IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 2, April 2026, pp. 1441~1450
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i2.pp1441-1450

TMA-Net: a transformer-based multi-modal attention network for abnormal behavior detection

Huong-Giang Doan[1], Ngoc-Trung Nguyen[2]
[1]Faculty of Control and Automation, Electric Power University, Hanoi, Vietnam
[2]Department of Personnel Organization and Administration, Electric Power University, Hanoi, Vietnam

Article Info

Article history:

Received Sep 18, 2024
Revised Jan 9, 2026
Accepted Jan 25, 2026

Keywords:

Abnormal dectection
Attention network
Convolutional neural network
Spatial-temporal
Transformer

ABSTRACT

Abnormal behavior detection in crowded environments remains challenging due to complex motion patterns, occlusions, and domain variability. This paper presents transformer-based multi-modal attention network (TMA-Net), a unified framework that integrates red, green, and blue (RGB), optical flow (OF), and heat map (HM) modalities through a dual-stage attention fusion mechanism. The system employs you only look once version 11 (YOLOv11) for human localization and vision transformer (ViT)-B/16 for feature encoding, followed by intra-modal self-attention and cross-modal fusion to capture fine-grained spatial–temporal and motion energy dependencies. Extensive experiments on six public benchmarks as UMN, Crowd-11, UBNormal, ShanghaiTech, CUHK Avenue, UCSD Ped2, and EPUAbN dataset, demonstrate that TMA-Net achieves up to 97.5% area under the curve (AUC) and 96–100% accuracy, outperforming previous other state-of-the-art approaches. These results highlight the framework's strong generalization and robustness across both single- and cross-dataset evaluations, underscoring its potential for reliable deployment in real intelligent surveillance systems.

Corresponding Author:

Ngoc-Trung Nguyen
Department of Personnel Organization and Administration, Electric Power University
No. 235 Hoang Quoc Viet Street, Nghia Do Ward, Hanoi City, Vietnam
Email: trungnn@epu.edu.vn

1. INTRODUCTION

In visual tasks using convolutional neural networks (CNNs), it can be challenging for the models to process all of the input data because of its size and complexity. In order to solve this challenge, attention mechanisms are proposed to help CNNs focus on the most relevant features of the input and ignore the irrelevant ones and thereby improving the accuracy and efficiency of the learning process. Depending on different CNN architectures and the learning targets we have different types of attention mechanisms that can be deployed. For human abnormal behavior detection, the popular attention mechanisms added in CNNs are temporal attention, spatial attention, and the combination of spatial and temporal attentions.

Spatial attention mechanisms answer the question of where to pay attention to the image. They are mainly added to CNN modules as additional layers for extracting important spatial features from the CNN outputs. The framework of CNN-long-short term memory (LSTM) with attention units is proposed for human abnormal behavior detection from videos [1]. Firstly, the input images sampled from the videos are pre-processed by converting to grayscale, equalizing the histogram, and reshaping to a smaller size. They are put into CNN layers, followed by attention units, and finally directed to the LSTM layer for interpreting the features obtained from CNN. The attention units work as supplemental feature extraction layers of CNN. However, random sampling of frames from videos may skip frames containing unusual behaviors. The more complex attention structure is proposed in [2]. In this work, models of AttM-CNN-AG and AttM-CNN-Porn

are proposed for child sexual abuse content detection. In this, Inception and ResNet deep neural networks are deployed as basic units. Two attention modules are added to these deep neural models to help automatically focus on key regions in the input frames. The attention module contains a 1×1 convolution layer, followed by an element-wise dot product with the feature vector of the respective layer. This result is then normalized by the SoftMax operation. The normalized result can be considered the coefficients of the attention grid, which represent the importance of the elements in the feature maps at the chosen layer of the CNN. Although some positive results have been achieved on self-collected databases, there are still some limitations to this work. The detection results of child sexual abuse depend on the age-group classification module, which relies on the human face but no other helpful features. This has led to some child sexual abuse images being misclassified by failure of age-group classification.

