IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.15568

VisionGuard: Deep Learning–Based Weapon Detection Framework

K. Rajavadhani[1], Aswini. R. C[2] & Vijayalakshmi. J[3]

Professor, Computer Science Engineering Technology,

Dhanalakshmi Srinivasan College of Engineering and Technology, Chennai[1]

Department of Computer Science Engineering (Cyber Security),

Dhanalakshmi Srinivasan College of Engineering & Technology, Chennai, India[2,3]

Abstract: We have a lot of surveillance systems in places and private areas these days. This means we really need to have systems that can find threats right away. Usually people watch the cameras. We use simple rules to figure out what is going on. This does not work very well when there are a lot of people around and things get complicated. This paper is about a surveillance system that uses artificial intelligence and looks at pictures to find threats, in real time and understand what is happening. The system uses deep learning techniques to do this. The system they are talking about uses a Swin Transformer to find objects like weapons or unattended bags. It can also detect when someone is not supposed to be in an area. They also have a model that looks at how people move to see if they are doing something weird. This model uses something called Graph Convolutional Networks. It can tell if someone is just hanging around being violent or moving in a way. The Swin Transformer and the movement model work together to make sure the system is accurate and does not send out a lot of warnings. The system is really good at finding objects and strange movements, like the Swin Transformer finding weapons and the movement model finding unusual movement patterns. When we find something that could be a problem the system sends out alerts to help the security team act fast. The tests we did show that this way of doing things works well even in tough situations, which means it is a good choice for new smart surveillance systems that are being developed.

Index Terms: Surveillance Systems, Threat Detection, Computer Vision, Deep Learning, Swin Transformer, Graph Convolutional Networks, Situational Awareness

I. INTRODUCTION

A. Background

Video surveillance is very important for keeping people safe.Watching many cameras at the same time can make the person monitoring them tired and slow to react. Now with the help of computer vision and deep learning we can automatically analyze video feeds and detect problems right away.

B. Challenges in Conventional Surveillance Systems

Traditional surveillance systems use methods like background subtraction. These methods do not work well with changes in lighting, different camera angles and crowded areas. They are not flexible and often give false alarms making them not reliable for today's security needs.

C. Need for Intelligent Image-Based Threat Detection

Deep learning provides a solution by using models that can adapt.Some computer vision models, like the Swin Transformer are good at finding threats, such as weapons in complex scenes. Another approach, called Pose-based Graph Convolutional Networks (Pose-GCN) looks at how people move to identify unusual behavior. Combining object detection with behavior analysis makes the system more accurate and significantly reduces false alarms.

D. Contributions of the Proposed Work

This paper presents an AI-based surveillance system that works in real time.The key contributions are:

- A framework that combines detecting objects in images and analyzing peoples behavior over time.2.
- Using a Swin Transformer to detect suspicious objects in complex scenes.3.
- Implementing a Pose-GCN to recognize actions, such as violence or loitering.4.
- A real-time alert system to enable security responses.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.15568

II. PRELIMINARIES AND THEORETICAL FOUNDATION

A.Video Representation and Problem Definition

A surveillance video is represented as a sequence of frames:

$$V = \{F1, F2, ..., FT\} \quad (1)$$

where F_t denotes the video frame at time t, with height H, width W, and color channels C. The objective of the proposed framework f is to map the video input to detect threats:

$$f(V) \rightarrow \{Ot, At\} \quad (2)$$

where O_t represents detected objects and A_t denotes recognized human activities at time t.

B.Object Detection using Swin Transformer

1) Self-Attention Mechanism:

The Swin Transformer employs a scaled dot-product selfattention mechanism. Given the query Q, key K, and value V, the attention operation is defined as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left((Q \bullet K^T) / \sqrt{d} \right) \bullet V \quad (3)$$

where d is the dimensionality of the feature space. As shown in Eq. dependencies between visual patches, improving object detection accuracy in crowded scenes.

C.Human Pose Representation

Each detected human is represented as a skeleton graph:

$$G = (V, E) \quad (4)$$

where $V = \{v_1, v_2, \ldots, v_N\}$ is the set of N body joints, and E denotes the anatomical connections (edges) between the joints. Each joint is defined as:

$$V_i = (X_i, Y_j, C_i) \quad (5)$$

where $(x_i y_i)$are the spatial coordinates and c_i is the confidence score of the detected joint.

