Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

A Transformer-Enhanced Human-Centric Unsupervised Framework for Multi-Person Video Anomaly Detection

Hiba Mohsin Abdulameer [*1] ***Ali Mohsin Al-juboori*** [1]

[1] *College of Computer Science and Information Technology,University of Al-Qadisiyah, Al-Qadisiyah, Iraq*

[*] *Corresponding author's Email: master.student2414@qu.edu.iq*

ARTICLE INFO

Article history:
Received: 2 /03/2026
Rrevised form: 11/04/2026
Accepted : 12 /04/2026
Available online: 30 /06/2026

Keywords:

Anomaly detection,

Skeleton-based analysis,

HuVAD,

Video surveillance

ABSTRACT

Video anomaly detection in surveillance environments is still difficult. This is because abnormal events do not happen often, take different forms, and depend on the complex nature of real scenes. In addition, methods that depend on visual appearance are affected by changes in lighting, camera angles, and background conditions. These issues can reduce detection accuracy and also cause privacy problems. For this reason, recent studies focus more on motion-based representations that describe human behavior and reduce the effect of unnecessary visual details.

In this work, a framework for video anomaly detection is proposed. Spatial motion features are extracted using a 2D convolutional neural network. These features are then passed to a GRU network to model motion over time. A Transformer module is also used to help capture longer temporal relationships in motion sequences.

The proposed framework is able to handle scenes that include more than one person. During training and testing, all detected persons are considered within fixed-length temporal windows. Information from each person is then combined to produce an anomaly score that represents the overall scene behavior. This helps the model detect abnormal activities even in crowded surveillance scenes.

The proposed model uses an unsupervised one class learning approach. Training is performed using normal motion data only, Abnormal events are identified by observing deviations from the learned patterns of normal behavior. The experiments were carried out on real surveillance datasets using standard evaluation metrics. The model achieved an AUC-ROC of 93% at the frame level, indicating stable and consistent performance across different cases. The integration of spatial and temporal features contributed to a more accurate representation of complex motion patterns and reduced the likelihood of confusing abnormal behavior with normal activity.

https://doi.org/10.29304/jqcsm.2026.18.22700

1.Introduction

Video surveillance systems play a vital role in enhancing public safety in modern environments such as airports, shopping malls, campuses, and smart cities [1]. With the rapid increase in the number of deployed cameras, continuous manual monitoring has become impractical, inefficient, and highly dependent on human attention. This limitation has driven significant research into automated video anomaly detection, which aims to identify abnormal or suspicious events that deviate from regular patterns of human behavior.

Human-centric anomalies in surveillance is for environments are influenced by multiple factors beyond motion patterns alone [2], [3]. Figure 1 illustrates the main elements involved in describing anomalous human behavior from a human-centered perspective.