Temporal attention mechanisms answer the question of when to pay attention or which frames should be focused in a frame sequence of the video. Temporal attention modules are normally applied for video processing. It relates to the motion patterns that are commonly extracted by the recurrent neural network (RNN) network. In the combined architecture of CNN and LSTM [3], CNN model is used for producing the spatial features from the input frame. These features are then directed into the LSTM module to generate temporal features. The feature maps of the LSTM component are then fed into an attention module to capture valuable and informative features in the frame of the video. The actions are recognized by the informative features using the SoftMax module.

Chong and Tay [4] proposed a spatiotemporal autoencoder using ConvLSTM to jointly model spatial and temporal information in video sequences. The model learns normal motion patterns in an unsupervised manner and detects abnormal events by measuring reconstruction errors on unseen video frames. The extensive experiments on the UCF-Crime [5], UMN [6], and Avenue [7] datasets indicate the better results compared to other state-of-the-art (SOTA) models. The above-mentioned works deploy temporal attention mechanisms in the same manner: spatial feature extraction first, then temporal feature extraction, and finally an attention mechanism is applied for weighting the temporal features. Chang *et al.* [8] proposed a clustering-driven deep autoencoder framework for video anomaly detection. In this approach, spatiotemporal features are extracted from red, green, and blue (RGB) frames and optical flow (OF) using two separate 3D CNN networks, and subsequently fused to form unified video segment representations. A deep autoencoder is employed to learn compact feature embeddings, while clustering is introduced to exploit the intrinsic structure of normal and abnormal events in a weakly supervised manner. To further enhance anomaly discrimination, multiple constraints, including event separation and temporal smoothness, are incorporated during training. Experimental results on the UCF-Crime dataset demonstrate that the proposed method achieves superior performance compared to existing approaches in detecting anomalous events.

Spatial attention aims to emphasize discriminative regions within individual video frames, while temporal attention focuses on identifying informative frames or segments in a video sequence. For abnormal human behavior detection, which relies on both appearance and motion cues, jointly modeling spatial and temporal information is essential for improving detection performance. The combination of spatial and temporal attention enables adaptive selection of important regions and moments from videos. Li *et al.* [9] proposed a spatio-temporal attention network for action recognition and detection, where spatial and temporal attention modules are embedded into a CNN to enhance discriminative feature representations. The spatial attention module highlights informative regions in video frames, while the temporal attention module assigns importance weights to key frames in a video sequence. Experimental results on the HMDB51 and UCF101 datasets demonstrate that incorporating spatio-temporal attention significantly improves performance compared to models without attention mechanisms. Building upon attention-based modeling, Chen *et al.* [10] introduced a spatial–temporal graph attention network for video anomaly detection. In this approach, spatiotemporal features extracted by a 3D CNN backbone are organized into a spatial–temporal graph, where graph attention mechanisms are employed to capture spatial relationships among regions and temporal dependencies across frames. Experimental results on the UCF-Crime dataset and a vehicle theft dataset show that the fusion of spatial and temporal graph attention outperforms using either attention alone as well as methods without attention. However, due to the reliance on local graph structures, the method may exhibit limited capability in modeling long-range temporal dependencies in extended video sequences. To address long-term temporal modeling, Liu *et al.* [11] proposed a temporal segment transformer framework with an embedded spatial–temporal attention mechanism for abnormal behavior recognition. By sampling video segments over longer time spans, the model captures long-range temporal dependencies while suppressing irrelevant frames and regions. Experimental results on UCF101, HMDB51, JHMDB, and THUMOS14 datasets demonstrate that incorporating spatial–temporal attention significantly improves recognition performance compared to models without attention mechanisms. In this work, we deploy both spatial and temporal attention units for abnormal behavior detection. However, it is different from other published methods, our proposed frameworks apply attention units on three inputs of RGB, OF, and heat map (HM) images. The attention feature

vectors from these inputs are then optimally combined to give out the final ones for classification. This allows effective exploitation and focuses on the important image features that need to be detected from many input sources. In addition, we also apply the knowledge distillation technique to the proposed framework. This aims at reducing the computing time of the detection system. The enhanced experiments are implemented on several benchmark datasets and our dataset using both single-dataset and cross-dataset evaluation strategies. The results show the outperformance of our proposed framework in detection accuracy compared to other SOTA methods. Furthermore, we also demonstrate through the experiments that using knowledge distillation technique not only reduces computation cost but also maintains high accuracy in abnormal behavior detection.

