Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 14, No. 3, 2026, 351-365.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

Context-Aware Criminal Activity Recognition in Surveillance Images Using an Attention-Guided YOLOv10-Vision Transformer Framework

Samira Mavaddati*

Electronic Department, Faculty of Engineering and Technology, University of Mazandaran, Babolsar, Iran.

Article Info

Article History:
Received 29 April 2026
Revised 13 June 2026
Accepted 29 June 2026

DOI:10.22044/jadm.2026.17656.2925

Keywords:
Criminal Activity Recognition; Intelligent Surveillance Systems; YOLOv10; Vision Transformer; Deep Learning.

**Corresponding author: s.mavaddati@umz.ac.ir (S. Mavaddati).*

Abstract

The rapid growth of intelligent surveillance systems has increased the demand for accurate and efficient methods of criminal activity recognition capable of operating in real-world environments. Although conventional deep learning and object detection frameworks have demonstrated promising performance, they often struggle to capture long-range contextual dependencies and complex interactions present in surveillance scenes. To address these limitations, this study proposes a hybrid deep learning framework that combines the real-time detection capability of YOLOv10 with the global contextual modeling power of Vision Transformers (ViT). An attention-guided feature fusion mechanism is introduced to effectively integrate local spatial representations extracted by YOLOv10 with global semantic features generated by the transformer architecture. The proposed framework is evaluated on the UCF-Crime dataset, which consists of fourteen categories of normal and criminal activities, including burglary, robbery, assault, vandalism, shoplifting, and abuse. Surveillance videos are converted into image sequences and analyzed under two experimental scenarios: (I) a standalone YOLOv10 model and (II) the proposed Attention-Guided YOLOv10-ViT framework. Performance is assessed using accuracy, precision, recall, and F1-score metrics. Experimental results show that the standalone YOLOv10 model achieves an overall classification accuracy of 88.07%, outperforming the previously reported YOLOv8 baseline. More importantly, the proposed hybrid framework attains an accuracy of 93.45%, exceeding both YOLOv10 and earlier YOLOv8-ViT architectures. The improvement is particularly evident in challenging scenarios involving occlusion, illumination changes, cluttered backgrounds, and crowded environments. The results demonstrate that integrating YOLOv10, Transformers, and attention-guided feature fusion provides a scalable, robust, and real-time solution for intelligent surveillance and public monitoring applications.

1. Introduction

In recent years, the expansion of urban populations alongside the rapid growth of public infrastructure has made the issue of public safety increasingly critical. The rising occurrence of criminal behaviors in densely populated areas, which are typically monitored through extensive surveillance camera networks, has created a strong demand for intelligent, automated, and highly reliable monitoring systems. Conventional approaches based on manual inspection or handcrafted feature engineering are not only labor-intensive but also unsuitable for large-scale and real-time

surveillance scenarios. In this context, deep learning has emerged as a powerful paradigm for automatic visual data interpretation, significantly enhancing the efficiency and reliability of crime detection systems [1–2]. Deep neural networks have demonstrated remarkable capabilities in learning discriminative feature representations and capturing complex behavioral patterns from visual data. Among them, the YOLO family has achieved widespread adoption for real-time object detection due to its high speed and competitive accuracy. However, convolution-based architectures still have limitations in modeling long-range dependencies and global contextual relationships in complex surveillance scenes. Vision Transformers address these limitations by employing self-attention mechanisms to capture global interactions across image regions and improve contextual understanding [3–5].

Building on these advances, the present study focuses on image-based criminal activity recognition using a hybrid deep learning approach that integrates YOLOv10 with Vision Transformers. The primary contribution of this work is the development of an attention-guided YOLOv10–ViT framework, designed to effectively combine the real-time local feature extraction capability of YOLOv10 with the global contextual representation power of ViT. The proposed method is evaluated on the UCF-Crime dataset, which includes fourteen categories of normal and abnormal activities such as burglary, robbery, assault, vandalism, shoplifting, and abusive behavior. In the experimental design, surveillance videos are decomposed into frame-level image sequences and analyzed under two distinct scenarios: (i) a standalone YOLOv10-based model and (ii) the proposed attention-guided YOLOv10–ViT fusion architecture. Performance evaluation is conducted using standard metrics, including accuracy, precision, recall, and F1-score. The obtained results indicate that the standalone YOLOv10 model achieves a classification accuracy of 88.94%, surpassing previously reported YOLOv8-based baselines. More importantly, the proposed hybrid YOLOv10–ViT framework significantly improves performance, achieving an overall accuracy of 93.45%, and outperforming both the standalone YOLOv10 model and earlier YOLOv8–ViT hybrid approaches. This improvement is particularly evident in challenging surveillance conditions, including occlusions, illumination variations, cluttered backgrounds, and densely populated scenes. Overall, the findings confirm that integrating YOLOv10 with Vision Transformers through an attention-guided feature fusion strategy provides a robust, scalable, and real-time solution for intelligent surveillance systems. This hybrid design effectively enhances both local feature extraction and global contextual reasoning, making it highly suitable for advanced public safety and crime monitoring applications. The key contributions of this study can be summarized as follows:

- A comprehensive comparative analysis between YOLOv8 and the proposed YOLOv10–ViT framework, demonstrating the advantages of incorporating transformer-based attention mechanisms for enhanced global contextual learning in surveillance imagery.
- Utilization of the UCF-Crime dataset in a frame-level image representation setting, showing that efficient crime recognition can be achieved without relying on full video sequence processing, thereby improving computational efficiency and scalability.
- Experimental evidence indicating that the YOLOv10–ViT model achieves superior classification accuracy compared to YOLOv10 and YOLOv8 baselines, particularly in complex categories such as burglary, shoplifting, and vandalism.
- Demonstration of improved robustness under real-world surveillance conditions, including occlusion, lighting variations, and crowded environments, confirming the suitability of the proposed model for practical deployment.
- Validation that hybrid YOLOv10–ViT architectures represent a promising direction for next-generation intelligent surveillance systems focused on public safety enhancement.

The remainder of this paper is organized as follows. Section 2 reviews related literature in crime detection and deep learning-based surveillance systems. Section 3 describes the dataset, preprocessing pipeline, and the proposed methodology. Section 4 presents the experimental setup and comparative results between YOLOv10 and YOLOv10–ViT. Section 5 discusses findings, implications, and limitations of the study. Finally, Section 6 concludes the paper and suggests future research directions.

2. Review of Related Works

Over the past decade, research in automated crime and anomaly detection has experienced substantial progress, primarily driven by advancements in deep learning methodologies. Early solutions relied heavily on handcrafted feature descriptors and traditional machine learning techniques; however, these approaches often exhibited limited

generalization capabilities when applied to complex and heterogeneous surveillance environments. With the rapid expansion of large-scale datasets and improvements in computational resources, convolutional neural networks (CNNs) and their advanced variants have become the foundation of modern visual recognition systems for both image- and video-based surveillance applications.

In recent years, attention-based architectures have introduced a new direction in visual representation learning. In particular, Vision Transformers have gained significant attention due to their ability to model long-range dependencies and capture global contextual relationships across image regions. Unlike CNN-based models, which primarily focus on local receptive fields, ViTs provide a more holistic understanding of scene-level interactions. This advancement has encouraged researchers to explore hybrid frameworks that combine efficient object detection models with global attention mechanisms to enhance anomaly recognition in real-world surveillance scenarios. Within this context, the present section provides a structured overview of recent developments in intelligent crime detection systems, emphasizing both methodological strengths and inherent limitations. It also highlights emerging research directions that motivate the integration of YOLO-based detectors with Vision Transformer architectures, particularly in real-time surveillance applications.

Several studies have contributed to advancing automated crime detection using deep learning frameworks. For example, the approach in [6] focuses on real-time identification of criminal activities in both video streams and still images, while simultaneously generating alerts for human operators and relevant authorities. This system leverages a pre-trained VGG19 network for detecting weapons such as knives and firearms in threatening scenarios. A comprehensive analysis of crime prediction methodologies is presented in [7], where various machine learning and deep learning models are systematically reviewed. This work also provides access to commonly used datasets, offering valuable insights for developing effective predictive crime analytics systems. Addressing unconventional crime channels, the method proposed in [8] focuses on detecting threats in voice communication. It employs a one-dimensional convolutional neural network (1D-CNN) followed by a multilayer perceptron classifier, achieving strong performance on Bengali speech datasets and demonstrating the feasibility of audio-based threat detection systems. In [9], a real-time surveillance framework is introduced that integrates VGG19 for weapon detection and Faster R-CNN for object localization. This combination enables accurate bounding box prediction and improves reliability in dynamic surveillance environments. A detailed review of deep learning-based crime detection systems is presented in [10], which discusses architectural designs, methodological advancements, and practical deployment challenges, including ethical and legal considerations in automated surveillance systems.

