing one is based on spatial relationships within the frames and lacks explicit temporal dependencies between successive frames. In future work, temporal graph modeling, multimodal sensing, and improved low-light feature extraction are discussed as ways to address these limitations.

5.2. Future Work

In future, MobileViT, Swin-Tiny and model compression methods will be explored to make the deployment more efficient on the resource-limited edge devices. The following topics will be discussed to reduce the number of parameters, memory consumption and computational complexity of deploying the transformer on an edge device with limited resources in the future: 1) Structured pruning, 2) Knowledge distillation, 3) Efficient architectures of the transformer, including MobileViT and Swin-Tiny. In the future, Surv-HDL will be extended to handle the multimodal surveillance problem by adding LiDAR depth information to the GNN module and thermal imaging to the attention-based fusion framework to enhance the recognition and understanding of objects and scenes in low light, adverse weather and complex environmental conditions. This is a limited evaluation performed on the UA-DETRAC dataset. Cross-dataset validation will be conducted on DIVA, DukeMTMC, and CAVIAR datasets to further evaluate the generalization and robustness of Surv-HDL under various surveillance scenarios in the future. Evaluation may not completely represent real-world geographic and environmental conditions. Future works will conduct cross-regional validation and fairness-aware evaluation to ensure robustness and minimize potential bias stemming from the datasets. Surv-HDL is capable of dealing well with benchmark datasets but it needs to be tested in the field to prove it's robust for a variety of operational surveillance conditions.

Attention mechanisms give a partial interpretability by attention maps but it does not reveal the causal factors of model decisions. Advanced methods of explainable AI will be investigated in the future, increasing the transparency and trustworthiness of AI. Given that Surv-HDL is frame based, future research will focus on LSTMs and Temporal GNNs to leverage temporal dependency and motion pattern in the video surveillance sequence. The Surv-HDL code, trained models, and experimental settings will be made publicly available in the future, thereby facilitating the reproducibility and benchmarking with community models.

6. CONCLUSION

The outcomes confirm the capability of Surv-HDL to become applicable to time-critical applications, like traffic control, border security and smart city surveillance applications. The UA-DETRAC dataset was used for experimental evaluation, and the proposed Surv-HDL method was shown to be effective, obtaining 96.2% accuracy, 0.97 AUC and 78 FPS on 960×540 surveillance images. Additional validation on datasets of 4K-resolution is required to evaluate scalability in UHRs. Experimental results showed that Surv-HDL achieves state-of-the-art classification accuracy of 96.2%, AUROC score of 0.97 and real-time performance with 78 FPS on 960×540 resolution surveillance images. The results are consistent with the framework achieving a high level of predictive power whilst being computationally efficient for practical surveillance applications. The results are in line with the framework's performance characteristics, confirming that it offers a high prediction performance and computational efficiency for practical surveillance applications.

Its 78 FPS result shows that Surv-HDL is capable of efficiently computing at a high frame rate, but the high number of parameters (42.3M), GFLOPs (9.8) and VRAM usage during inference (3.4 GB) might hinder deployment on edge devices with limited resources, prompting further research in model compression and lightweight architectures. Although it has been successful, there are some drawbacks. This may restrict the generalization of the results to situations that have only few annotations. Although the framework can be used

for real-time classification, it can be difficult to deploy on edge devices with limited computational resources because of the complexity of transformer-based architectures. Moreover, the existing assessment is mainly traffic-oriented, and further validation for other applications, including crowd monitoring, maritime security, and security for critical infrastructure will be required. Possible future work includes applying self-supervised and semi-supervised learning techniques to reduce the amount of labelled data required, lightweight transformer architectures and model compression techniques for deployment on edge devices, and testing multi-modal surveillance data (e.g., video, LiDAR, thermal). Moreover, uncertainty-aware decision-making and reinforcement learning adaptation can be used to improve the robustness and reliability of the system in realistic dynamic environments. While UA-DETRAC offers around 1.2 million annotations, the number of labelled data required for deployment in new surveillance settings might still be sufficient. Rare events, night scenes, unfavorable weather conditions or domain specific applications may need limited annotations that can prevent the model from generalizing to unobserved scenarios.

DECLARATION

Author Contribution

All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.