The remainder of this paper is organized as follows: section 2 firstly explains the proposed evaluation scheme. The experimental results and discussions are analyzed in section 3. Finally, section 4 concludes the proposed research directions for future works.

2. METHOD

The proposed framework for abnormal behavior detection, illustrated in Figure 1, is adapted and extended from our previous work [12], [13]. It takes three input modalities that consist of RGB, OF, and HM images to comprehensively represent spatial, temporal, and motion-energy information. An additional attention block that is highlighted in pink in Figure 1. This framework is incorporated to enhance multi-modal feature interaction and improve detection performance with vision transformer (ViT) [14] feature extraction and cross attention modalities strategies.

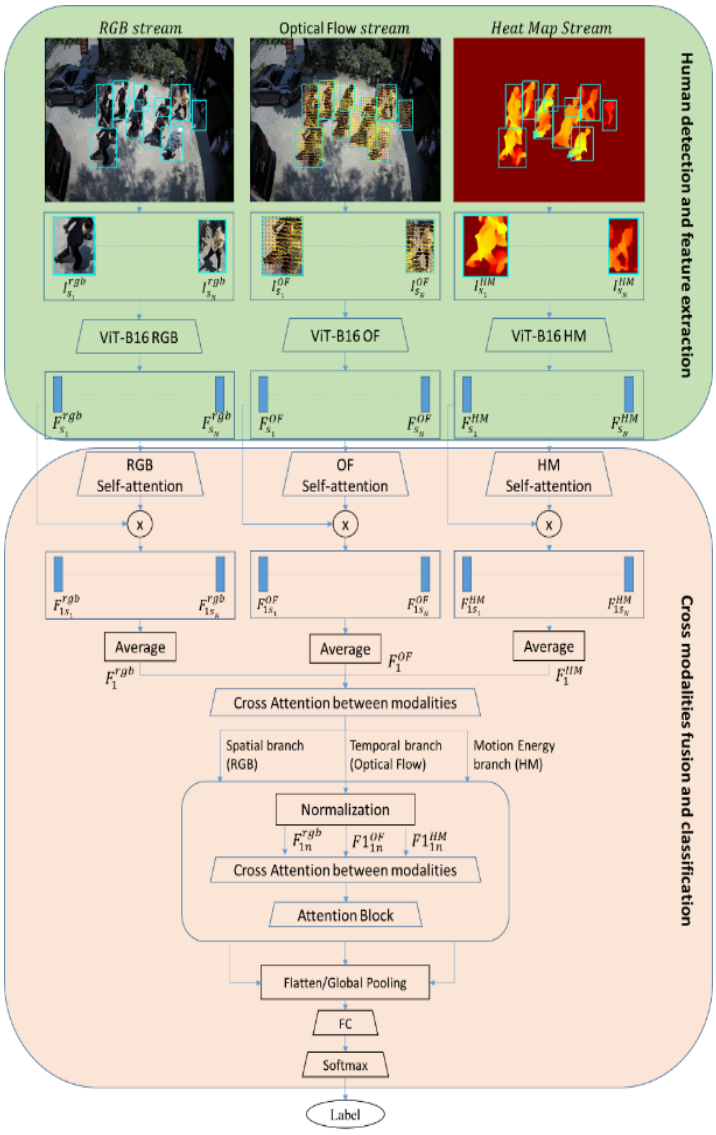

Figure 1. The ViT and cross attention modalities-based framework for abnormal detection

2.1. Human detection and feature extraction

The initial stage of the proposed system employs a you only look once (YOLO)-based detection module to localize human regions in each frame derived from RGB, OF, and HM modalities. Given an input frame $I_t \in R^{HxWx3}$, YOLO predicts a set of bounding boxes $B = \{b_i, i = (1, \dots, N)\}$ with corresponding confidence scores p_i, where each box denotes the center coordinates and dimensions of a detected person as (1).

$$B = M_{YOLOv11}(I_t), \qquad p_i = \sigma\left(W_p f(b_i) + b_p\right) \tag{1}$$