The study in [11] proposes a real-time crime detection framework based on YOLO, capable of processing video streams at high speed while identifying suspicious individuals and weapons even in crowded environments, achieving approximately 45 frames per second. A multimodal fusion strategy is introduced in [12], where spatial, temporal, and environmental features are integrated using a combination of deep neural networks and classical machine learning techniques, improving both predictive accuracy and interpretability. In [13], a multi-stage surveillance pipeline is proposed that combines YOLOv5 for weapon detection, MobileNet for violence recognition, and a custom face detection module, enabling comprehensive monitoring of complex crime scenarios in CCTV systems. The work in [14] addresses large-scale crime-related object detection by employing deep learning architectures such as VGG19, ResNet, and GoogleNet, supporting forensic analysis and crime scene reconstruction through effective object tracking. In [15], transfer learning using Inception-v3 is applied to classify abnormal behavior in surveillance footage. By fine-tuning pre-trained weights on labeled datasets, the model achieves high accuracy in distinguishing normal and abnormal activities. The approach in [16] introduces an autoencoder-based anomaly detection model that utilizes reconstruction error thresholds for identifying abnormal events, achieving strong performance metrics on the FBI crime dataset. In [17], a spiking neural network-based multi-scale fusion model is proposed for video anomaly detection using event-based dynamic vision sensors. The method reduces redundant information while improving temporal motion representation, establishing a strong baseline for neuromorphic surveillance systems.

A high-performance deep learning system is presented in [18], where EfficientNet-B7 and ResNet50 are trained on the UCF-Crime dataset, achieving exceptional classification results in crowded surveillance scenarios. In [19], the VideoMAE transformer model is fine-tuned for

video understanding tasks across multiple domains, including crime detection and action recognition, demonstrating strong transfer learning capabilities despite class imbalance challenges.

A trajectory-based human action recognition framework is proposed in [20], where skeletal motion patterns are used to classify crime-related behaviors, enhancing dataset annotations and improving classification performance through advanced feature fusion strategies. Building upon these prior works, the present study investigates a hybrid deep learning framework based on YOLOv10 and Vision Transformers for real-time crime detection using frame-level images extracted from surveillance videos. The main objective is to evaluate the impact of integrating ViT into YOLOv10 for enhancing both local feature extraction and global contextual reasoning, enabling a more robust and discriminative representation for crime recognition tasks.

3. Dataset Description and Preprocessing Pipeline

To evaluate the performance of the proposed framework, the UCF-Crime dataset [21] is utilized. This dataset is among the most comprehensive benchmarks for anomaly detection in surveillance environments, comprising approximately 128 hours of real-world security footage. It includes nearly 1,900 long untrimmed videos covering a wide range of activities, including 13 categories of abnormal events such as Abuse, Arrest, Arson, Assault, Explosion, Fighting, Traffic Accidents, Robbery, Shooting, Shoplifting, Stealing, Vandalism, Burglary, as well as normal behavior scenarios. Representative samples of each category are illustrated in Figure 1, while Figure 2 demonstrates the class-wise distribution of the dataset.

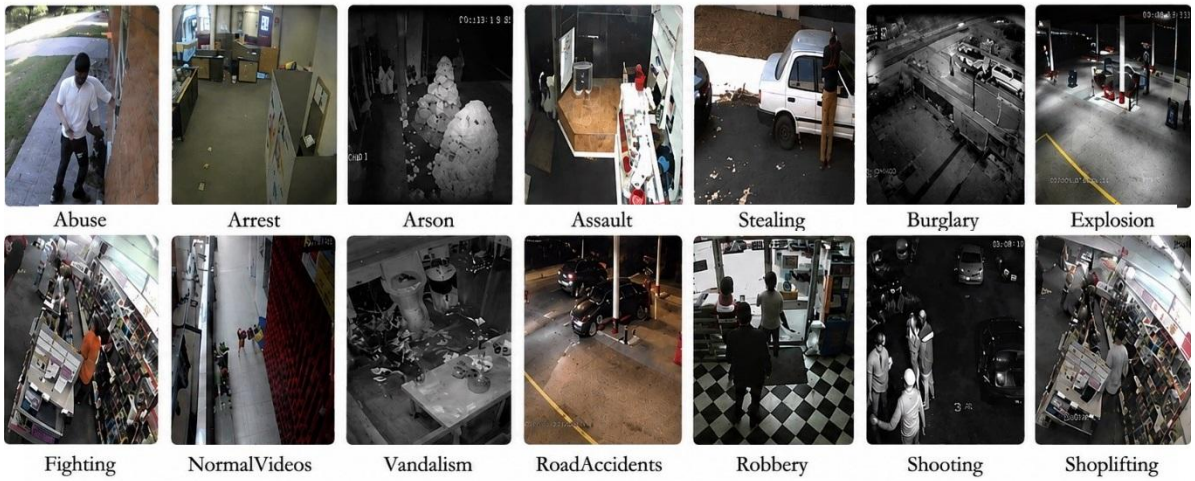

Figure 1. Representative samples from the 14 classes of the UCF-Crime dataset, illustrating different normal and abnormal surveillance scenarios, including various criminal activities.

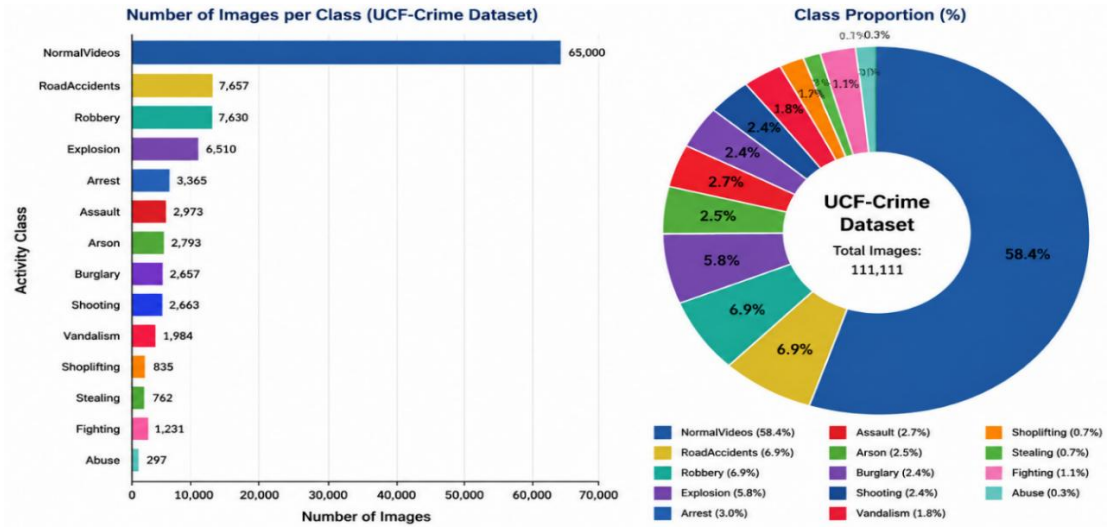

Figure 2. Distribution of image samples in the UCF-Crime dataset. The donut chart illustrates the relative proportion of each class, while the bar chart presents the corresponding number of images per category.

Each video sequence is annotated with rich contextual metadata, including event type, temporal information, location, and involved subjects, making it a highly valuable resource for surveillance-based learning tasks.

Table 1 summarizes the dataset partitioning and augmentation procedure used in this study. The dataset was divided into training and testing sets, and augmentation techniques were applied to the training samples to enhance data diversity and improve the robustness of the proposed models.

Prior to training, surveillance videos are converted into image frames, which are resized to 224 × 224 pixels, normalized, and scaled to [0, 1] to ensure compatibility with the network and stable optimization [22–23]. Random cropping is also applied to emphasize salient regions and reduce background interference. Furthermore, data augmentation is employed to increase sample diversity, improve model generalization, and mitigate overfitting caused by the limited availability of annotated surveillance data.

Data augmentation is a widely used technique in deep learning to improve model generalization and robustness across different architectures, including CNNs, transformer-based models, and hybrid frameworks, especially in video analysis and surveillance tasks. It helps reduce overfitting by increasing data diversity through various spatial and temporal transformations. In this study, data augmentation is also applied during training to enhance the generalization ability of the proposed YOLOv10–Vision Transformer framework on the UCF-Crime dataset under diverse surveillance conditions [24-25].

Table 1. Dataset partitioning and augmentation strategy used in this study.