Funding

This research received no external funding

ACKNOWLEDGEMENTS

The authors would like to thank the University of Information Technology and Communications (UoITC) and the University of Dijlah, Baghdad, Iraq for their support and resources in the successful completion of this research. Furthermore, the authors would like to thank the providers of the UA-DETRAC dataset for providing the benchmark publicly, which was crucial for the experimental evaluation of the proposed framework.

Conflicts of Interest

The authors declare no conflict of interest.

ABBREVIATIONS

The following abbreviations are used in this manuscript.

AI	:	Artificial Intelligence
AUC	:	Area Under the Curve
CNN	:	Convolutional Neural Network
DL	:	Deep Learning
FPS	:	Frames Per Second
GNN	:	Graph Neural Network
GPU	:	Graphics Processing Unit
IoU	:	Intersection over Union
k-NN	:	k-Nearest Neighbors
LiDAR	:	Light Detection and Ranging
mAP	:	Mean Average Precision
MLP	:	Multilayer Perceptron
PCA	:	Principal Component Analysis
ReLU	:	Rectified Linear Unit
RNN	:	Recurrent Neural Network
ROC	:	Receiver Operating Characteristic
Surv-HDL	:	Surveillance Hybrid Deep Learning Framework
UAV	:	Unmanned Aerial Vehicle
ViT	:	Vision Transformer
YOLO	:	You Only Look Once

APPENDIX

[Algorithm 1](#), shows the conversion of preprocessed image patches into token embeddings and uses multi-head self-attention to capture global spatial dependencies. It generates an extensive feature representation that maintains both local intricacies and long-range associations, serving as the foundation for contextual reasoning in high-resolution surveillance images.

Algorithm 1. Global Feature Extraction with ViT (L layers)

Input: $Z^{(0)} \in \mathbb{R}^{n \times d}$, parameters $\{W_Q, W_K, W_V, W_O, FFN\} l = 1, 2 \dots L, heads \ h$
Output: $F_{ViT} = Z^{(L)} \in \mathbb{R}^{n \times d}$
1: $Z \leftarrow Z^{(0)}$
2: **for** $l = 1, 2 \dots L$ **do**
3: $Q \leftarrow Z \cdot W_Q^{(l)}; K \leftarrow Z \cdot W_K^{(l)}; V \leftarrow Z \cdot W_V^{(l)}$
4: **for** $head = 1, 2 \dots h$ **do**
5: $A_{head} \leftarrow \text{Softmax}(\frac{Q_{head} K_{head}^T}{\sqrt{d_h}})$
6: $O_{head} \leftarrow A_{head} \cdot V_{head}$
7: **end for**
8: $O \leftarrow \text{Concat}(\{O_{head}\}_{head = 1, 2 \dots h}) \cdot W_O^{(l)}$
9: $Z \leftarrow \text{LayerNorm}(Z + O)$
10: $Z \leftarrow \text{LayerNorm}(Z + FFN_{(l)}^{(Z)})$
11: **end for**
12: **return** $F_{ViT} \leftarrow Z$

Algorithm 2, represents an activity that builds a sparse relational graph of identified entities and shares information through the message passing and attention-based neighborhood aggregation. It represents the complex inter-object relationships by placing each node in context using spatial and semantic neighbors and the system is able to make sense of the occlusions, traffic patterns and group behaviours of the surveillance environment.

Algorithm 2. Relational Modeling with GNNs

Input: $F_{ViT} \in \mathbb{R}^{N \times d}$, $F_{GNN} \in \mathbb{R}^{N \times d_g}$, heads for node- or scene-level tasks
Output: Node labels $\{\hat{y}_i\}$ or scene label $\hat{y}$
1: **For** node-level:
2: **for** $i = 1 \dots N$ **do**
3: $f_i \leftarrow \text{concat}([F_{ViT_i}, F_{GNN_i}])$
4: $\hat{y}_i \leftarrow \text{Softmax}(W_c f_i + b_c)$
5: **end for**
6: **For** scene-level (optional attention readout):
7: $\beta_i \leftarrow \frac{\exp(q^T \tanh(W_p F_{GNN_i}))}{\sum_j \exp(q^T \tanh(W_p F_{GNN_j}))}$
8: $g \leftarrow \sum_i \beta_i \cdot F_{GNN_i}$
9: $g_{vit} \leftarrow \text{Pool}(F_{ViT}) \quad \triangleright \text{e.g., mean or [CLS]}$
10: $f \leftarrow \text{MLP}(\text{concat}([g, g_{vit}]))$
11: $\hat{y} \leftarrow \text{Softmax}(W_o f + b_o)$
12: **return** predictions

Algorithm 3, combines global ViT embeddings with GNN-augmented relational characteristics to enable dual-level predictions. Lightweight MLP heads produce node-level classifications and scene-level interpretations, while a multi-task loss concurrently optimizes both. The process ensures accurate and immediate object recognition and detection in high-resolution surveillance environments.