Where $f(b_i)$ represents the feature vector of the candidate region and $\sigma(.)$ is the sigmoid activation. Regions with are retained as valid detections and cropped for subsequent processing. Each detected region b_i is resized to a fixed spatial resolution of 224×224 pixels as (2).

$$R_i = Resize(I_t[b_i], 224{,}224) \tag{2}$$

These human images then passed into a ViT-B/16 encoder to obtain a discriminative feature representation. Each image R_i is divided into non-overlapping 16×16 patches, flattened, and projected into a latent embedding space via a linear mapping as (3).

$$z_p = Ex_p + E_{pos} \tag{3}$$

Where E is the patch-embedding matrix and E_{pos} denotes the positional encoding. The patch sequence is then processed by multiple transformer encoder layers to model global dependencies across patches as (4).

$$F_i \in R^C = M_{ViT-B16}(R_i) \tag{4}$$

With C being the output embedding dimension (typically C=768). The resulting vector F_i encodes the semantic and spatial context of each detected human region. The feature sets extracted from all regions in each modality are denoted as as (5).

$$F^{rgb} = \left\{F_i^{rgb}\right\}, F^{OF} = \{F_i^{OF}\}, F^{HM} = \{F_i^{HM}\} \tag{5}$$

These feature embeddings serve as inputs to the subsequent multi-modal attention fusion module, which performs cross-attention and late fusion to integrate spatial, temporal, and motion-energy information for abnormal human action recognition. The combination of YOLO and ViT-B/16 leverages the strengths of both models: i) YOLO provides efficient, real-time object localization, ensuring precise human-region extraction and background suppression; and ii) ViT-B/16 encodes global contextual dependencies within each cropped region through its self-attention mechanism. This hybrid design enables the model to capture both local spatial detail and global semantic context, forming a robust feature foundation for subsequent cross-modal fusion.

2.2. Multi-modal attention fusion and classification

2.2.1. Intra-branch self-attention

Each modality is first refined independently through self-attention mechanism to enhance intra-modal relationships. Given three modalities $m \in \{RGB, OF, HM\}$. Query, key and values are computed as (6).

$$Q^m = F^m W_Q^m,\ K^m = F^m W_K^m,\ V^m = F^m W_V^m$$
$$SA(F^m) = softmax\left(\frac{Q^m (K^m)^T}{\sqrt{d_k}}\right) V^m \tag{6}$$

Where W_Q^m, W_K^m, W_V^m are learnable projection matrices and d_k is the key dimension. This operation emphasizes the most informative spatial or temporal regions within each modality.

2.2.2. Cross-modal attention fusion (early fusion)

To synchronize contextual information among modalities, the outputs of the self-attention modules are fused via a cross-attention between modalities block. For instance, the RGB branch attends to OF and HM features as (7).

$$CA(F^{rgb}, F^{OF}, F^{HM}) = softmax\left(\frac{Q^{rgb}[K^{OF}, K^{HM}]^T}{\sqrt{d_k}}\right)[V^{OF}, V^{HM}] \quad (7)$$

And analogously for $CA(F^{OF})$ and $CA(F^{HM})$. This stage represents early fusion, aligning multi-modal semantics across spatial, temporal, and motion energy domains.

2.2.3. Global attention block (late fusion)

The aligned representations are then fed into a global attention block that performs both self-attention and cross-attention to produce a unified joint representation as (8).

$$F^{fusion} = Norm\left(\sum_{j=1}^{3} \alpha_j SA\left(F'^j + \sum_{j \neq k} \beta_{jk} CA(F'^j, F'^k)\right)\right) \quad (8)$$

Where F'^j are outputs of early-fusion stage, α_i, β_{jk} are learnable weights, and Norm denotes layer normalization. This late-fusion step captures higher-level inter-modal dependencies for robust representation learning.

2.2.4. Feature aggregation and classification

The fused feature map F^{fusion} is flattened or globally pooled and passed through a fully connected layer followed by a SoftMax activation as (9).

$$y = softmax(W_{FC} Flatten(F^{fusion}) + b) \quad (9)$$

Yielding the posterior probability vector y = [abnormal, non-abnormal]. This determines the final prediction of the scene state.