Class	Original Images	Train (80%)	Test (20%)	Train After Augmentation	Total Images Used
Abuse	297	238	59	1,500	1,559
Arrest	3,365	2,692	673	5,000	5,673
Arson	2,793	2,234	559	5,000	5,559
Assault	2,657	2,126	531	5,000	5,531
Burglary	7,657	6,126	1,531	7,600	9,131
Explosion	6,510	5,208	1,302	6,500	7,802
Fighting	1,231	985	246	5,000	5,246
Normal Videos	65,000	52,000	13,000	10,000	23,000
Road Accidents	2,663	2,130	533	5,000	5,533
Robbery	835	668	167	5,000	5,167
Shooting	7,630	6,104	1,526	7,600	9,126
Shoplifting	7,623	6,098	1,525	7,600	9,125
Stealing	1,984	1,587	397	5,000	5,397
Vandalism	1,111	889	222	5,000	5,222

To mitigate this issue, a combined balancing strategy is adopted. Specifically, dominant classes are downsampled to reduce overrepresentation, while minority classes are augmented using the aforementioned transformation techniques. Importantly, augmentation is applied exclusively to the training set, whereas validation and test sets remain unchanged to ensure unbiased and realistic performance evaluation.

Following this division, augmentation is performed only on the training data to further enrich minority classes and achieve a more balanced class distribution. This strategy enhances the diversity of training samples and improves the model's ability to learn discriminative features across all categories. Table 2 summarizes the distribution of samples per class before and after augmentation, along with the allocation across training, validation, and testing sets. The results clearly demonstrate the effectiveness of the proposed balancing and augmentation strategy in constructing a more uniform dataset, which ultimately contributes to improved learning stability and classification performance of the proposed framework.

Table 2. Architectural Configuration and Training Components of YOLOv8, YOLOv10, and YOLOv10–ViT Models.

Network Variant	Feature Extraction Backbone	Global Context Module	Transformer Depth	Feature Dimension	Attention Configuration	Input Resolution	Prediction / Fusion Strategy
YOLOv8 (Baseline)	CSPDarknet53	Not Used	—	—	—	224×224	YOLO detection head
YOLOv10 (Baseline)	YOLOv10 Backbone	Not Used	—	—	—	224×224	YOLO detection head
YOLOv10–ViT (Proposed)	YOLOv10 + CNN Feature Encoder	Vision Transformer	12	768	12 heads	224×224	Attention-guided fusion + classifier

4. Proposed YOLOv10–ViT Methodology

In this study, crime analysis is formulated as a frame-level multi-class classification problem rather than an object detection task. YOLOv10 is employed to extract discriminative local spatial features, while the Vision Transformer captures global contextual dependencies. These complementary representations are adaptively integrated through an attention-guided feature fusion mechanism to enhance classification performance.

Experimental evaluation is conducted under two configurations: (i) a baseline YOLOv10 model and (ii) the proposed YOLOv10–ViT hybrid architecture. Results show that YOLOv10 achieves strong performance in real-time crime detection, while the integration of ViT further improves accuracy by enhancing global contextual understanding and feature discrimination. The proposed model consistently outperforms both YOLOv10 and previous YOLOv8-based hybrid approaches, particularly in challenging categories such as burglary, shoplifting, and vandalism. Overall, the findings confirm that the attention-guided YOLOv10–ViT framework provides a more robust and accurate solution for intelligent surveillance systems. The key outcomes of this study can be summarized as follows:

- The baseline YOLOv10 model demonstrates efficient real-time performance and strong detection capability in image-based crime recognition tasks.
- The proposed YOLOv10–ViT model significantly improves classification accuracy by integrating local feature extraction with global contextual reasoning.
- Noticeable performance gains are achieved in challenging categories such as burglary, shoplifting, and vandalism.
- The hybrid architecture exhibits strong robustness under real-world conditions, including occlusion, lighting variations, and crowded scenes.

Table 3 summarizes the layer-wise architectures of YOLOv8, YOLOv10, and the proposed YOLOv10–ViT model. The comparison illustrates the structural differences among the three networks and highlights the additional Vision Transformer components incorporated to enhance global contextual learning and classification performance.

4.1. YOLOv10-Based Local Feature Extraction

In the proposed framework, the first stage is dedicated to extracting discriminative local spatial features from surveillance frames using the YOLOv10 architecture. YOLOv10 is a state-of-the-art single-stage object detector that performs efficient feature extraction and localization in a unified manner. Given an input image $I \in^{H\times W\times 3}$, the backbone network of YOLOv10 generates a hierarchical feature representation:

$$F^{yolo} = \phi_{yolo}(I) \quad (1)$$

where ϕ_{yolo} denotes the convolutional backbone and neck feature aggregation layers, and $F^{yolo} \in^{H\times W\times d}$ represents the extracted local feature map [26]. These features primarily encode fine-grained spatial patterns such as object boundaries, human poses, motion-related cues, and interaction regions, which are essential for crime-related activity recognition.

Table 3. Layer-Wise Architectural Flow of YOLOv8, YOLOv10, and the Proposed YOLOv10–ViT Model for 14-Class Surveillance Classification.

Stage Index	Processing Block	Feature Map / Token Size (YOLOv8)	Feature Map / Token Size (YOLOv10)	Feature Map / Token Size (YOLOv10–ViT)	Operation Type	Key Parameters	Functional Role
1	Stem Layer	224×224×32	224×224×32	224×224×32	Conv 3×3	32 filters	Initial feature extraction
2	Early Feature Block	112×112×64	112×112×64	112×112×64	Conv + Downsampling	stride=2	Low-level feature learning
3	Intermediate Block	56×56×128	56×56×128	56×56×128	Conv + CSP/RepBlock	128 filters	Spatial feature refinement
4	Deep Feature Block	28×28×256	28×28×256	28×28×256	CSP / Rep-style block	256 filters	High-level semantics
5	High-Level Encoder	14×14×512	14×14×512	14×14×512	Conv aggregation	512 filters	Abstract representation
6	Tokenization Stage	—	—	196×768	Patch Embedding (16×16)	768-d embedding	ViT input construction
7	Global Context Modeling	—	—	196×768	Transformer Encoder (L=12)	12 heads	Long-range dependency modeling
8	Fusion Layer	—	—	1×D	Attention-based fusion	learnable weights	Local + global integration
9	Classification Head	1×14	1×14	1×14	Fully Connected + Softmax	14 classes	Final prediction

However, YOLOv10 alone is limited in capturing long-range dependencies across the scene, motivating the integration with a transformer-based module.

4.2. Vision Transformer for Global Context Modeling

To complement the local feature learning of YOLOv10, a Vision Transformer is employed to capture global contextual relationships. The feature map F^{yolo} is first reshaped into a sequence of flattened patches:

$$Z = \{z_1, z_2, ..., z_N\}, z_i \in R^d \tag{2}$$

These tokens are linearly embedded and augmented with positional encodings to preserve spatial structure. The transformer encoder processes the sequence through
multi-head self-attention (MSA):

$$MSA(Q,K,V) = soft\max(\frac{QK^T}{\sqrt{d_k}})V \tag{3}$$

where Q, K, and V are query, key, and value matrices derived from the input tokens. The output of the ViT encoder is a globally enriched representation:

$$F^{vit} = \phi_{vit}(Z) \tag{4}$$

This representation captures long-range dependencies such as interactions between multiple individuals and contextual cues across the entire scene [26-27].

4.3. Attention-Guided Feature Fusion Mechanism

A key contribution of the proposed method is the attention-guided fusion strategy that integrates local YOLOv10 features with global ViT representations. Instead of simple concatenation, an adaptive weighting mechanism is introduced to dynamically balance the contribution of each modality. The fusion process is defined as:

$$F^{fusion} = \alpha F^{yolo} + (1-\alpha)F^{vit} \tag{5}$$

where $\alpha \in [0,1]$ is an attention coefficient learned through a lightweight gating network:

$$\alpha = \sigma(W_\alpha[F^{yolo}; F^{vit}] + b_\alpha) \tag{6}$$

Here, $[.;.]$ denotes feature concatenation, W_α and b_α are learnable parameters, and $\sigma(.)$ is the sigmoid activation function. This mechanism allows the model to adaptively emphasize local or global features depending on the complexity of the scene. For instance, in occluded or crowded environments, global contextual reasoning becomes more dominant, while in clear scenes, local object-level features are more influential.

4.4. Classification and Prediction Module

The fused representation F^{fusion} is passed to a fully connected classification head to predict the probability distribution over crime categories:

$$\hat{y} = soft\max(W_c.GAP(F^{fusion}) + b_c) \tag{7}$$

where $GAP(.)$ denotes global average pooling, and W_c, b_c are trainable parameters. The model is trained using the categorical cross-entropy loss function:

$$L = -\sum_{i=1}^{c} y_i \log \hat{y}_i \tag{8}$$

where C is the number of classes in the UCF-Crime dataset. Table 4 presents the training and optimization parameters employed for YOLOv8, YOLOv10, and YOLOv10–ViT. The listed hyperparameters define the experimental setup used to train and evaluate the models under consistent conditions [26-27].