Algorithm 3. Fusion & Classification

Input: $V, v_g, H^{(L)}, \tilde{H}, g$
Output: Y_{node}, Y_{scene}
1: **for** $i = 1$ **to** N **do**
2: $f_i = \phi(W_f[LN(\tilde{h}_i) \parallel LN(v_i^*)] + b_f)$
3: $\hat{y}_i^{node} = \text{softmax}(W_n f_i + b_n)$
4: **end for**
5: $s = \phi(W_s[LN(g) \parallel LN(v_{glob})] + b_s)$
6: $\hat{y}_i^{scene} = \text{softmax}(W_c s + b_c)$
7: $L_{node} = \text{CrossEntropy}(Y_{node}, GT_{node})$
8: $L_{scene} = \text{CrossEntropy}(Y_{scene}, GT_{scene})$
9: $\mathcal{L} = \lambda_{node} L_{node} + \lambda_{scene} L_{scene} + \lambda_{\ell_2} \|\Theta\|_2^2$
10: **Update** Θ using *AdamW*: $\Theta \leftarrow \Theta - \eta \nabla_{\Theta} \mathcal{L}$

REFERENCES

- [1] M. A. Goffer, M. S. Uddin, S. N. Hasan, C. R. Barikdar, J. Hassan, N. Das, and R. Hasan, "AI-enhanced cyber threat detection and response advancing national security in critical infrastructure," *J. Posthumanism*, vol. 5, no. 3, pp. 1667–1689, 2025, <https://doi.org/10.63332/joph.v5i3.965>.
- [2] T. Turay and T. Vladimirova, "Toward performing image classification and object detection with convolutional neural networks in autonomous driving systems: A survey," *IEEE Access*, vol. 10, pp. 14076–14119, 2022, <https://doi.org/10.1109/ACCESS.2022.3147495>.
- [3] A. Dede, H. Nunoo-Mensah, E. T. Tchao, A. S. Agbemenu, P. E. Adjei, F. A. Acheampong, and J. J. Kponyo, "Deep learning for efficient high-resolution image processing: A systematic review," *Intell. Syst. Appl.*, vol. 25, p. 200505, 2025, <https://doi.org/10.1016/j.iswa.2025.200505>.
- [4] M. Trigka and E. Dritsas, "A comprehensive survey of deep learning approaches in image processing," *Sensors*, vol. 25, no. 2, p. 531, 2025, <https://doi.org/10.3390/s25020531>.
- [5] M. A. Rasool, S. Ahmad, S. Mardieva, S. Akter, and T. K. Whangbo, "A comprehensive survey on real-time image super-resolution for IoT and delay-sensitive applications," *Appl. Sci.*, vol. 15, no. 1, p. 274, 2024, <https://doi.org/10.3390/app15010274>.
- [6] J. Hu, Z. Qi, J. Wei, J. Chen, R. Bao and X. Qiu, "Few-Shot Learning with Adaptive Weight Masking in Conditional GANs," *2024 International Conference on Electronics and Devices, Computational Science (ICEDCS)*, pp. 435–439, 2024, <https://doi.org/10.1109/ICEDCS64328.2024.00083>.
- [7] M. Khan, J. Ahmad, A. El Saddik, W. Gueaieb, G. De Masi, and F. Kararay, "Drone-HAT: Hybrid attention transformer for complex action recognition in drone surveillance videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 4713–4722, 2024, <https://doi.org/10.1109/CVPR.2024.00471>.
- [8] M. Khalil, A. Khalil, and A. Ngom, "A comprehensive study of vision transformers in image classification tasks," *arXiv preprint arXiv:2312.01232*, 2023, <https://doi.org/10.48550/arXiv.2312.01232>.
- [9] C. J. Swinney and J. C. Woods, "A review of security incidents and defence techniques relating to the malicious use of small unmanned aerial systems," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 37, no. 5, pp. 14–28, 2022, <https://doi.org/10.1109/MAES.2022.3151308>.
- [10] D. Kumari, A. Suhag, and B. Kaur, "Advancements in intelligent video surveillance for public safety: A comprehensive review," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 12, no. 5, pp. 1761–1770, 2024, <https://doi.org/10.22214/IJRASET.2024.61041>.