2.3. Datasets and scenarios

In this work, several benchmark datasets and one self-constructed dataset were employed to evaluate the proposed model such as: the UMN dataset [6] contains three indoor and outdoor scenes with a total of 4 min 17 s of video at 30 fps (320×240 px). Each sequence starts with normal activities and ends with abnormal panic behavior. The Crowd-11 dataset [15] defines 11 crowd motion patterns in 6,000 video clips (about 100 frames each), partly collected from WWW crowd dataset [16], CUHK [17], Violent-Flows [18], WorldExpo10 [19], AgoraSet [20], PETS [21], UMN [6], and Hockey Fight [22]. The UCF_CC_50 dataset [5] includes 50 highly crowded images with 63,974 annotated pedestrians (94 to 4,543 per image), providing a challenging benchmark for crowd-density estimation. The UCSD Ped2 dataset [23] contains 2,000 frames of a single pedestrian scene, with 11 to 46 people per frame and 49,885 labeled instances. The UBNormal dataset [24] has 236,902 synthetic frames generated from 29 natural scenes (streets, stations, offices), evenly containing both normal and abnormal events. The ShanghaiTech dataset [25] includes 437 surveillance videos (317,398 frames, 13 scenes) with 158 anomalies in 11 categories, widely used for large-scale anomaly detection. The CUHK Avenue dataset [7] consists of 15 sequences (35,240 frames) with 14 unusual events such as running, throwing, and loitering. Finally, the EPUAbN dataset (self-built) comprises 300 RGB videos captured outdoors (2688×1520 pixels, 30 fps) using fixed HiKVision DS-2CD2643G2-IZS cameras. This dataset defined 11 abnormal crowd behaviors, including fighting, robbery, fire, smoke, weapon carrying, falling objects, and sudden vehicle entry, with 5 to 25 participants per scene.

2.4. The evaluation criteria

The performance of the proposed model is evaluated using micro-area under the curve (AUC), macro-AUC [24], and micro/macro accuracy [26], [27]. The final prediction score obtained from the SoftMax layer is threshold (α equals 0.1 to 1.0) to classify each frame as normal or abnormal, and a receiver operating characteristic (ROC) curve is constructed based on the true positive rate (TPR) and false positive rate (FPR). The micro-AUC reflects the overall detection performance across all test samples, while the macro-AUC averages the AUC scores over individual videos. Accuracy is computed at α=0.5, where micro-accuracy measures the global classification correctness and macro-accuracy represents the mean accuracy per video. These metrics collectively assess both global and per-scene detection effectiveness. These metrics were concerned in detail in our previous research [13].

3. EXPERIMENTAL RESULTS

All evaluation experiments are implemented in Python using the PyTorch deep learning framework and executed on a workstation equipped with an NVIDIA GPU with 18 GB memory. Our models are trained

for 100 epochs, early stopping mode, batch size 32 and a learning rate between 10^{-6} to 10^{-4}. The proposed method is evaluated on several challenging benchmark datasets as presented in section 2.3. Two evaluation strategies are adopted: single dataset evaluation and cross dataset evaluation. In single dataset evaluation, each dataset is divided into training and testing splits according to its original protocol. In cross dataset evaluation, one dataset is used entirely for training, while another is used for testing to examine the model's cross-domain generalization capability. The proposed framework is evaluated under both strategies, and their results are presented in the following sections.

3.1. Single-dataset evaluation

The single dataset evaluation were conducted using AUC [24] and accuracy [26], [27] metrics across six benchmark anomaly detection datasets: UBNormal, ShanghaiTech, CUHK Avenue, UMN, UCSD Ped2, and the proposed EPUAbN dataset. Figures 2 and 3 report the corresponding micro and macro results for AUC and accuracy, respectively, comparing the proposed TMA-Net model with prior methods including ROHAC [13], ROHAC-KD [12], ROHAC V2 [13], and ROHAC-KD V2 [13].