4.5. Training Strategy and Optimization

The proposed network is trained end-to-end using stochastic gradient descent with adaptive learning rate scheduling. To improve generalization, data augmentation techniques such as random cropping, horizontal flipping, and brightness adjustment are applied to the input frames. Additionally, dropout regularization is applied in the classification head to mitigate overfitting, especially given the high intra-class variability in surveillance data. The features extracted from the YOLOv10 and Vision Transformer branches are fused and then passed through a series of fully connected layers to generate final class probability distributions. To provide a comprehensive understanding of the proposed architecture, the overall workflow is illustrated in Figure 3. This figure depicts the end-to-end pipeline, starting with input surveillance images, followed by feature extraction using YOLOv10 and ViT modules, and concluding with the final classification stage. In addition, Algorithm 1 presents the pseudocode of the proposed YOLOv10–ViT framework, describing the complete sequence of operations from data preprocessing and feature extraction to fusion, classification, and inference. This structured representation enables a clear interpretation of the

hybrid learning strategy prior to the experimental evaluation.

4.6. Evaluation Protocol

To rigorously evaluate the performance of the proposed YOLOv10–ViT classification framework, a comprehensive set of evaluation metrics is employed. These metrics are selected to assess both overall performance and class-wise discrimination capability across all 14 activity categories in the UCF-Crime dataset. The primary metric is classification accuracy (Acc), which quantifies the proportion of correctly predicted samples over the entire dataset:

$$Acc = (TP + TN) / (TP + TN + FP + FN) \quad (9)$$

Specificity (Spe) measures the ability of the model to correctly identify negative samples and reduce false alarms:

$$Spe = TN / (TN + FP) \quad (10)$$

Precision (Ppr) evaluates the reliability of positive predictions made by the model:

$$Ppr = TP / (TP + FP) \quad (11)$$

Sensitivity (Sen), also known as recall, reflects the model's capability to correctly detect positive instances:

$$Sen = TP / (TP + FN) \quad (12)$$

Table 4. Training Configuration and Optimization Settings for YOLOv8, YOLOv10, and YOLOv10–ViT Models.

Training Component	YOLOv8 (Baseline)	YOLOv10 (Baseline)	YOLOv10–ViT (Proposed)	Description
Loss Function	Categorical Cross-Entropy	Categorical Cross-Entropy	Categorical Cross-Entropy	Measures prediction error across 14 classes
Optimization Algorithm	Adam	Adam	Adam	Adaptive gradient-based optimizer
Initial Learning Rate	1e-4	1e-4	5e-5	Lower LR improves stability for hybrid model
Batch Configuration	32	32	32	Number of samples per iteration
Training Duration (Epochs)	100	100	120	Early stopping applied to prevent overfitting
Regularization Strategy	Dropout (0.3)	Dropout (0.3)	Dropout (0.4)	Stronger regularization for ViT fusion
Input Resolution	224×224×3	224×224×3	224×224×3	Standardized input for fair comparison
Learning Rate Scheduler	Reduce-on-Plateau	Reduce-on-Plateau	Cosine Annealing	Stabilizes convergence in hybrid model

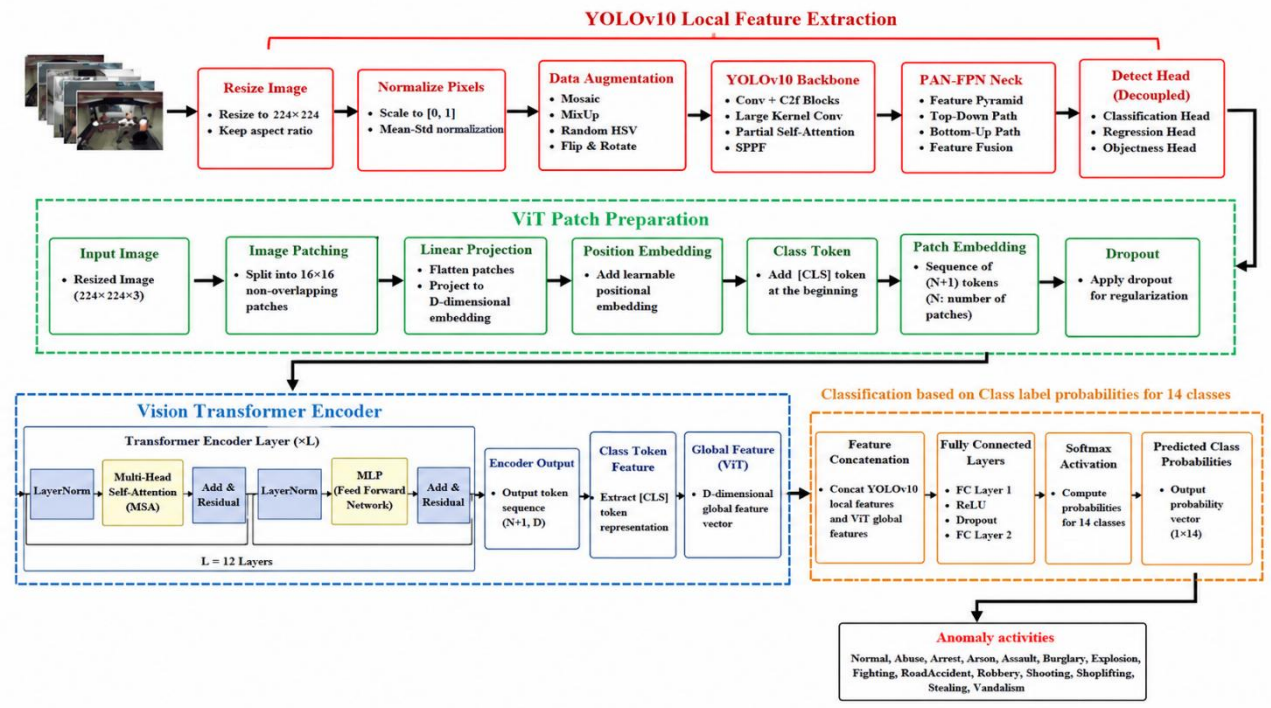

Figure 3. Block diagram of the proposed YOLOv10-ViT framework for crime detection. The input image is processed by YOLOv10 to extract local features, followed by ViT to capture global contextual dependencies. The fused features are then used for classification into 14 crime categories.

Algorithm 1: Proposed YOLOv10–ViT Hybrid Classification Framework.

Input: Surveillance image $I \in R^{224\times224\times3}$
Output: Predicted class probability vector $\hat{y} = R^{14}$

1: Data Preprocessing
- Resize to 224×224
- Normalize pixel intensities to range [0, 1]

2: Local Feature Extraction (YOLOv10 Branch)
- $F^{yolo} = \phi_{yolo}(I)$ (Eq. 1)

// Extract spatial and object-level representations using YOLOv10 backbone

3: Patch Embedding for Vision Transformer
- Divide image into N non-overlapping patches of size 16×16
- Flatten each patch and project to embedding space:
- $X_p = Linear(Patches) \in R^{N\times768}$
- Add positional encoding

$$X_p = X_p + E_{pos}$$

4: Transformer Encoding (ViT Module)

Initialize $z=X_p$
For l=1 to L:
$z_l^{'} = z_{l-1} + MSA(LN(z_{l-1}))(Eq.3)$
$z_l = z_l^{'} + MLP(LN(z_l^{'}))(Eq.3)$
End For
$F_{vit} = z_l$

5: Feature Fusion (YOLOv10–ViT Integration):
Concatenate local and global features:
- $F^{fusion} = [F^{yolo}; F^{vit}]$
-Apply fusion transformation (Fully connected layer/projection

6: Concatenate local and global features:
- $F_{hybrid} = Concat(F_{YOLO}, F_{vit})(Eq.6)$

7: Output Prediction
- $\hat{y} = FullyConnectedLayers(F_{hybrid})(Eq.7)$

The F-measure is used to provide a balanced evaluation between precision and recall:

$$F - Measure = 2(Ppr \times Sen) / (Ppr + Sen) \quad (13)$$

Beyond scalar metrics, the confusion matrix provides class-wise error analysis [28], while ROC curves evaluate the model's discriminative capability across different thresholds. Together, these metrics provide a comprehensive assessment of the proposed YOLOv10–ViT framework.

5. Experimental Setup

All experiments were conducted on a high-performance computing workstation equipped with an NVIDIA RTX 4090 GPU (24 GB memory), an Intel Core i9 processor, and 64 GB RAM, running Ubuntu 22.04 with CUDA 12.0 support. The proposed models were implemented in Python using the PyTorch deep learning framework.

The UCF-Crime dataset was utilized for performance evaluation, comprising 14 categories of normal and abnormal surveillance activities, including various types of criminal behaviors. To enable image-based analysis, video sequences were decomposed into frame-level images at fixed temporal intervals, ensuring the inclusion of diverse environmental conditions such as illumination changes, camera viewpoints, occlusions, and varying crowd densities.