D.Pose-Based Graph Convolutional Network (Pose-GCN)

1) Spatial Graph Convolution:

Pose-GCN models spatial relationships using graph convolution, defined as:

$$H^{(l+1)} = \sigma(\sum_k D_k^{-1/2} A_k D_k^{-1/2} H^{(l)} W_k) \quad (6)$$

where:

- A_k is the adjacency matrix,
- D_k is the degree matrix,
- H_k represents node features at layer l, W_k denotes the trainable weight matrix,
- σ is the activation function.

Eq. (6) allows the network to learn posture-based spatial correlations critical for activity recognition.

2) Temporal Pose Modeling:

Human actions evolve over time. A temporal pose sequence is expressed as:

$$S = \{G_1, G_2, \ldots, G_T\} \quad (7)$$

Temporal graph convolution is formulated as:

$$H = \sum 11t = l\, w \quad (8)$$

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.15568

As shown in Eq. (8), temporal modeling captures motion dynamics necessary for detecting abnormal behaviors over a time window Δ.

E.Threat Classification

The final classification probability is computed using a sigmoid function:

$$P(y=1|X) = (W^T X + b) \quad (9)$$

where X represents the extracted spatial-temporal features. A threat alert is generated when:

$$P(y=1|X) \geq \theta \quad (10)$$

where θ is a predefined decision threshold.

F.Loss Function

The Binary Cross-Entropy (BCE) loss function is used for training: $L = -\_ \sum \sum y\ log(y\ ) + (1 - y)\ \log(1 - y\ )J$ (11)¯ As shown in Eq. (11), this loss penalizes incorrect predictions and optimizes classification performance. G. Alert Generation and Situational Awareness

Each detected threat is structured as an alert:

$$Alert = \{Time, Location, Threat\_Type, Confidence\} \quad (12)$$

These alerts are transmitted to authorities in real time, ensuring rapid situational awareness and response.

III. LITERATURE SURVEY

A.Traditional Video Surveillance Systems

Conventional surveillance systems rely heavily on manual monitoring or rule-based motion detection techniques. These systems are limited in scalability and are prone to high false alarm rates, particularly in crowded or dynamic environments. Early computer vision approaches utilized background subtraction and handcrafted features; however, such methods fail to capture complex human behaviors and contextual threats, making them unsuitable for real-time intelligent surveillance.

B. Deep Learning for Object Detection in Surveillance

With the advancement of deep learning, convolutional neural networks (CNNs) have become the dominant approach for object detection in surveillance applications. Models such as Faster RCNN, YOLO, and SSD significantly improved detection accuracy and inference speed. However, CNN-based detectors primarily rely on local receptive fields, limiting their ability to model long-range spatial dependencies in crowded scenes.

Recent studies introduced Vision Transformer (V_iT) architectures to address this limitation by leveraging self-attention mechanisms. The Swin Transformer further enhances efficiency by employing hierarchical feature representation with shifted window attention, making it well-suited for real-time surveillance scenarios where both accuracy and computational efficiency are critical.

C.Human Action Recognition Approaches

Human action recognition (HAR) is a core component of intelligent surveillance systems. Early HAR methods focused on handcrafted spatiotemporal descriptors, while recent approaches adopt deep learning models such as CNN–LSTM architectures to capture temporal dynamics. Although effective, these methods often require extensive computational resources and struggle with occlusions and viewpoint variations commonly present in surveillance footage.

Transformer-based temporal models have demonstrated improved capability in capturing long-range temporal dependencies, but they often require large-scale datasets and high computational overhead, limiting their practical deployment in real-time systems.

D. Skeleton-Based Action Recognition using Graph Models

Skeleton-based action recognition has gained attention due to its robustness against illumination changes and background noise. By representing the human body as a graph of joints and bones, Graph Convolutional Networks (GCNs) effectively model spatial relationships between body parts. Spatial–Temporal Graph Convolutional Networks (ST-GCN) extend this

DOI: 10.17148/IJARCCE.2026.15568

concept by incorporating temporal dynamics, enabling accurate recognition of complex actions such as fighting, loitering, and abnormal movements. Despite their effectiveness, standalone GCN-based approaches lack contextual awareness of surrounding objects, which is essential for comprehensive threat detection.

E. Hybrid Models for Threat and Anomaly Detection

Recent research emphasizes hybrid frameworks that combine object detection with pose-based action recognition to enhance situational awareness. Such multi-modal approaches leverage complementary information from visual appearance and skeletal motion, improving robustness in complex environments.