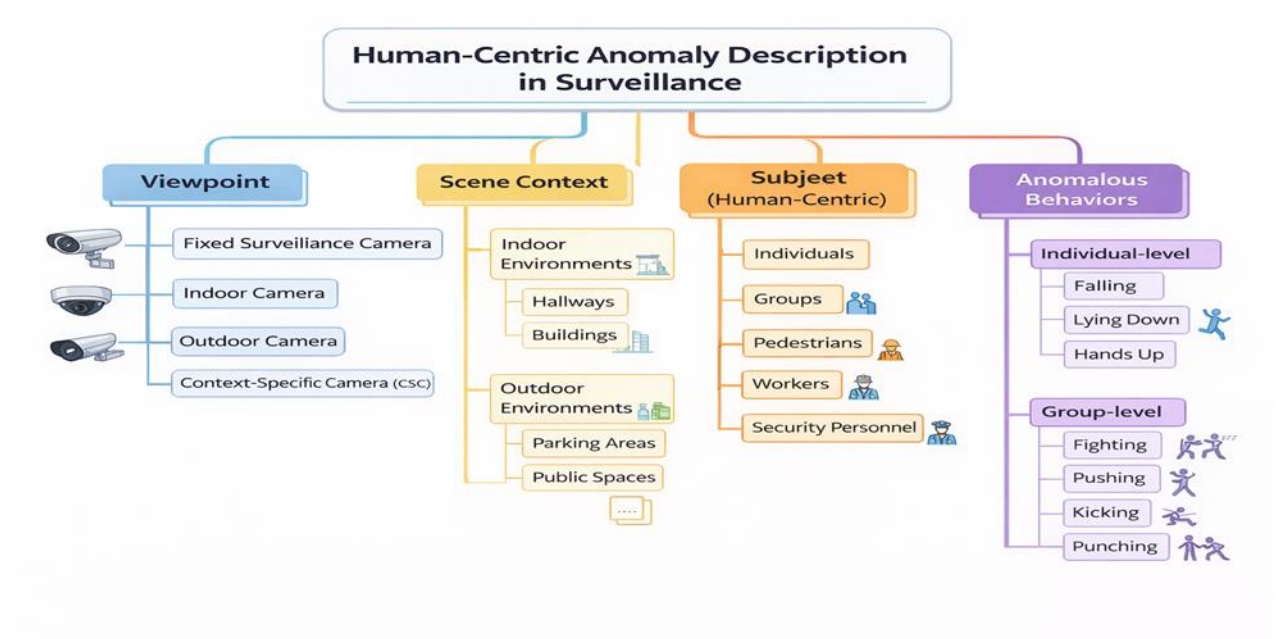


Figure 1. Taxonomy of human-centric anomaly description elements in surveillance

Despite the large amount of work in this area, video anomaly detection is still considered a difficult problem. This is because abnormal events do not happen often, take different forms ،and depend heavily on the surrounding situation [4]. For this reason ،it is not easy to define clear anomaly classes in advance. In real surveillance scenes ، the task becomes more complex due to background clutter ،changes in lighting ،different camera angles ،and the presence of several people at the same time. All these factors make human behavior harder to model and reduce the accuracy of many detection methods [5], [6].

In earlier studies ،anomaly detection was mainly based on handcrafted features and appearance information taken directly from video frames. Although these methods worked reasonably well in simple or controlled settings ،they were strongly affected by changes in the scene background. In addition ،appearance-based features often capture scene-specific details ،which limits their ability to work well in new environments and also raises privacy concerns. Because of these issues ،recent research has moved toward approaches that focus on motion and behavior rather than visual appearance [7], [8].

Using deep learning, better results have been achieved in learning spatial and temporal information from video data. Convolutional neural networks are commonly used to extract spatial features ،while recurrent models such as LSTM and GRU are used to represent motion over time. However, recurrent models often face difficulties when dealing with long time sequences, especially in complex surveillance scenes where abnormal events may occur at different time scales.

In recent years, Transformers have been applied to reduce these limitations. Their attention mechanism helps the model focus on important time steps and understand both short and long temporal patterns.[9], [10]. This is useful for detecting abnormal events that involve weak or irregular motion patterns.[11]

Handling more than one person in the same scene is not easy. When the model looks at one person only, it may miss what others are doing. In crowded scenes, any person may show abnormal behavior, so all people should be used.

Abnormal data is often limited and not clear.[12] For this reason, the model is trained on normal data only, and abnormal behavior is detected as a change.

Some limits still exist. Some methods use only spatial or only temporal modeling. Models like GRU can follow motion, but they struggle with long sequences. A transformer can handle long time, but may miss small motion details.[13]

Many studies still use one person only or a small number of people. Real scenes have many people. Some methods ignore motion features. Others need labeled data, so they are hard to use.[14]

Here, normal motion is learned using a one-class unsupervised setup.[15] Simple spatial features and temporal information are used, and multiple people are considered during training and testing. This reduces false detections in crowded scenes and addresses limits seen in earlier HuVAD based work.[16].

The main contributions of this work are :-

- The proposed model combines a convolutional neural network, a Transformer encoder, and a bidirectional GRU to capture complementary spatial and temporal characteristics of human motion.
- More than one person was used within each time window, and this helps to work better in crowded scenes.
- A motion-based representation is utilized by incorporating joint positions, velocity, and acceleration, which helps the model notice small changes in behavior.
- The framework is designed in a one-class unsupervised setting, where only normal data are used during training, making it suitable for real-world applications with limited labeled anomalies.
- Extensive experiments on the HuVAD dataset demonstrate that the proposed approach achieves stable and competitive performance across multiple camera views

The proposed framework differs from existing approaches by jointly integrating multi-person modeling, motion-based representation, and hybrid temporal learning within a unified one-class learning setting

NOMENCLATURE

T	temporal window length
K	number of persons per window (Top-K)
J	number of skeleton joints
P	pose sequence
X	joint coordinates

2. Related Work

HuVAD [17]is a The dataset focuses on human motion represented through skeletal information. It includes long-duration recordings captured from multiple camera views rather than short video clips. for analyzing surveillance scenarios in which individuals move continuously over time,This structure makes it well suited ، both in indoor and outdoor environments, and where more than one person may appear [18], [19].
These features allow abnormal behavior to be studied without depending on visual appearance. By using skeleton-based motion data, the focus stays on movement patterns instead of sensitive visual details. The continuous recordings also help in analyzing crowded scenes, where people interact frequently and normal and abnormal actions may happen at the same time [20].

Several previous studies have tested pose-based anomaly detection methods on the HuVAD dataset. Earlier approaches, such as MPED-RNN, focused on modeling human motion patterns using recurrent networks to distinguish between normal and abnormal behavior.[21], used recurrent neural networks to model motion over time. These methods showed that skeletal information can represent human movement effectively, but their performance decreases when scenes become crowded or when interactions between individuals are involved [22].

Later studies used graph based models, such as GEPC, to represent the spatial relationships between body joints. These approaches improved joint modeling, but they often assume relatively regular motion patterns. When movements become more complex, their performance may decrease. More recent methods, such as STG-NF, have tried to address these limitations.[23] and TSGAD[24], examined density and reconstruction methods; however, crowded environments and long-term motion continue to pose challenges.[20], [25].