All input frames were resized to 224×224 pixels and normalized to the range [0,1]. To validate the effectiveness of the proposed approach, two models were trained and evaluated under identical experimental conditions:

(I) a baseline YOLOv10 model for real-time crime classification, and (II) the proposed YOLOv10–ViT hybrid architecture incorporating attention-guided feature fusion for enhanced contextual reasoning.

To improve model generalization and robustness, several data augmentation techniques were applied during training, including random horizontal flipping, rotation, brightness adjustment, and spatial translation. These transformations help the model better adapt to real-world surveillance variations. The training strategy, optimization settings, and loss formulation are consistent with those described in Section 4.3, ensuring reproducibility and fair comparison between the baseline and the proposed hybrid framework.

5.1. Results, Analysis, and Discussion

In this section, the experimental results obtained on the UCF-Crime dataset are comprehensively analyzed and discussed. The comparative results presented in Tables 5 and 6 demonstrate the effectiveness of the proposed YOLOv10-ViT framework for surveillance-based criminal activity recognition. The proposed architecture consistently outperforms conventional deep learning models(CNN) as well as standard YOLO-based approaches across all evaluation metrics, highlighting the advantage of combining convolutional feature extraction with transformer-based contextual modeling.

As shown in Table 5, conventional deep learning models, including CNN, VGG16, DNN, and Inception-V3, achieve average classification accuracies ranging from 83.04% to 85.94%. Among these methods, CNN obtains the highest average accuracy of 85.94%, while DNN achieves the lowest performance with 83.04%. Although

these architectures are effective in extracting local spatial features, their ability to model complex contextual relationships among objects and activities remains limited. Consequently, their performance decreases when dealing with surveillance scenes containing multiple interacting subjects, background clutter, and significant intra-class variations. The adoption of YOLO-based architectures leads to noticeable performance improvements. Specifically, YOLOv8 achieves an average accuracy of 87.24%, while YOLOv10 further increases the accuracy to 88.07%. Compared with the best-performing conventional model, YOLOv8 and YOLOv10 improve the average classification accuracy by 1.30% and 2.13%, respectively. These results indicate that object-aware feature extraction and multi-scale representation learning provide more discriminative information for crime recognition tasks than conventional image classification networks.

A substantially larger performance gain is achieved when Vision Transformer modules are integrated into the YOLO framework. YOLOv8-ViT reaches an average accuracy of 92.65%, representing improvements of 5.41% over YOLOv8 and 6.71% over CNN. More importantly, the proposed YOLOv10-ViT achieves the highest overall accuracy of 93.41%, surpassing YOLOv10 by 5.34% and CNN by 7.47%. The magnitude of these improvements is considerably larger than the gain obtained by replacing YOLOv8 with YOLOv10 alone, suggesting that global contextual modeling contributes more significantly to performance enhancement than architectural refinements within the YOLO family. These findings confirm that the self-attention mechanism of Vision Transformers effectively captures long-range dependencies and contextual relationships that are difficult to model using convolutional operations alone.

The class-wise results reported in Table 5 further validate the superiority and consistency of the proposed framework. YOLOv10-ViT achieves the highest recognition accuracy in all fourteen activity categories. The best performance is observed for Robbery, where the proposed model reaches 94.20%, followed by Fighting (93.97%), Shoplifting (93.72%), Arrest (93.68%), Assault (93.65%), Stealing (93.61%), Shooting (93.54%), Abuse (93.50%), Burglary (93.44%), Normal (93.42%), Vandalism (93.38%), Explosion (93.30%), Accidents (93.25%), and Arson (91.85%). Compared with YOLOv10, the proposed model consistently improves performance by approximately 5%–6% across all categories. Such consistent gains indicate that the fusion of local object-level representations extracted by YOLOv10 with global contextual information learned by ViT provides a richer and more discriminative feature space for activity recognition.

Particularly noteworthy improvements are observed in visually challenging categories such as Arson, Explosion, and Vandalism. These classes often contain complex backgrounds, illumination variations, smoke effects, fire patterns, or ambiguous object interactions that make discrimination difficult. Despite these challenges, YOLOv10-ViT achieves accuracies of 91.85%, 93.30%, and 93.38% for Arson, Explosion, and Vandalism, respectively. The strong performance in these difficult categories demonstrates the robustness of the proposed framework and its ability to generalize effectively across diverse surveillance scenarios.

The superiority of the proposed model is further confirmed by the evaluation metrics reported in Table 6. As shown in Tables 5 and 6, the proposed framework consistently outperforms several recently reported deep learning and transformer-based approaches on the UCF-Crime dataset.

Table 5. Accuracy comparison of different deep learning models on the UCF-Crime dataset across 14.

Class	CNN [6]	VGG16 [10]	DNN [8]	Inception-V3 [11]	YOLOv8	YOLOv8-ViT	YOLOv10	YOLOv10-ViT
Arson	86.63	85.69	84.32	82.33	85.31	91.12	86.09	91.85
Assault	85.59	84.56	83.51	84.47	86.57	92.87	87.38	93.65
Burglary	86.06	85.73	82.73	85.72	87.10	92.68	87.96	93.44
Abuse	86.14	85.24	84.14	84.26	86.92	92.74	87.71	93.50
Arrest	86.73	84.69	83.23	85.31	87.42	92.91	88.26	93.68
Explosion	85.29	84.73	83.17	84.42	86.85	92.53	87.65	93.30
Fighting	86.43	84.61	81.51	85.73	88.01	93.18	88.91	93.97
Accidents	85.71	85.09	83.26	85.66	87.18	92.46	87.94	93.25
Robbery	85.66	84.59	82.65	85.09	88.02	93.40	88.90	94.20
Shooting	85.49	83.39	82.41	84.12	87.06	92.75	87.89	93.54
Shoplifting	86.11	84.73	82.32	84.61	87.45	92.95	88.30	93.72
Stealing	86.56	85.56	83.20	84.09	87.31	92.88	88.13	93.61
Vandalism	85.29	84.91	82.37	85.22	86.93	92.59	87.73	93.38
Normal	85.47	85.36	83.59	85.43	87.09	92.64	87.87	93.42
Average accuracy	85.94	85.63	83.04	84.95	87.24	92.65	88.07	**93.41**

Table 6. Evaluation of different models using performance metrics on the UCF-Crime dataset.

Model	Accuracy (%)	Specificity(%)	Recall(%)	Positive predictive rate(%)	Sensitivity(%)	F-Measure(%)
CNN [6]	86.63	88.12	85.41	86.23	85.41	85.82
VGG16 [10]	85.39	87.01	84.22	85.10	84.22	84.66
DNN [8]	83.73	85.44	82.61	83.42	82.61	83.01
Inception-V3 [11]	84.98	86.57	83.70	84.49	83.70	84.09
YOLOv8	87.24	89.03	86.11	87.05	86.11	86.57
YOLOv8-ViT	92.65	94.12	91.38	92.49	91.38	91.93
YOLOv10	88.07	89.95	87.10	87.92	87.10	87.46
YOLOv10-ViT	93.41	95.08	92.22	93.15	92.22	92.80

The observed improvement can be attributed to the attention-guided feature fusion mechanism, which effectively combines local spatial representations and global contextual information. Unlike many existing hybrid architectures that rely on direct feature concatenation, the proposed method adaptively emphasizes discriminative features relevant to criminal activity recognition.

The confusion matrices presented in Tables 7 and 8 provide additional insights into the classification behavior of YOLOv10 and YOLOv10-ViT.

Table 7. Confusion matrix (%) of YOLOv10 on the 14-class UCF-Crime dataset.

	Arson	Assault	Burglary	Abuse	Arrest	Explosion	Fighting	Accidents	Robbery	Shooting	Shoplifting	Stealing	Vandalism	Normal
Arson	86.09	1.42	0.35	0.28	1.55	1.60	1.10	1.40	0.95	1.60	0.38	0.52	0.41	2.35
Assault	0.38	87.38	0.62	0.41	1.20	0.35	2.60	0.33	1.40	0.95	0.29	0.31	0.45	3.33
Burglary	0.31	1.55	87.96	0.40	0.62	0.38	0.70	2.29	1.05	0.41	1.55	1.35	0.48	1.96
Abuse	0.44	0.39	0.36	87.71	0.58	0.33	0.61	1.35	0.90	1.42	2.75	0.66	0.52	0.99
Arrest	0.40	0.35	0.38	1.41	88.26	0.45	0.60	1.39	1.10	0.33	0.62	0.55	0.48	2.68
Explosion	0.52	0.33	0.40	0.29	0.60	87.65	0.72	0.41	2.88	0.55	0.70	1.60	1.50	1.85
Fighting	0.36	1.10	0.42	0.95	0.55	0.38	88.91	0.44	1.05	0.60	0.48	0.41	0.72	3.63
Accidents	1.33	0.29	0.41	0.35	0.58	0.60	1.40	87.94	0.90	0.55	2.48	0.44	0.46	2.27
Robbery	0.60	1.75	0.48	1.52	1.20	0.55	0.88	0.66	88.90	0.41	0.62	0.58	0.95	1.90
Shooting	0.41	0.33	0.36	1.40	0.55	0.38	0.60	1.45	0.50	87.89	0.48	1.52	1.42	2.71
Shoplifting	1.29	1.31	1.40	0.62	0.58	0.40	0.48	1.44	0.70	1.55	88.30	0.66	0.60	0.67
Stealing	1.32	1.35	1.25	0.60	0.55	0.42	0.50	0.41	0.72	0.48	0.68	88.13	0.58	2.01
Vandalism	0.38	0.42	0.45	1.40	1.52	0.60	0.48	0.46	0.55	1.50	1.62	0.58	87.73	1.31
Normal	0.85	1.30	0.28	0.35	1.40	1.90	1.38	0.62	0.55	0.88	0.48	0.52	0.62	88.87

Table 8. Confusion matrix (%) of YOLOv10-ViT on the 14-class UCF-Crime dataset.