3.1.1. AUC-based evaluation

As illustrated in Figure 2, the proposed TMA-Net achieves the highest AUC values across all datasets, consistently outperforming the existing ROHAC-based frameworks. In particular, on UBNormal and ShanghaiTech, TMA-Net yields improvements of approximately 0.9% micro-AUC and 1.5% macro-AUC over ROHAC V2 [13], highlighting its enhanced sensitivity to subtle abnormal motion cues in complex synthetic and real-world crowd scenes. For CUHK Avenue, UMN, and UCSD Ped2, the AUC scores of TMA-Net are nearly saturated, reaching 97% to 99%, comparable to the highest reported results in literature. On the EPUAbN dataset that collected under diverse outdoor conditions, TMA-Net achieves 99.5% micro-AUC and 99.6% macro-AUC, confirming its robustness to illumination changes, motion clutter, and scale variations. These AUC results demonstrate that TMA-Net effectively captures multi-level spatiotemporal dependencies through its dual attention fusion strategy. The consistent gain across heterogeneous datasets indicates a strong capacity for generalizing both spatial structures from static scenes and dynamic motion correlations from temporal patterns.

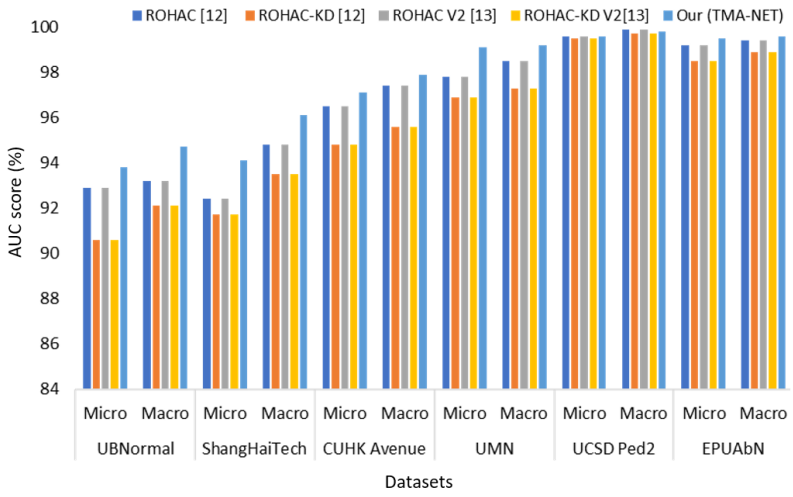

Figure 2. The micro-AUC (%) and macro-AUC (%) results on single dataset evaluation

3.1.2. Accuracy-based evaluation

The accuracy results in Figure 3 reinforce the AUC findings. TMA-Net again attains the best overall performance, with micro-accuracy and macro-accuracy exceeding those of all baseline models. Specifically, it achieves 95.1 to 96.7% on UBNormal, 94.6 to 95.2% on ShanghaiTech, and up to 98.9% to 100% on UMN, UCSD Ped2, and EPUAbN datasets. Compared to ROHAC V2, TMA-Net improves by an average of 1.2 to 2.5%, while maintaining stable results across all domains. The gain in accuracy demonstrates that the proposed fusion architecture not only enhances anomaly discrimination at the feature level but also yields more reliable final classification decisions. Notably, TMA-Net achieves perfect accuracy at 100% on UCSD Ped2 and

EPUAbN, suggesting that its attention-based feature alignment successfully preserves temporal coherence even in simpler or well-structured environments.

The superior results of TMA-Net across both AUC and accuracy metrics confirm its ability to extract discriminative representations and generalize across diverse data distributions. The improvements are particularly prominent in challenging datasets such as UBNormal and ShanghaiTech, where scene complexity, camera angles, and human density vary significantly. By combining spatial, temporal, and motion energy cues through dual-stage attention, TMA-Net mitigates overfitting to dataset-specific contexts and ensures consistent performance on unseen scenes. Overall, the single dataset evaluation results clearly indicate that TMA-Net outperforms all existing ROHAC variants in both frame-level detection precision and video-level stability. This establishes TMA-Net as a robust and scalable framework for "hand-in-wild" abnormal behavior detection across multiple environments and data domains.