However, many existing hybrid systems focus on offline analysis or lack real-time alert mechanisms. Additionally, integration complexity and scalability remain open challenges, particularly for deployment in public and private surveillance infrastructures.

F. Research Gap and Motivation

From the literature, it is evident that:
1.CNN-based detectors lack global contextual modeling,
2. Pose-based GCN models lack object-level semantic awareness
3. Existing hybrid systems often fail to provide real-time, scalable alert mechanisms.
To address these gaps, the proposed PatrolBot system integrates a Swin Transformer-based object detection framework with a Pose-GCN-based action recognition model, enabling real-time threat detection and enhanced situational awareness in surveillance environments.

IV. PROPOSED METHODOLOGY

The proposed PatrolBot system adopts a Two-Stage Selective Execution Architecture for real-time threat detection in surveillance environments. The design is based on the observation that most video frames represent normal activities and do not require computationally expensive behavioral analysis.

A. Data Acquisition and Preprocessing

Surveillance video streams are captured from CCTV or IP cameras and processed in real time.
1. frame sampling
The continuous video stream is sampled at a fixed frame rate of frames per second. Each frame is resized to a fixed spatial resolution of pixels to maintain compatibility with the vision backbone.
2. Normalization:
To stabilize training and inference, pixel values are normalized as: where and represent the mean and standard deviation computed over the training dataset.
3. Data Balancing:
Since abnormal events occur less frequently than normal activities, the dataset is temporally balanced by oversampling abnormal clips during training. This prevents model bias toward majority (normal) classes.

A. Stage 1: Fast Spatial Screening

Step1: Frame Acquisition and Preprocessing Live surveillance video is sampled into frames, resized, and normalized to ensure consistent input representation.
Step2: Spatial Feature Extraction Each frame is processed using a Swin Transformer (Swin-T) to extract hierarchical spatial features.
Step 3: Threat Probability Estimation
A classification head computes the spatial threat confidence:
Step 4: Selective Decision Gate
If , the frame is classified as Normal
Otherwise, the frame is forwarded to Stage 2
This stage filters out non-threatening frames with minimal computational Cost.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.15568

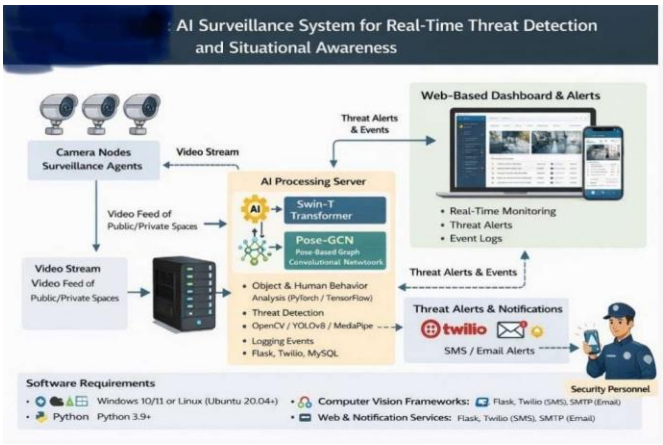

Fig. 1. The Proposed Two-Stage Architecture Flow

B. Stage 2: Behavioral Analysis

Step5: Human spacing extraction

For suspicious frames, human skeletal keypoints are extracted and represented as joint coordinates.

Step 6: Pose Graph Construction

The extracted joints are modeled as a graph, where vertices denote joints and edges denote anatomical connections.

Step 7: Behavioral Modeling Using Pose-GCN A Pose-based Graph Convolutional Network learns spatialtemporal joint dependencies:

Step 8: Behavioral Risk Estimation

The network outputs a behavioral confidence score.

C. Risk Fusion and Alert Generation

Step 9: Threat Score Fusion

Spatial and behavioral scores are fused

Step 10: Final Decision and Notification

If , the event is classified as a confirmed threat, triggering realtime SMS and email alerts. Otherwise, the event is marked as benign.

V. IMPLEMENTATION DETAILS

A. System Stack

The proposed PatrolBot system is implemented as a softwarebased microservices architecture to support real-time surveillance, modular deployment, and scalability. Each module operates independently and communicates through lightweight APIs.