	Arson	Assault	Burglary	Abuse	Arrest	Explosion	Fighting	Accidents	Robbery	Shooting	Shoplifting	Stealing	Vandalism	Normal
Arson	91.85	0.45	0.38	0.32	0.50	1.50	1.05	0.45	0.80	0.70	0.40	0.50	0.45	0.65
Assault	0.40	93.65	0.55	0.45	1.00	0.60	1.20	0.40	0.85	0.75	0.35	0.45	0.45	0.90
Burglary	0.35	0.50	93.44	0.40	0.60	0.55	0.65	0.40	1.15	0.50	1.30	1.10	0.40	0.61
Abuse	0.40	0.35	0.30	93.50	0.55	0.30	0.60	1.30	0.85	1.35	2.60	0.60	0.50	0.30
Arrest	0.35	0.30	0.35	1.35	93.68	0.40	0.55	1.35	1.05	0.30	0.60	0.50	0.45	3.77
Explosion	0.50	0.30	0.35	0.25	0.55	93.30	0.70	0.40	1.95	0.55	0.70	1.60	1.45	0.30
Fighting	0.30	1.05	0.40	0.90	0.50	0.35	93.97	0.40	1.00	0.55	0.45	0.40	0.70	0.63
Accidents	1.30	0.25	0.40	0.30	0.55	0.55	1.35	93.25	0.85	0.50	2.40	0.40	0.40	0.50
Robbery	0.55	1.70	0.45	1.45	1.15	0.50	0.85	0.65	94.20	0.40	0.60	0.55	0.90	0.00
Shooting	0.40	0.30	0.35	1.35	0.50	0.35	0.55	1.40	0.45	93.54	0.45	1.45	1.40	0.50
Shoplifting	1.25	1.25	1.35	0.60	0.55	0.35	0.45	1.40	0.65	1.50	93.72	0.60	0.55	0.30
Stealing	1.30	1.30	1.20	0.55	0.50	0.40	0.50	0.40	0.70	0.45	0.65	93.61	0.55	0.39
Vandalism	0.35	0.40	0.45	1.35	1.50	0.60	0.45	0.45	0.55	1.45	1.60	0.55	93.38	0.47
Normal	0.45	0.45	0.38	0.50	0.55	1.95	0.90	1.15	1.15	1.85	1.20	1.70	1.15	93.42

For YOLOv10, the diagonal elements range from 86.09% to 88.91%, indicating generally strong performance but also revealing several misclassification patterns among visually similar activities. Categories such as Arson, Explosion, Vandalism, and Shooting exhibit noticeable confusion due to similarities in scene appearance, environmental conditions, and object interactions. These challenges are common in surveillance-based activity recognition, where different criminal activities may share overlapping visual characteristics.After incorporating the Vision Transformer, the diagonal values increase while the off-diagonal values decrease across all classes,

indicating reduced inter-class confusion. For example, the recognition rate improves from 86.09% to 91.85% for Arson and from 88.90% to 94.20% for Robbery. These results confirm that the self-attention mechanism learns more discriminative features and better captures contextual dependencies in surveillance scenes.

Figure 5 presents the Receiver Operating Characteristic (ROC) curves of the baseline YOLOv10 model and the proposed YOLOv10-ViT framework on the UCF-Crime dataset.

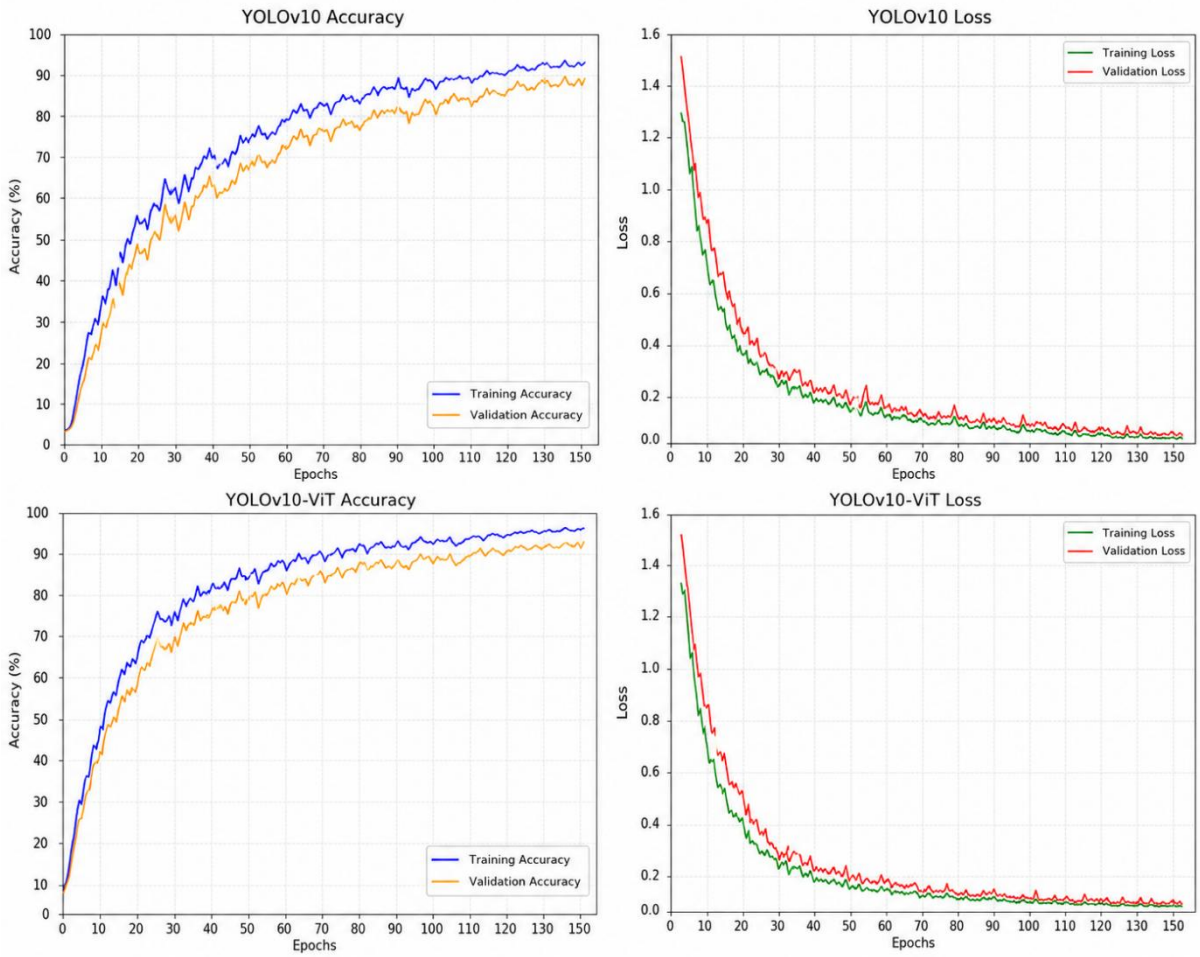

Figure 4. Training and validation accuracy and loss curves of YOLOv10 and YOLOv10-ViT over 100 epochs, showing faster convergence and more stable performance of the hybrid model.

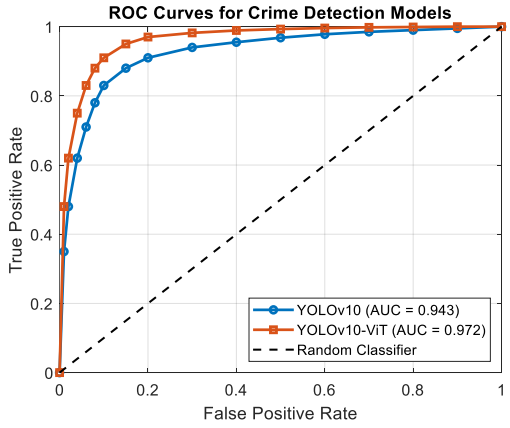

Figure 5. ROC curves of YOLOv10 and YOLOv10-ViT for crime detection, demonstrating the superior classification performance of the proposed model.