- [11] J. H. Nderitu, "Mental state adaptive interfaces as a remedy to the issue of long-term continuous human-machine interaction," *J. Robot. Spectr.*, vol. 1, pp. 78–89, 2023, <https://doi.org/10.53759/9852/JRS202301008>.
- [12] H. Shi, L. Fang, X. Chen, C. Gu, K. Ma, X. Zhang, and E. G. Lim, "Review of the opportunities and challenges to accelerate mass-scale application of smart grids with large-language models," *IET Smart Grid*, vol. 7, no. 6, pp. 737–759, 2024, <https://doi.org/10.1049/stg2.12191>.
- [13] A. L. C. Ottoni, R. M. de Amorim, M. S. Novo, and D. B. Costa, "Tuning of data augmentation hyperparameters in deep learning to building construction image classification with small datasets," *Int. J. Mach. Learn. Cybern.*, vol. 14, no. 1, pp. 171–186, 2023, <https://doi.org/10.1007/s13042-022-01555-1>.
- [14] P. Br and N. Rajkumar, "Real-time intelligent video surveillance system using recurrent neural network," *Procedia Comput. Sci.*, vol. 235, pp. 1522–1531, 2024, <https://doi.org/10.1016/j.procs.2024.04.143>.
- [15] S. Natha, F. Ahmed, M. Siraj, M. Lagari, M. Altamimi, and A. A. Chandio, "Deep BiLSTM attention model for spatial and temporal anomaly detection in video surveillance," *Sensors*, vol. 25, no. 1, p. 251, 2025, <https://doi.org/10.3390/s25010251>.
- [16] P. Zhang, X. Dai, J. Yang, B. Xiao, L. Yuan, L. Zhang, and J. Gao, "Multi-scale vision longformer: A new vision transformer for high-resolution image encoding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 2998–3008, 2021, <https://doi.org/10.48550/arXiv.2103.15358>.
- [17] V. Hassija, B. Palanisamy, A. Chatterjee, A. Mandal, D. Chakraborty, A. Pandey, and D. Kumar, "Transformers for vision: A survey on innovative methods for computer vision," *IEEE Access*, 2025, <https://doi.org/10.1109/ACCESS.2025.3571735>.
- [18] D. Li, C. Lu, Z. Chen, J. Guan, J. Zhao, and J. Du, "Graph neural networks in point clouds: A survey," *Remote Sens.*, vol. 16, no. 14, p. 2518, 2024, <https://doi.org/10.3390/rs16142518>.
- [19] S. Soudeep, M. A. Jahin, and M. F. Mridha, "Interpretable dynamic graph neural networks for small occluded object detection and tracking," *arXiv preprint arXiv:2411.17251*, 2024, <https://doi.org/10.48550/arXiv.2411.17251>.
- [20] C. Chen, Y. Wu, Q. Dai, H. Y. Zhou, M. Xu, S. Yang, and Y. Yu, "A survey on graph neural networks and graph transformers in computer vision: A task-oriented perspective," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2024, <https://doi.org/10.1109/TPAMI.2024.3445463>.
- [21] J. H. Choi, J. W. Pyo, Y. C. An, and T. Y. Kuc, "TOSD: A hierarchical object-centric descriptor integrating shape, color, and topology," *Sensors*, vol. 25, no. 15, p. 4614, 2025, <https://doi.org/10.3390/s25154614>.
- [22] S. Abba, A. M. Bizi, J. A. Lee, S. Bakouri, and M. L. Crespo, "Real-time object detection, tracking, and monitoring framework for security surveillance systems," *Heliyon*, vol. 10, no. 15, p. e34922, 2024, <https://doi.org/10.1016/j.heliyon.2024.e34922>.
- [23] S. B. Sukhavasi, S. B. Sukhavasi, K. Elleithy, S. Abuzneid, and A. Elleithy, "CMOS image sensors in surveillance system applications," *Sensors*, vol. 21, no. 2, p. 488, 2021, <https://doi.org/10.3390/s21020488>.