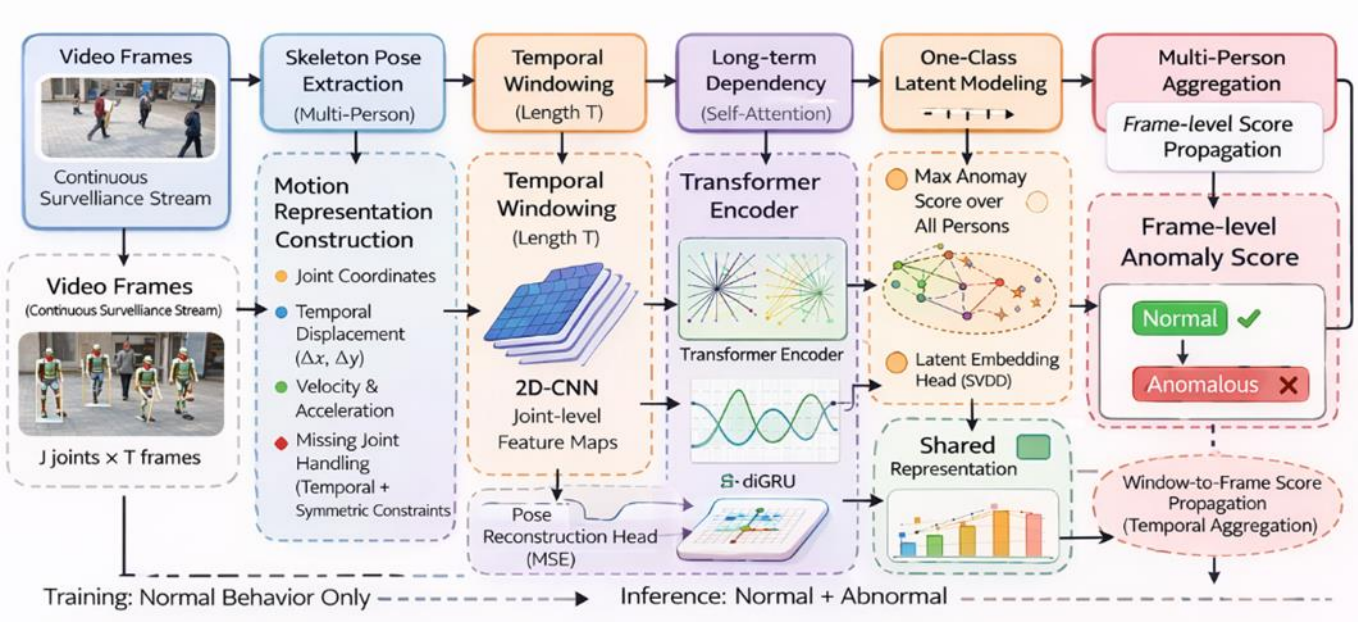
3.Methodology

3.1. HuVAD Dataset Description

The HuVAD dataset is used in this work. It relies on human motion data and includes long recordings from seven camera views. Six cameras correspond to standard surveillance scenes, whereas the seventh camera captures context specific security situations. These camera views reflect different real surveillance settings. Examples of the camera views used in this work are illustrated in Figure 1[1] .The camera views excluding people. The ratio has been adjusted to fit the manuscript

1. Figure 1 provides example frames captured by the CSC camera, which represents a specialized surveillance context

In contrast to many existing benchmarks composed of short, isolated video segments, HuVAD provides long-duration recordings that more closely resemble practical surveillance conditions. The scenes cover both indoor and outdoor settings and exhibit variations in crowd density, occlusion, and interaction patterns. The annotated abnormal events range from individual behaviors, such as falling, lying down, and raising hands, to interactions involving multiple individuals, including pushing, punching, kicking, and strangling.
To preserve privacy, HuVAD provides only anonymized human-related information, including bounding boxes, tracking IDs, and skeletal pose key points, without sharing raw video frames. This makes the dataset suitable for studies that is focus on skeletal motion and behavior analysis rather than visual appearance. Figure 2 shows the general structure of the proposed framework and illustrates The primary stages involved in transforming raw video input into frame-level anomaly detection results.



2. Figure 2 Architecture of the proposed unsupervised one-class multi-person video anomaly detection framework based on skeleton motion, Transformer attention, and GRU temporal modeling

3.2. Annotation Structure and Data Organization

Each video in the HuVAD dataset comes with an annotation file.
This file contains frame-level information for all detected people.
For each frame, it includes the person ID, bounding box, and 2D joint positions with confidence values.
Each frame also has a label.
Normal frames are labeled 0, and abnormal frames are labeled 1. In this work, each camera is treated separately.
Each scene stays as it is. The data is divided using the standard HuVAD setup.
Only normal data is used for training. Abnormal data is used for evaluation only.

3.3. Data Preprocessing and Pose Normalization

Skeletal data from surveillance videos may contain errors, such as missing joints or wrong positions due to occlusion or fast movement. For this reason, simple steps are applied before training.

The coordinates are centered around one joint to reduce the effect of the camera. Normalization is also applied within each time window to reduce differences in size and distance.

When joints are missing, they are estimated from nearby frames. Body symmetry can also be used in some cases. Segments with little or no movement are removed so the model focuses on useful motion.