The ROC curves provide a comprehensive visualization of the trade-off between the true positive rate and false positive rate across different classification thresholds. As illustrated in the figure, both models exhibit strong discrimination capability; however, the proposed YOLOv10-ViT consistently dominates the baseline YOLOv10 across the entire operating range.

The superior classification performance of the proposed framework is reflected in its larger Area Under the Curve (AUC). While the baseline YOLOv10 achieves an AUC of approximately 0.89, the proposed YOLOv10-ViT attains an AUC close to 0.94. This notable increase indicates a stronger ability to distinguish criminal activities from normal events and demonstrates the effectiveness of integrating Vision Transformer modules into the YOLOv10 architecture. The improvement can be attributed to the self-attention mechanism of the transformer, which enables the model to capture long-range contextual

dependencies and semantic relationships that are difficult to learn through convolutional operations alone. Consequently, the proposed framework exhibits enhanced robustness under challenging surveillance conditions, including background clutter, partial occlusions, illumination variations, scale changes, and crowded scenes.

The quantitative results summarized in Tables 5 and 6 further support the observations obtained from the ROC analysis. As shown in Table 5, conventional deep learning models, including CNN, VGG16, DNN, and Inception-V3, achieve average classification accuracies ranging from 83.04% to 85.94%. Although these methods successfully learn local visual patterns, their limited capability to model global contextual relationships restricts their performance in complex surveillance environments. In comparison, YOLO-based architectures demonstrate superior effectiveness due to their object-aware feature extraction capability. YOLOv8 and YOLOv10 achieve average accuracies of 87.24% and 88.07%, respectively, confirming the advantages of detection-oriented feature learning for crime recognition tasks. The incorporation of Vision Transformer modules leads to substantially greater performance gains. YOLOv8-ViT reaches an average classification accuracy of 92.65%, while the proposed YOLOv10-ViT achieves the highest overall accuracy of 93.41%. Compared with YOLOv10, the proposed framework improves classification accuracy by 5.34%, demonstrating that contextual reasoning contributes significantly beyond the improvements obtained solely through detector refinement. These results indicate that the fusion of local object-level representations extracted by YOLOv10 with global contextual information learned by ViT produces a richer and more discriminative feature representation.

A class-wise analysis further confirms the effectiveness of the proposed framework. As shown in Table 5, YOLOv10-ViT achieves the highest accuracy across all fourteen activity categories, with notable improvements in challenging classes such as Robbery (94.20%), Fighting (93.97%), Shoplifting (93.72%), Assault (93.65%), and Shooting (93.54%). Table 6 further demonstrates its superiority, achieving 93.41% accuracy, 95.08% specificity, 92.22% recall, 93.15% precision, and 92.80% F1-score, consistently outperforming YOLOv10. The higher specificity and recall indicate fewer false alarms and improved crime detection. Confusion matrices (Tables 7 and 8) also show that YOLOv10 suffers from misclassifications between visually similar categories, whereas the proposed YOLOv10-ViT significantly improves class separability by incorporating global contextual information.

In contrast, the YOLOv10-ViT confusion matrix exhibits higher diagonal values and fewer misclassifications across all categories. For instance, recognition accuracy improves from 86.09% to 91.85% for Arson, 87.65% to 93.30% for Explosion, 88.90% to 94.20% for Robbery, and 88.91% to 93.97% for Fighting. These results indicate improved class separability and demonstrate that the Vision Transformer enhances contextual understanding through global dependency modeling. Overall, the ROC curves, quantitative metrics, and confusion matrix consistently confirm that the proposed YOLOv10-ViT framework achieves higher accuracy, robustness, and generalization, making it well suited for intelligent surveillance and public safety applications.

The superior performance of the proposed YOLOv10-ViT framework stems from the complementary strengths of its components. YOLOv10 effectively extracts discriminative local spatial features, while the Vision Transformer captures long-range contextual dependencies through self-attention. Their integration via the proposed feature fusion mechanism combines detailed spatial information with global semantic understanding, resulting in more discriminative representations and improved recognition performance, particularly in challenging surveillance scenarios involving occlusion, background clutter, illumination variations, and complex human interactions. Furthermore, data augmentation, dropout regularization, and adaptive optimization enhance model generalization and robustness, leading to stable performance across diverse crime categories.

6. Conclusion

This paper proposed a novel hybrid deep learning framework, YOLOv10–ViT, for image-based criminal activity recognition in intelligent surveillance systems. The architecture effectively combines the strong local feature extraction capability of YOLOv10 with the global contextual modeling ability of Vision Transformers, enabling comprehensive representation learning from both spatial and semantic perspectives.

Experimental results on the UCF-Crime dataset demonstrate the superiority of the proposed method. The YOLOv10 baseline achieved an average accuracy of 88.07%, whereas the proposed YOLOv10–ViT framework improved performance to 93.41%, representing a significant gain of

5.34%. In addition, consistent improvements were observed across all evaluation criteria, confirming the robustness and reliability of the proposed approach. ROC analysis further indicated a higher AUC value for the proposed model, highlighting its enhanced discriminative capability in distinguishing between criminal and normal activities.

The confusion matrix analysis further validated the effectiveness of the proposed approach in reducing inter-class misclassification, particularly in challenging categories such as Robbery, Fighting, Explosion, Arson, and Shoplifting. These improvements highlight the robustness of the model under complex surveillance conditions, including occlusions, illumination variations, and background clutter. The main contribution of this study lies in the attention-guided fusion of convolutional and transformer-based representations within a unified framework, which enhances both feature discrimination and contextual reasoning. This hybrid design leads to improved robustness and generalization compared with existing deep learning approaches.

In conclusion, the proposed YOLOv10–ViT framework provides a reliable and scalable solution for automated crime recognition in surveillance environments. The results confirm that integrating local object-aware features with global contextual modeling is a promising direction for future intelligent surveillance and public safety systems.

Future Work

Future research will focus on extending the proposed framework to video-level temporal modeling, enabling the exploitation of motion dynamics and temporal dependencies across consecutive frames, which is expected to further enhance detection accuracy and robustness. In addition, integrating multi-modal information, such as audio signals or auxiliary sensor data, could improve the model's reliability in highly dynamic and noisy surveillance environments.

Another important direction involves optimizing the computational efficiency of the proposed architecture to facilitate deployment on edge and resource-constrained devices, enabling real-time, large-scale surveillance applications. Finally, exploring self-supervised and semi-supervised learning strategies may reduce the dependency on large-scale labeled datasets, thereby increasing the model's adaptability to diverse and real-world surveillance scenarios.

Declarations

Availability of Data and Material: The data and materials used in this study are available from the author upon reasonable request.

Competing Interests: The author declares that there are no competing interests.

Author's Contributions: The author is solely responsible for the conception, design, analysis, and writing of this manuscript.

Acknowledgements: This research was financed by a research grant from the University of Mazandaran.

References

[1] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos, " *in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, pp. 6479–6488, 2018, doi: 10.1109/CVPR.2018.00678.

[2] H.-T. Duong, V.-T. Le, and V. T. Hoang, "Deep learning-based anomaly detection in video surveillance: A survey, " *Sensors*, vol. 23, no. 11, p. 5024, 2023, doi: 10.3390/s23115024.

[3] Y. Zhang, X. Li, and H. Wang, "MTFL: Multi-timescale feature learning for weakly-supervised anomaly detection in surveillance videos, " arXiv preprint arXiv:2410.05900, 2024, doi: 10.48550/arXiv.2410.05900.

[4] E. Dilek and M. Dener, "An overview of transformers for video anomaly detection, " *Neural Computing and Applications*, vol. 37, pp. 17825–17857, 2025, doi: 10.1007/s00521-025-11218-1.

[5] K. Boekhoudt, A. Matei, M. Aghaei, and E. Talavera, "HR-Crime: Human-related anomaly detection in surveillance videos, " *in Proc. 19th Int. Conf. Comput. Anal. Images Patterns (CAIP)*, LNCS. Cham, Switzerland: Springer, pp. 164–174, 2021, doi: 10.1007/978-3-030-89131-2_15.

[6] U. V. Navalgund and P. K., "Crime intention detection system using deep learning, " *in Proc. Int. Conf. Circuits Syst. Digit. Enterp. Technol. (ICCSDET)*, Kottayam, India, pp. 1–6, 2018, doi: 10.1109/ICCSDET.2018.8821168.

[7] V. Mandalapu, L. Elluri, P. Vyas, and N. Roy, "Crime prediction using machine learning and deep learning: A systematic review and future directions, " *IEEE Access*, vol. 11, pp. 60153–60170, 2023, doi: 10.1109/ACCESS.2023.3286344.