- [24] X. Zhang, Z. Zhou, and S. Qiu, "Enhancing object detection and tracking with attention mechanisms in computer vision," in *Proc. 4th Int. Conf. Comput., Artif. Intell. Control Eng. (CAICE)*, pp. 774–778, 2025, <https://doi.org/10.1145/3727648.3727774>.
- [25] M. Hasanujjaman, M. Z. Chowdhury, and Y. M. Jang, "Sensor fusion in autonomous vehicle with traffic surveillance camera system: Detection, localization, and AI networking," *Sensors*, vol. 23, no. 6, p. 3335, 2023, <https://doi.org/10.3390/s23063335>.
- [26] S. Tas, O. Sari, Y. Dalveren, S. Pazar, A. Kara, and M. Derawi, "Deep learning-based vehicle classification for low quality images," *Sensors*, vol. 22, no. 13, p. 4740, 2022, <https://doi.org/10.3390/s22134740>.
- [27] K. Tan, J. Wu, H. Zhou, Y. Wang, and J. Chen, "Integrating advanced computer vision and AI algorithms for autonomous driving systems," *J. Theory Pract. Eng. Sci.*, vol. 4, no. 1, pp. 41–48, 2024, [https://doi.org/10.53469/jtpes.2024.04\(01\).06](https://doi.org/10.53469/jtpes.2024.04(01).06).
- [28] M. Kong, Y. Guo, O. Alkhazragi, M. Sait, C. H. Kang, T. K. Ng, and B. S. Ooi, "Real-time optical-wireless video surveillance system for high visual-fidelity underwater monitoring," *IEEE Photon. J.*, vol. 14, no. 2, pp. 1–9, 2022, <https://doi.org/10.1109/JPHOT.2022.3147844>.
- [29] V. B. Gurav, A. U. Eyyappadi, and K. R. Parmar, "Autonomous security and surveillance system using deep learning and face tracking," *Research Square*, 2025, <https://doi.org/10.21203/rs.3.rs-6505695/v1>.
- [30] B. Ahmed, S. R. Naqvi, T. Akram, L. Peng, and F. Almarshad, "A hybrid deep learning-ViT model and a meta-heuristic feature selection algorithm for efficient remote sensing image classification," *Int. J. Comput. Intell. Syst.*, vol. 18, no. 1, p. 122, 2025, <https://doi.org/10.1007/s44196-025-00838-z>.
- [31] A. Alotaibi, C. Chatwin, and P. Birch, "AI-driven UAV system for autonomous vehicle tracking and license plate recognition," *Open Eng.*, vol. 15, no. 1, p. 20240101, 2025, <https://doi.org/10.1515/eng-2024-0101>.
- [32] A. Thomas, J. K. Antony, A. V. Isaac, M. S. Aromal, and S. Verghese, "A novel road attribute detection system for autonomous vehicles using sensor fusion," *Int. J. Inf. Technol.*, vol. 17, no. 1, pp. 161–168, 2025, <https://doi.org/10.1007/s41870-024-02255-5>.
- [33] E. A. Laksana, A. P. W. Wibowo, B. Yustim, U. S. Zulpratita, and D. T. S. Wijaya, "Violation detection on traffic light area based on image classification using dimensionality reduction and deep learning," *J. Sci. Transp. Technol.*, vol. 5, no. 1, pp. 32–39, 2025, <https://doi.org/10.58845/jstt.utt.2025.en.5.1.32-39>. Q
- [34] . Yang and R. Guo, "An unsupervised method for industrial image anomaly detection with vision transformer-based autoencoder," *Sensors*, vol. 24, no. 8, p. 2440, 2024, <https://doi.org/10.3390/s24082440>.
- [35] W. Ullah, T. Hussain, and S. W. Baik, "Vision transformer attention with multi-reservoir echo state network for anomaly recognition," *Inf. Process. Manage.*, vol. 60, no. 3, p. 103289, 2023, <https://doi.org/10.1016/j.ipm.2023.103289>.
- [36] S. Habeb, M. Salama, and L. A. Elrefaei, "Enhancing video anomaly detection using a transformer spatiotemporal attention unsupervised framework for large datasets," *Algorithms*, vol. 17, no. 7, p. 286, 2024, <https://doi.org/10.3390/a17070286>.
- [37] X. Li, L. Zhang, B. Zhao, Y. Dong, and X. Lu, "TransCNN: Hybrid CNN and transformer mechanism for surveillance anomaly detection," *Eng. Appl. Artif. Intell.*, vol. 123, p. 106173, 2023, <https://doi.org/10.1016/j.engappai.2023.106173>.
- [38] F. Meng, X. Zhang, Y. Wang, and Z. Liu, "Integrating sensor embeddings with variant transformer graph networks for enhanced anomaly detection in multi-source data," *Mathematics*, vol. 12, no. 17, p. 2612, 2024, <https://doi.org/10.3390/math12172612>.
- [39] F. Zoghلامي, M. A. Aganj, and R. A. Aganj, "ViGLAD: Vision graph neural networks for logical anomaly detection," *IEEE Access*, vol. 12, pp. 173304–173315, 2024, <https://doi.org/10.1109/ACCESS.2024.3502514>.
- [40] V.-T. Le and Y.-G. Kim, "Attention-based residual autoencoder for video anomaly detection," *Appl. Intell.*, vol. 53, no. 3, pp. 3240–3254, 2023, <https://doi.org/10.1007/s10489-022-03613-1>.
- [41] P. Mishra, R. Verk, D. Fornasier, C. Picciarelli, and G. L. Foresti, "VT-ADL: A vision transformer network for image anomaly detection and localization," *arXiv preprint arXiv:2104.10036*, 2021, <https://doi.org/10.48550/arXiv.2104.10036>.
- [42] E. Caville, W. W. Lo, S. Layeghy, and M. Portmann, "Anomal-E: A self-supervised network intrusion detection system based on graph neural networks," *Knowledge-based systems*, vol. 258, p. 110030, 2020, <https://doi.org/10.1016/j.knosys.2022.110030>.
- [43] J. Li, X. Wang, Y. Zhao, and Z. Zhang, "CVTGAD: Simplified transformer with cross-view attention for unsupervised graph-level anomaly detection," *arXiv preprint arXiv:2405.02359*, 2024, <https://doi.org/10.48550/arXiv.2405.02359>.
- [44] N. Madan *et al.*, "Self-Supervised Masked Convolutional Transformer Block for Anomaly Detection," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 1, pp. 525–542, Jan. 2024, <https://doi.org/10.1109/TPAMI.2023.3322604>.
- [45] P. Bergmann, K. Batzner, M. Fauser, D. Sattlegger, and C. Steger, "The MVTec anomaly detection dataset: A comprehensive real-world dataset for unsupervised anomaly detection," *Int. J. Comput. Vis.*, vol. 129, no. 4, pp. 1038–1059, 2021, <https://doi.org/10.1007/s11263-020-01400-4>.
- [46] Y. Zou, J. Jeong, L. Pemula, D. Zhang, and O. Dabeer, "Spot-the-difference self-supervised pre-training for anomaly detection and segmentation," In *European conference on computer vision*, pp. 392–408, 2022, https://doi.org/10.1007/978-3-031-20056-4_23.