Simple motion values are also added. The difference between two frames gives velocity, and the second difference gives acceleration. These are combined with the coordinates so the model focuses more on motion than shape.

3.4. Temporal Windowing and Multi-Person Representation

Skeleton sequences are divided into fixed windows of length T. Each person is trained separately. In testing, the highest anomaly score among people in the same window is used as the final result, which helps in crowded environments

3.5. Network Architecture

The model uses spatial and temporal information, where a 2D CNN is used to extract spatial features. Then Transformer and Bi-GRU. After that, global pooling gives a fixed output.

The network has two parts. One for reconstruction. One for feature embedding one-class learning.
The structure of each part is shown in detail to make the model clear and easy to follow.in **Table 1**.

TABLE 1-DETAILED ARCHITECTURE OF THE PROPOSED MODEL

No.	Layer	Type	Output Shape	Param	Activation	Notes
1	Conv2d	Convolution	(64, T, J)	3,520	GELU	kernel=3, stride=1, padding=1
2	BatchNorm2d	Normalization	(64, T, J)	128	—	num_features=64
3	GELU	Activation	(64, T, J)	0	GELU	Non-linear activation
4	BatchNorm2d	Normalization	(128, T, J)	256	—	num_features=128
5	GELU	Activation	(128, T, J)	0	GELU	Non-linear activation
6	Conv2d	Convolution	(128, T, J)	16,512	—	kernel=1, stride=1
7	Mean	Aggregation	(T, 128)	0	—	Mean over joint dimension
8	TransformerEncoder	Transformer	(T, 128)	1,319,424	—	4 layers, 8 heads, FF=1024, dropout=0.1
9	GRU	Recurrent	(T, 256)	494,592	—	2 layers, bidirectional=True
10	Linear	Fully Connected	(256)	131,328	GELU	in=512, out=256
11	LayerNorm	Normalization	(256)	512	—	normalized_shape=256
12	GELU	Activation	(256)	0	GELU	Non-linear activation
13	Linear	Reconstruction Head	(J×2)	8,738	—	in=256, out=34
14	Linear	Latent Projection	(128)	32,896	—	in=256, out=128
15	LayerNorm	Normalization	(128)	256	—	normalized_shape=128

Table 2-Training hyperparameters

Parameter	Value
Window size (T)	32
Max persons (K)	5
Batch size	32
Learning rate	1e-4
Optimizer	AdamW
Epochs	60 (with early stopping)
Weight decay	0.02
Latent dimension	128
Transformer heads	8
Transformer layers	2
GRU layers	2

3.6. Training Objective and One-Class Learning

The model is trained using normal data only in a one-class setting. During training, two loss terms are used. The first helps the model learn to reconstruct normal movement. This allows it to understand what normal motion looks like. The second loss keeps the features close to each other in the feature space. This follows the basic idea of SVDD.
Together, these two objectives help is for the model learn a compact representation of normal motion and maintain stable feature distributions.

3.7. Learning Objective

The proposed model is trained using a hybrid objective that combines reconstruction loss and a latent-space regularization inspired by Support Vector Data Description (SVDD).

Reconstruction Loss

The reconstruction loss measures the difference between the predicted joint representation and the target joint coordinates:

$$\mathcal{L}_{rec} = \frac{1}{N}\sum_{i=1}^{N} \| x_i - \hat{x}_i \|^2 \quad (1)$$

where:

- $x_i$represents the ground-truth joint coordinates,
- $\hat{x}_i$denotes the reconstructed output,
- $N$is the number of samples in the batch.

In this work, reconstruction is applied to aggregated joint representations, where joint coordinates are averaged over the temporal dimension.

SVDD-based Latent Regularization

To enforce compactness of normal samples in the latent space, a center vector $c$is introduced:

$$\mathcal{L}_{SVDD} = \frac{1}{N}\sum_{i=1}^{N} \| z_i - c \|^2 \quad (2)$$

where:

- $z_i$is the latent embedding produced by the model,
- $c$is the center of normal data distribution,
- $N$is the batch size.

This constraint encourages normal samples to lie close to the center, while abnormal samples are expected to deviate from it.

Total Loss Function

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda \mathcal{L}_{SVDD} \quad (3)$$

where:

λ is used to balance reconstruction and latent compactness.

Anomaly Score Calculation

During inference, anomaly scores are computed based on reconstruction error:

$$Score = \max_{j} \| x_j - \hat{x}_j \|^2 \quad (4)$$

where:

- $j$indexes the joints within a sequence window.

For multi-person scenarios, the maximum score across all persons is selected.
Finally, scores are normalized to the range $[0, 1]$using min-max normalization

This formulation ensures that the model simultaneously learns accurate reconstruction and a compact representation of normal behavior

3.8. Anomaly Scoring and Frame-Level Evaluation

During inference, reconstruction errors are computed for each individual within a temporal window. The maximum error across individuals is used as the window-level anomaly score, which is then propagated to frame-level scores through temporal aggregation of overlapping windows. Final anomaly scores are normalized and evaluated using standard metrics, including AUC-ROC, AUC-PR, Equal Error Rate (EER), and TPR at fixed false positive rates.
Algorithm 1: Unsupervised One-Class Multi-Person Video Anomaly Detection
Input:
Skeletal pose sequences of detected persons, temporal window length T, maximum number of persons per window K, and training data containing only normal behavior.
Output:
Frame-level anomaly scores.