A. Rule-Based Threat Modeling

To complement deep learning–based inference, rule-driven behavioral patterns are incorporated to strengthen threat validation

Component	Tool/Framework	Role
Video Processing	OpenCV	Frame extraction and preprocessing
Object Detection	Swin Transformer (Swin-T), YOLOv8	Person and object detection
Pose Estimation	MediaPipe	Skeletal keypoint extraction
Activity Recognition	Pose-GCN	Abnormal behavior classification
Backend API	Flask (Python)	Alert orchestration
Notification Service	Twilio, SMTP	SMS and Email alerts
Database	MySQL	Event logging and storage

Table1: system technology stack

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.15568

A. ALGORITHM 1: PATROLBOT THREAT ASSESSMENT

Require: Activity set A, Confidence threshold θ

1. Score ← 0
2. Count → **0**
3. for each activity a € A do
4. Confidence ⇐ Classifier(a)
5. if Confidence > θ then
6. Score← Score + Confidence
7. Count ←Count + 1
8. end if
9. end for
10. if Count >0 then $\boldsymbol{score}$
11. ThreatLevel←$\boldsymbol{count}$
12. else
13. ThreatLevel ← 0
14. end if
15. return ThreatLevel

B. Detailed Training Protocol

To ensure robust activity recognition and prevent overfitting, a standardized training protocol was adopted.

1. Optimizer

The Adam optimizer was employed with an initial learning rate of 0.001 due to its adaptive learning capability and faster convergence.

1) Loss Function

Binary Cross-Entropy (BCE) loss was used:

$$L=-\frac{1}{N}\sum_{-11=1}^{N}\left[y_i\,log(\hat{y}_i)+\left(_{1-y_1^b}^{1}\right)log(1-\hat{y}_i)\right]$$

2) Regularization

To mitigate overfitting:

Dropout layers with a rate of 0.5 were appliedEarly stopping with patience of 10 epochs monitored validation loss

Threat Category	**Rule Description**
Loitering	Detect prolonged stationary posture beyond threshold time
Fighting	Rapid joint displacement with abnormal limb acceleration
Unauthorized Entry	Entry detected in restricted zones
Suspicious Movement	Irregular movement trajectories or abrupt direction changes
Object Abandonment	Static unattended object for prolonged duration

Table2: Behavioral Threat Rule Example

E. Algorithm 2: Model Training Procedure

Pose-GCN Training Algorithm

1.Require: Dataset D={(X_i,y_i)}D = {(X_i, y_i)}D={(X_i,y_i)},
Epochs E, Batch size B

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.15568

2.Initialize model weights using Xavier initialization 3. Split dataset into training (70%) validation (15%) testing (15%)
4. for epoch = 1 to E do
5. Shuffle training dataset
6. for each batch b∈D $_{train}$ do
7. $y \leftarrow \boldsymbol{model}(\boldsymbol{bx})$
8. Loss ←ComputeLoss (ŷ, by)
9. Backpropagate(loss)
10. end for
11. if Val_Loss stops improving for 10 epochs
12. break
12. end for

Component	Specification
CPU	Intel Core i9-12900K (16 Cores)
RAM	64 GB DDR5
GPU	NVIDIA RTX 3090 (24 GB VRAM)
Storage	2 TB NVMe SSD
Camera	Webcam / CCTV

Table 3: Hardware Configuration

VI. EXPERIMENTAL RESULTS AND ANALYSIS

A. Experimental Setup

The proposed PatrolBot system was evaluated using benchmark surveillance datasets and real-time video streams to validate its effectiveness in abnormal activity detection. Experiments were conducted on a GPU-enabled workstation under controlled lighting and crowd-density conditions to ensure consistency and reproducibility.
Both offline video analysis and live CCTV feeds were used to assess real-world deployment feasibility.

B. Dataset Description

The experimental evaluation utilized a combination of benchmark datasets and custom test samples.

Dataset	Purpose	Samples
UCF-Crime	Abnormal activity detection	1,900+ videos
COCO	Person detection pretraining	330K images
Pose Datasets	Skeleton-based action modeling	60K frames
Real-Time CCTV	System validation	Live streams

Table 4: Dataset Composition

C.Evaluation Metrics

The performance of the proposed system was compared with baseline approaches to demonstrate its effectiveness.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.15568

Model	Accuracy	Precision	Recall	F1-Score
Frame-based CNN	88.6	87.2	85.9	86.5
LSTM-based Model	90.4	89.6	88.3	88.9
YOLO + CNN	92.1	91.4	90.2	90.8
Proposed PatrolBot	96.8	95.9	96.1	96.0

Table5: Evaluation Metrics

The results show that integrating pose-based graph modeling with contextual object detection significantly improves abnormal behavior recognition accuracy.