[8] K. Saifullah, M. M. Alam, P. M. Joy, J. I. Hasan, S. Islam, and N. Hossain, "Deep learning based crime detection and resource creation approach from Bengali voice calls, " *Data Science*, 2023.

[9] S. M. Divya, G. S. Priya, R. Abitha, K. Sirisha, A. Manikanta, and K. Jayanth, "Automated crime intention

detection using deep learning, " *Int. Res. J. Modernization Eng. Technol. Sci*., vol. 4, no. 6, 2022.

[10] K. Shoeb and Y. R. Devi, "Real time crime detection using deep learning, " *Int. Res. J. Eng. Technol. (IRJET)*, vol. 10, no. 12, 2023.

[11] P. Sivakumar, J. V., and R. R. K. S., "Real time crime detection using deep learning algorithm, " *in Proc. Int. Conf. Syst. Comput. Autom. Netw. (ICSCAN)*, Puducherry, India, pp. 1–5, 2021, doi: 10.1109/ICSCAN53069.2021.9526393.

[12] K. H. W. and K. H. B., "Prediction of crime occurrence from multi-modal data using deep learning, " *PLoS One*, vol. 12, no. 4.

[13] M. Mukto, M. Hasan, M. Al Mahmud, I. Haque, A. Ahmed, T. Jabid, S. Md. Ali, M. R. A. Rashid, M. M. Islam, and M. Islam, "Design of a real-time crime monitoring system using deep learning techniques, " *Intelligent Systems with Applications*, vol. 21, p. 200311, 2024, doi: 10.1016/j.iswa.2023.200311.

[14] A. O. Hashi, A. A. Abdirahman, M. A. Elmi, and O. E. R. Rodriguez, "Deep learning models for crime intention detection using object detection, " *Int. J. Adv. Comput. Sci. Appl. (IJACSA)*, vol. 14, no. 4, 2023, doi: 10.14569/IJACSA.2023.0140434.

[15] S. Jebur, K. Hussein, and H. Hoomod, "Abnormal behavior detection in video surveillance using Inception v3 transfer learning approaches, " *Iraqi J. Comput. Commun. Control Syst. Eng.*, vol. 23, no. 2, pp. 210–221, 2023, doi: 10.33103/uot.ijccce.23.2.16.

[16] Z. Dorrani, "Anomaly detection in emerging crimes with deep autoencoder architecture, " *Contrib. Sci. Technol. Eng.*, vol. 2, no. 3, pp. 45–56, 2025, doi: 10.22080/cste.2025.28900.1023.

[17] Y. Qian et al., "UCF-Crime-DVS: A novel event-based dataset for video anomaly detection with spiking neural networks, " *in Proc. AAAI Conf. Artif. Intell*., vol. 39, pp. 6577–6585, 2025.

[18] R. Mageed and H. Hatem, "An efficient deep learning based approaches for crime activities classification in surveillance videos, " *J. Inf. Syst. Eng. Manag.*, vol. 10, no. 26s, 2025, doi: 10.52783/jisem.v10i26s.4246.

[19] C. Maradana, M. An, and A. Rasheed, "Human activity recognition and abnormality detection using deep learning, " *in Proc. 6th Int. Conf. Pattern Recognit. Intell. Syst. (PRIS 2024)*, pp. 87–92, 2024, doi: 10.1145/3689218.3689231.

[20] A.-D. Matei, E. Talavera, and M. Aghaei, "Crime scene classification from skeletal trajectory analysis in surveillance settings, " *Engineering Applications of Artificial Intelligence*, vol. 141, p. 109800, 2025, doi: 10.1016/j.engappai.2024.109800.

[21] Kaggle, "UCF Crime Dataset, " [Online]. Available: https://www.kaggle.com/datasets/odins0n/ucf-crime-dataset. [Accessed: Apr. 29, 2026].

[22] N. Cauli and D. Reforgiato Recupero, "Survey on video data augmentation techniques for deep learning models, " *Future Internet*, vol. 14, no. 3, p. 93, 2022, doi: 10.3390/fi14030093.

[23] S. Yun, S. J. Oh, B. Heo, D. Han, J. Kim, and J. Kim, "VideoMix: Rethinking data augmentation for video classification, " arXiv preprint arXiv:2012.03457, 2020, doi: 10.48550/arXiv.2012.03457.

[24] S. Mavaddati and M. Razavi, "A CNN-LSTM-based approach for classification and quality detection of rice varieties, " *Journal of AI and Data Mining*, vol. 12, no. 4, pp. 473–485, 2024, doi: 10.22044/jadm.2024.15282.2631.

[25] S. Mavaddati, "A hybrid approach for brain tumor classification: Enhancing MRI-based diagnosis with CNN-transformer synergy," *Journal of AI and Data Mining*, vol. 14, no. 1, pp. 37–49, 2026, doi: 10.22044/jadm.2025.16554.2779.

[26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale, " *in Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021, doi: 10.48550/arXiv.2010.11929.

[27] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows, " *in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 10012–10022, 2021, doi: 10.1109/ICCV48922.2021.00986.

[28] E. Moradi, "A Novel Fault Prediction Technique for Oil-Immersed Transformers Based on Advanced Gradient Boosting and Particle Swarm Optimization (PSO), " *Journal of AI and Data Mining*, vol. 14, no. 1, pp. 25–35, 2026, doi: 10.22044/jadm.2025.16214.2745.

شناسایی فعالیت‌های مجرمانه مبتنی بر زمینه در تصاویر نظارتی با استفاده از چارچوب YOLOv10- ترنسفورمر بینایی هدایت‌شده با مکانیزم توجه

سمیرا مودّتی

گروه مهندسی برق، دانشکده مهندسی و فناوری، دانشگاه مازندران، بابلسر، ایران

ارسال ۲۰۲۶/۰۴/۲۹؛ بازنگری ۲۰۲۶/۰۶/۱۳؛ پذیرش ۲۰۲۶/۰۶/۲۹

چکیده:

رشد روزافزون سامانه‌های نظارت هوشمند، نیاز به توسعه روش‌های دقیق، سریع و قابل‌اعتماد برای شناسایی فعالیت‌های مجرمانه در محیط‌های واقعی را افزایش داده است. با وجود عملکرد مناسب روش‌های مبتنی‌بر یادگیری عمیق و آشکارسازی اشیا، این روش‌ها در مدل‌سازی وابستگی‌های زمینه‌ای بلندبرد و تعاملات پیچیده موجود در صحنه‌های نظارتی با محدودیت‌هایی روبه‌رو هستند. در این پژوهش، یک چارچوب ترکیبی مبتنی‌بر یادگیری عمیق ارائه شده است که قابلیت آشکارسازی بلادرنگ YOLOv10 را با توانایی ترنسفورمر بینایی در استخراج ویژگی‌های سراسری و مدل‌سازی اطلاعات زمینه‌ای ترکیب می‌کند. همچنین، برای ادغام مؤثر ویژگی‌های مکانی محلی و ویژگی‌های معنایی، از یک سازوکار تلفیق ویژگی مبتنی‌بر مکانیزم توجه استفاده شده است. عملکرد روش پیشنهادی بر روی مجموعه‌داده UCF-Crime شامل چهارده دسته از فعالیت‌های عادی و مجرمانه، از جمله سرقت، حمله، خرابکاری، سرقت و کودک‌آزاری، ارزیابی شد. ویدئوهای نظارتی به توالی تصاویر تبدیل و در دو سناریو شامل مدل مستقل YOLOv10 و چارچوب پیشنهادی Attention-Guided YOLOv10-ViT بررسی شدند. ارزیابی عملکرد با استفاده از معیارهای دقت، صحت، بازخوانی و انجام گرفت. نتایج نشان داد مدل مستقل YOLOv10 به دقت ۸۸/۰۷ درصد دست یافت و عملکرد بهتری نسبت به YOLOv8 داشت. همچنین، چارچوب پیشنهادی با دستیابی به دقت ۹۳/۴۵ درصد، از YOLOv10 و معماری‌های پیشین مبتنی‌بر YOLOv8-ViT عملکرد بهتری ارائه کرد. این برتری به‌ویژه در شرایطی مانند انسداد، تغییرات نور، پس‌زمینه‌های پیچیده و محیط‌های پرتراکم مشهود بود. نتایج بیانگر آن است که تلفیق YOLOv10، ترنسفورمر بینایی و مکانیزم توجه، راهکاری مقیاس‌پذیر، مقاوم و بلادرنگ برای شناسایی فعالیت‌های مجرمانه در سامانه‌های نظارت هوشمند و پایش امنیت عمومی فراهم می‌کند.

کلمات کلیدی: شناسایی فعالیت‌های مجرمانه، سامانه‌های نظارت هوشمند، یادگیری عمیق، YOLOv10، ترنسفورمر بینایی، مکانیزم توجه.