-
- [47] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2021, <https://doi.org/10.48550/arXiv.2010.11929>.
- [48] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," *arXiv preprint arXiv:2103.00020*, 2021, <https://doi.org/10.48550/arXiv.2103.00020>.
- [49] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992-10002, 2021, <https://doi.org/10.1109/ICCV48922.2021.00986>.
- [50] W. Wang, X. Zhang, Y. Wang, Z. Liu, and F. Meng, "N-STGAT: Spatio-temporal graph neural network for intrusion detection," *Remote Sens.*, vol. 15, no. 14, p. 3611, 2023, <https://doi.org/10.3390/rs15143611>.
- [51] A. A. M. Fatima, "A generative pattern extraction algorithm tackling ultra-imbalanced data in financial fraud detection systems," *PatternIQ Min. (PIQM)*, vol. 2, no. 4, pp. 19–41, 2025, <https://doi.org/10.70023/sahd/242502>.
- [52] A. G. Jaber, A. A. Ali, A. A. Mahmood, M. J. Salim, G. J. Mohammed, and K. A. Z. Ariffin, "Dynamic quantization-aware neural architecture search for real-time encrypted traffic classification in 5G networks," *Al-Iraqia J. Sci. Eng. Res.*, vol. 5, no. 1, pp. 34–48, 2026, <https://doi.org/10.58564/IJSER.5.1.2026.365>.
- [53] L. Wen *et al.*, "UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking," *Computer Vision and Image Understanding*, vol. 193, p. 102907, 2020, <https://doi.org/10.1016/j.cviu.2020.102907>.