Skeletal pose data are first extracted for all detected persons and organized into continuous temporal sequences for each camera view. The data are then preprocessed using root centered and scale normalization, while missing joints are handled through temporal consistency. Motion information is encoded using velocity and acceleration features.
The pose sequences are segmented into overlapping temporal windows of length T. During training, only windows representing normal behavior are used. For each person within a window, spatial motion features are extracted using a 2D convolutional network, followed by temporal modeling through a Transformer encoder and a bidirectional GRU. Normal behavior is learned by minimizing a combined objective that includes reconstruction During testing, up to K persons are used in each window. Each person is evaluated alone. The reconstruction error is computed for each one. The highest value is taken as the score for the window. These scores are then used to get frame-level scores. The results are measured using AUC-ROC, AUC-PR, EER, and TPR.

3.9. Data Split and Evaluation Protocol

To make the evaluation fair, the data was split over time. Each camera was handled alone so the scene stays the same. The first 70% was used for training, and the rest for validation and testing. The frame order was kept, and training and testing data were separate. Only normal data was used in training, so any part with abnormal frames was removed. A small part of the data was also used to follow the training and choose the model. Training was stopped when the error did not improve, and this helps reduce overfitting. In testing, both normal and abnormal data were used. The data was divided into overlapping windows. Each frame got a score based on these windows. The same preprocessing was used for all data.The results were measured at the frame level using AUC-ROC, AUC-PR, EER, and TPR.

TABLE 3-DATA SPLIT STRATEGY USED IN THE PROPOSED FRAMEWORK

Split	Data Type	Description
Training	Normal only	Used to learn normal motion patterns
Validation	Normal only	Used for early stopping and model tuning
Testing	Normal + Abnormal	Used for final evaluation

*3.*10*. Implementation Details*

The proposed model was implemented using the PyTorch framework, and all experiments relied on structural motion data extracted from the HuVAD database. The time series were divided into fixed windows of length 32 frames (T = 32) to achieve a suitable balance between representing short-term and During testing, a fixed number of persons was retained within each temporal window, where the Top-K.
K subjects were selected according to motion saliency. Keep people with more movement and remove the rest. This reduces noise in crowded scenes and helps the model focus on clear movement. The model to focus on the most informative motion patterns. In this work, K
K was set to 5 as a practical balance between representational coverage and computational efficiency

In the training phase, AdamW was used as the optimization algorithm with an initial learning rate of 8×10^{-5} and a weight decay of 0.02. A batch size of 32 was used, and the maximum number of training epochs was set at 200. Early stopping based on validation loss was employed to mitigate the risk of over-allocation, as training was stopped when validation performance did not improve after a specified number of consecutive epochs.

The Transformer has four layers. It uses eight attention heads. The embedding size is 128. The hidden layer part uses size 1024 with dropout 0.1. The Bi-GRU has two layers, with size 128 in each direction.

After combining the features, a dropout layer is used to get a compact representation. Then two outputs are produced. One is for reconstruction. The other is for feature representation used in one-class learning.

To keep the results stable, the random seed was fixed at 42. During training, stride = 2 was used. During testing, stride = 1 was used. These values were chosen after several trials to balance accuracy and speed. The model settings and training parameters are shown in.خطأ! لم يتم العثور على مصدر المرجع . This includes the CNN, Transformer, Bi-GRU, and training setup.

TABLE 4-DETAILED ARCHITECTURE AND TRAINING HYPERPARAMETERS OF THE PROPOSED FRAMEWORK

Component	Configuration
Input representation	$T \times J \times 6$skeletal motion features
Temporal window length	$T = 32$
Maximum selected persons	$K = 5$
CNN layer 1	Conv2D ($6 \rightarrow 64$), kernel 3×3, padding 1 + BatchNorm + GELU
CNN layer 2	Conv2D ($64 \rightarrow 128$), kernel 3×3, padding 1 + BatchNorm + GELU
CNN layer 3	Conv2D ($128 \rightarrow 128$), kernel 1×1
Transformer encoder layers	4
Attention heads	8
Embedding dimension	128
Dropout	0.1
Bi-GRU hidden size	128
Bi-GRU layers	2
Bi-GRU direction	Bidirectional
Reconstruction head	Linear ($256 \rightarrow J \times 2$)
Embedding head	Linear ($256 \rightarrow 128$) + LayerNorm
Optimizer	AdamW
Learning rate	8×10^{-5}
Weight decay	0.02
Batch size	32
Maximum epochs	200
Early stopping	Based on validation loss
Patience	8
Min delta	10^{-6}
Train stride	2
Test stride	1
Regularization weight λ	0.15
Random seed	42

4. Ablation Study

An ablation study was done to test each part of the model. Each part was removed or changed one by one. The same setup and metrics were used in all tests **Table 5** presents the results of different model variants, including the removal of the Transformer module, the GRU unit, and motion-based features such as velocity and acceleration.