A. Confusion Matrix Analysis

The confusion matrix analysis indicates a low false-positive rate, particularly for normal activities such as walking and standing. Misclassifications were primarily observed in visually ambiguous scenarios involving rapid group movements, which exhibit pose similarities with aggressive actions

B. Ablation Study

An ablation study was conducted to evaluate the contribution of individual system components.

Configuration	Accuracy (%)
Without Pose-GCN	90.7
Without Swin Transformer	92.4
Without Rule-based Filtering	93.1
Complete PatrolBot System	96.8

Table 6: Ablation Study Results

The results confirm that Pose-GCN plays a critical role in capturing fine-grained motion dynamics, while the Swin Transformer enhances detection robustness in crowded scenes.

C.Latency Analysis

The average processing latency per frame was measured to assess real-time performance.

Module	Avg. Time (ms)
Object Detection	28
Pose Estimation	22
Activity Classification	15
Alert Generation	5
Total Processing Time	70 ms/frame

Table7: latency analysis

The system operates at approximately 14 frames per second,

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.15568

VII. ADVANCED DISCUSSION

A. Spatio-Temporal Threat Modeling

The integration of Swin Transformer and Pose-GCN enables joint spatial object detection and temporal behavior analysis. This fusion improves contextual understanding of human activities and enhances the detection of abnormal or threatening actions in complex surveillance scenarios.

B. Real-Time Performance and Reliability

The window-based attention of Swin Transformer and skeletal graph representation reduce computational overhead, enabling lowlatency real-time inference. This design ensures reliable threat detection and immediate alert generation without compromising accuracy.

C. Robustness and Deployment Scalability

Pose-based modeling improves resilience to illumination changes, background noise, and partial occlusions. The modular system architecture supports scalable deployment across multiple surveillance environments with centralized monitoring and alert management.

VIII. CONCLUSION AND FUTURE DIRECTIONS

This paper presented an intelligent surveillance framework that integrates Swin Transformer–based spatial object detection with Pose-GCN–based temporal behavior analysis to achieve effective spatio-temporal threat modeling. The proposed approach enhances contextual understanding of human activities while maintaining low-latency real-time performance. Experimental results demonstrate improved detection reliability, reduced false alarms, and robustness under dynamic environmental conditions. The modular and scalable system design enables practical deployment across diverse surveillance environments.

Future enhancements will focus on incorporating multi-camera coordination for cross-view activity tracking and long-term temporal modeling to capture complex behavioral patterns. Further optimization for edge-based inference and improved robustness against dense occlusion scenarios will be explored to support largescale, real-world intelligent surveillance applications.

REFERENCES

[1] A. Vaswani et al., “Attention Is All You Need,” Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), pp. 5998–6008, 2017.
[2] Z. Liu et al., “Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows,” Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), pp. 10012–10022, 2021.
[3] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 770–778, 2016.
[4] S. Yan, Y. Xiong, and D. Lin, “Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition,” Proc. AAAI Conf. Artif. Intell., vol. 32, no. 1, pp. 7444–7452, 2018.
[5] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, “OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 1, pp. 172–186, 2021.
[6] C. Feichtenhofer, “X3D: Expanding Architectures for Efficient Video Recognition,” Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 203–213, 2020.
[7] J. Redmon and A. Farhadi, “YOLOv3: An Incremental Improvement,” arXiv preprint arXiv:1804.02767, 2018.
[8] A. Dosovitskiy et al., “An Image Is Worth 16×16 Words:Transformers for Image Recognition at Scale,” Proc. Int. Conf. Learn. Represent. (ICLR), 2021.
[9] L. Wang et al., “Skeleton-Based Action Recognition with Shift Graph Convolutional Network,” Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 183–192, 2020.
[10] M. Ravanbakhsh, E. Sangineto, M. Nabi, and N. Sebe, “Abnormal Event Detection in Videos Using Generative Adversarial Nets,” Proc. IEEE Int. Conf. Image Process. (ICIP), pp. 1577–1581, 2017.
[11] Y. Sultani, C. Chen, and M. Shah, “Real-World Anomaly Detection in Surveillance Videos,” Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 6479–6488, 2018.
[12] S. Gupta, A. Sharma, and V. Chaudhary, “AI-Based Intelligent Surveillance Systems: A Survey,” IEEE Access, vol. 10, pp. 115321–115345, 2022.