TABLE 5-ABLATION STUDY OF THE PROPOSED FRAMEWORK ON THE HUVAD DATASET

Model Variant	CNN	Transformer	Bi-GRU	Motion Features	AUC-ROC	AUC-PR	EER
CNN Only	✓	✗	✗	✗	86.3	58.2	0.38
Without Motion Features	✓	✓	✓	✗	89.8	61.5	0.34
Without Bi-GRU	✓	✓	✗	✓	90.5	63.9	0.31
Without Transformer	✓	✗	✓	✓	91.2	64.7	0.29
Full Model (Proposed)	✓	✓	✓	✓	93.6	68.4	0.24

In the **Table 5** Removing the Transformer reduced performance. Removing the Bi-GRU also lowered accuracy. Removing velocity and acceleration affected the results. Using only CNN gave the lowest results, which means spatial features alone are not enough in complex scenes.The full model gave the best results among all cases.

5. Discussion

The model learns normal movement and finds changes in surveillance videos. It does not use one frame only. It follows motion over time, so it sees small changes. It also works with more than one person. In training, each person is used alone. In testing, all people are used together.
The model is trained using normal data only. This fits real surveillance cases, where abnormal events are rare.

There are some limits. The results depend on the quality of the skeletal data, since missing joints or tracking errors can affect performance. The model was tested on HuVAD only. The results were similar across cameras.

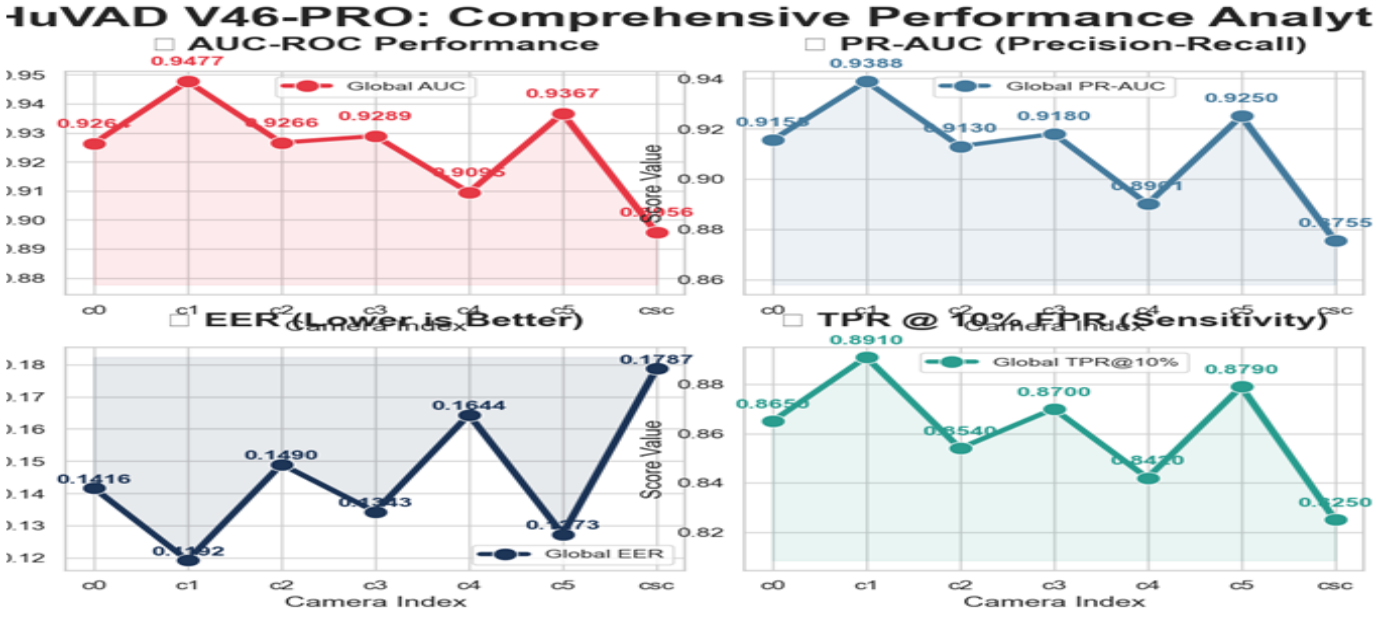
Using motion only helps with privacy, but it may not be enough in some cases. In the future, the model can be tested on more data. Simple information like audio or context can also be added.**Table 6** shows a comparison between previous studies and the proposed method on the HuVAD dataset.

TABLE 6- COMPARISON OF THE RESULTS OF THE USED METHOD AND SOME PREVIOUS METHODS ACROSS ALL CAMERAS OF THE HUVAD DATASET.

		C0				C1				C2			
Model	**Conf.**	**AUC**	**PR**	**EER**	**10ER**	**AUC**	**PR**	**EER**	**10ER**	**AUC**	**PR**	**EER**	**10ER**
MP ED-RNN	**CVPR19**	79.57	46.76	0.26	0.37	83.57	53.62	0.22	0.39	73.36	47.66	0.32	0.57
GEPC	**CVPR20**	59.07	28.00	0.44	0.71	56.27	23.20	0.45	0.78	55.09	30.13	0.45	0.79
STG-NF	**ICCV23**	58.96	83.45	0.46	0.84	47.82	78.31	0.52	0.89	51.06	74.20	0.49	0.94
TSGAD	**WACV24**	64.18	31.93	0.40	0.66	68.88	39.02	0.35	0.72	62.81	35.53	0.39	0.73
Proposed	**Ours**	92.83	71.62	0.132	0.14	94.30	67.32	0.136	0.13	94.62	69.41	0.146	0.14
		C3				**C4**				**C5**			
Model	**Conf.**	**AUC**	**PR**	**EER**	**10ER**	**AUC**	**PR**	**EER**	**10ER**	**AUC**	**PR**	**EER**	**10ER**
MPED-RNN	**CVPR19**	83.62	63.87	0.23	0.42	49.20	41.80	0.45	0.86	74.59	44.55	0.29	0.31
GEPC	**CVPR20**	52.40	27.09	0.50	0.77	42.10	39.50	0.48	0.90	72.44	30.71	0.29	0.99
STG-NF	**ICCV23**	49.15	74.10	0.50	0.90	41.30	45.60	0.47	0.88	72.24	93.95	0.28	0.71
TSGAD	**WACV24**	54.64	24.25	0.48	0.76	47.80	43.20	0.44	0.85	75.11	37.13	0.28	1.00
Proposed	**Ours**	94.68	62.70	0.175	0.17	92.81	67.53	0.165	0.16	93.09	63.25	0.221	0.23
		CSC				**Combined**							
Model	**Conf.**	**AUC**	**PR**	**EER**	**10ER**	**AUC**	**PR**	**EER**	**10ER**				
MPED-RNN	**CVPR19**	56.83	37.13	0.43	0.69	76.05	42.83	0.28	0.49				
GEPC	**CVPR20**	58.32	41.09	0.42	0.86	62.25	28.62	0.41	0.67				
STG-NF	**ICCV23**	53.60	66.28	0.47	0.92	57.57	83.77	0.46	0.90				
TSGAD	**WACV24**	58.91	43.28	0.43	0.86	68.00	34.61	0.36	0.64				
Proposed	**Ours**	92.90	67.60	0.186	0.84	93.60	68.40	0.24	0.35				

The results in Table 4 show different performance levels across cameras. The model performs better on C1 and C3, while lower results are observed on the CSC camera, where the scenes are more crowded and complex. Although the max-aggregation strategy helps detect the most suspicious individual, it can sometimes increase the effect of noisy pose estimates in very crowded scenes.

To better illustrate these differences, Fig. 3 presents the camera-wise performance of the proposed model using several evaluation metrics.



3. Figure 3Camera-wise performance of the proposed unsupervised anomaly detection framework on the HuVAD dataset

6. Conclusion and Future Work

This work studies anomaly detection in surveillance scenes with more than one person. The method is unsupervised and uses only skeleton motion.

The method was tested on HuVAD with different cameras. Time windows were used, and the results became better. This was useful in crowded scenes where people move at the same time.

There are some limits. The results depend on the skeletal data, since missing joints or tracking errors can affect the output. The results were similar across cameras.

Acknowledgments

The author is grateful to colleagues and researchers who provided valuable discussion and feedback, allowing to improve this work.

Reference

[1] A. Alnajjar et al., "Anomaly Detection Based on Hierarchical Federated Learning with Edge-Enabled Object Detection for Surveillance Systems in Industry 4.0 Scenario," International Journal of Intelligent Engineering and Systems, vol. 17, no. 4, 2024, doi: 10.22266/ijies2024.0831.49.
[2] P. Geetha and V. Narayanan, "Multi-Modal Video Summarization," International Journal of Intelligent Engineering and Systems, vol. 7, no. 3, 2014.
[3] E. Mofreh, A. Abozeid, H. Farouk, and K. A. Eldahshan, "Multi-Object Semantic Video Detection and Indexing Using a 3D Deep Learning Model," International Journal of Intelligent Engineering and Systems, vol. 15, no. 3, p. 2022, doi: 10.22266/ijies2022.0630.23.
[4] D. Karunya Sampath and K. Kumar, "Abnormal Crowd Behaviour Detection in Surveillance Videos Using Spatiotemporal Inter-Fused Autoencoder," International Journal of Intelligent Engineering and Systems, vol. 16, no. 6, p. 2023, doi: 10.22266/ijies2023.1231.39.
[5] G. Pang, C. Shen, L. Cao, and A. Van Den Hengel, "Deep Learning for Anomaly Detection: A Review," ACM Comput. Surv., vol. 54, no. 2, pp. 1–36, 2022, doi: 10.1145/3439950.
[6] J. Ren, F. Xia, Y. Liu, and I. Lee, "Deep Video Anomaly Detection: Opportunities and Challenges".
[7] K. Doshi and Y. Yilmaz, "Towards Interpretable Video Anomaly Detection," Proceedings - 2023 IEEE Winter Conference on Applications of Computer Vision, WACV 2023, pp. 2654–2663, 2023, doi: 10.1109/WACV56688.2023.00268.
[8] J. Tatiya, R. Makhija, M. Pathe, S. Late, and Prof. M. Pathak, "Anomaly Detection for Video Surveillance," Int. J. Sci. Res. Sci. Technol., pp. 82–89, 2021, doi: 10.32628/ijsrsr21869.
[9] J. Liu et al., "Networking Systems for Video Anomaly Detection: A Tutorial and Survey," 2025, [Online]. Available: http://arxiv.org/abs/2405.10347
[10] K. Faber, R. Corizzo, B. Sniezynski, and N. Japkowicz, "Lifelong Continual Learning for Anomaly Detection: New Challenges, Perspectives, and Insights," IEEE Access, vol. 12, no. March, pp. 41364–41380, 2024, doi: 10.1109/ACCESS.2024.3377690.
[11] J. Tatiya, R. Makhija, M. Pathe, S. Late, and Prof. M. Pathak, "Anomaly Detection for Video Surveillance," Int. J. Sci. Res. Sci. Technol., pp. 82–89, May 2021, doi: 10.32628/ijsrsr21869.
[12] S. Late, M. Pathe, R. Makhija, J. Tatiya, and P. M. Pathak, "Anomaly Detection through Video Surveillance using Machine Learning," pp. 1–9.
[13] S. Late, M. Pathe, R. Makhija, J. Tatiya, and M. Pathak, "Anomaly Detection through Video Surveillance using Machine Learning," 2021, doi: 10.32628/IJSRST.
[14] A. Flaborea, G. M. D. di Melendugno, S. D'Arrigo, M. A. Sterpa, A. Sampieri, and F. Galasso, "Contracting skeletal kinematics for human-related video anomaly detection," Pattern Recognit., vol. 156, pp. 1–32, 2024, doi: 10.1016/j.patcog.2024.110817.
[15] Y. Huang et al., "Track Any Anomalous Object: A Granular Video Anomaly Detection Pipeline", [Online]. Available: https://tao-25.github.io/
[16] S. Nou, J. S. Lee, N. Ohyama, and T. Obi, "Human pose feature enhancement for human anomaly detection and tracking," International Journal of Information Technology (Singapore), vol. 17, no. 3, pp. 1311–1320, 2025, doi: 10.1007/s41870-024-02363-2.
[17] A. D. Pazho, S. Yao, G. A. Noghre, B. R. Ardabili, V. Katariya, and H. Tabkhi, "Towards Adaptive Human-centric Video Anomaly Detection: A Comprehensive Framework and A New Benchmark," pp. 1–9, 2024, [Online]. Available: http://arxiv.org/abs/2408.14329
[18] G. A. Noghre, A. D. Pazho, and H. Tabkhi, "Human-Centric Video Anomaly Detection Through Spatio-Temporal Pose Tokenization and Transformer," 2025, [Online]. Available: http://arxiv.org/abs/2408.15185
[19] G. A. Noghre et al., "An Exploratory Study on Human-Centric Video Anomaly Detection through Variational Autoencoders and Trajectory Prediction," Proceedings - 2024 IEEE Winter Conference on Applications of Computer Vision Workshops, WACVW 2024, pp. 995–1004, 2024, doi: 10.1109/WACVW60836.2024.00109.
[20] J. Xiao, T. Liu, and G. Ji, "Human Kinematics-inspired Skeleton-based Video Anomaly Detection," Sep. 2023, [Online]. Available: http://arxiv.org/abs/2309.15662
[21] R. Morais, V. Le, T. Tran, B. Saha, M. Mansour, and S. Venkatesh, "Learning regularity in skeleton trajectories for anomaly detection in videos," Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2019-June, pp. 11988–11996, 2019, doi: 10.1109/CVPR.2019.01227.
[22] G. T. de Araújo and A. L. F. de Almeida, "PARAFAC-Based Channel Estimation for Intelligent Reflective Surface Assisted MIMO System," Jan. 2020, [Online]. Available: http://arxiv.org/abs/2001.06554
[23] Z. Qin et al., "Fusing Higher-Order Features in Graph Neural Networks for Skeleton-Based Action Recognition," IEEE Trans. Neural Netw. Learn. Syst., vol. 35, no. 4, pp. 4783–4797, Apr. 2024, doi: 10.1109/TNNLS.2022.3201518.
[24] X. Xiang, X. Li, X. Liu, Y. Qiao, and A. El Saddik, "A GCN and Transformer complementary network for skeleton-based action recognition," Computer Vision and Image Understanding, vol. 249, Dec. 2024, doi: 10.1016/j.cviu.2024.104213.
[25] Z. K. Abbas and A. A. Al-Ani, "DETECTION OF ANOMALOUS EVENTS BASED ON DEEP LEARNING-BILSTM," Iraqi Journal of Information and Communication Technology, vol. 5, no. 3, pp. 34–42, Dec. 2022, doi: 10.31987/ijict.5.3.207.