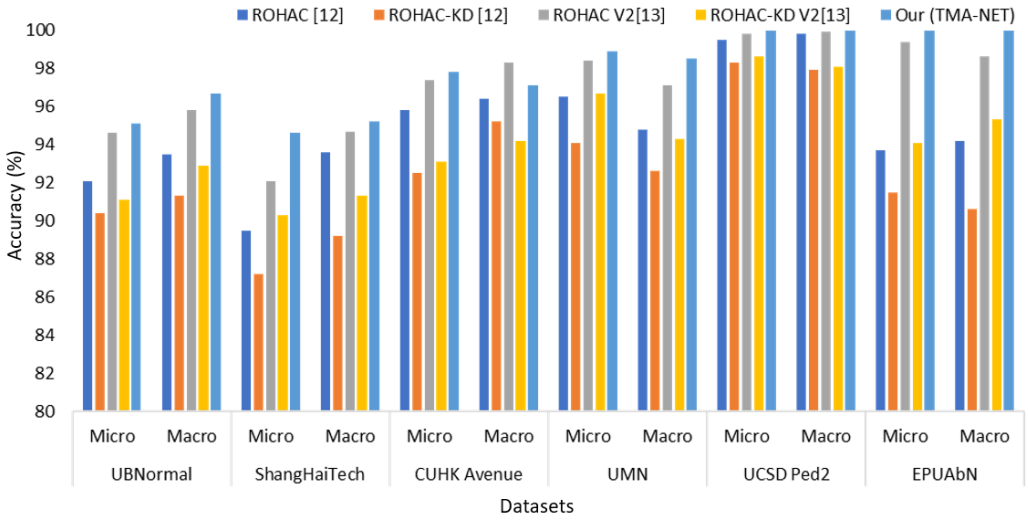

Figure 3. The micro accuracy (%) and macro accuracy (%) results on single dataset evaluation

3.2. Cross-dataset evaluation

Cross-dataset experiments were conducted to evaluate the generalization capability of the proposed models under domain shifts between training and testing datasets. Following the same experimental setup as in [28], one dataset was used for training while another was reserved for testing. Each experiment was repeated five times, and the average micro-AUC and macro-AUC scores were reported. The results over three benchmark datasets, such as: CUHK Avenue, ShanghaiTech, and UCSD Ped2 which are summarized from Tables 1 to 3. Overall, both ROHAC V2 [13] and the proposed TMA-Net demonstrate consistent superiority compared with previous methods, including Georgescu *et al.* [28] and ROHAC [12]. Across all dataset pairs, TMA-Net achieves the highest scores in both micro-AUC and macro-AUC, confirming its strong adaptability across heterogeneous environments.

As shown in Table 1, when trained on ShanghaiTech or UCSD Ped2 and tested on CUHK Avenue, the proposed TMA-Net achieves 97.2% micro-AUC and 97.5% macro-AUC, surpassing [28] by 4.9% and 7.1%, respectively. Compared with ROHAC V2 [13], the gains are smaller but consistent, indicating that TMA-Net preserves the stability of ROHAC while improving cross-domain feature generalization. This improvement suggests that the dual-stage attention mechanism effectively captures high-level spatiotemporal semantics invariant to dataset-specific differences.

Table 1. The micro-AUC and macro-AUC (%) results on cross dataset evaluation of CUHK Avenue dataset

Method	CUHK Avenue		ShanghaiTech		UCSD Ped2	
	Micro-AUC	Macro-AUC	Micro-AUC	Macro-AUC	Micro-AUC	Macro-AUC
Georgescu *et al.* [28]	92.3	90.4	83.6	81	-	-
ROHAC [12]	93.7	94.8	93.8	94.5	92.5	95.2
ROHAC V2 [13]	96.1	95.7	95.4	97.2	95.3	96.7
Our (TMA-NET)	97.2	97.5	96.4	98.1	96.9	97.1

When evaluating the ShanghaiTech dataset (Table 2), the domain gap becomes more challenging due to complex crowd dynamics and diverse camera viewpoints. Despite this, TMA-Net achieves 96.4% micro-AUC and 98.1% macro-AUC, outperforming Georgescu *et al.* [28] by 12.8% and 17.1%, respectively, and exceeding ROHAC V2 [13] by approximately 1%. These results demonstrate the robustness of TMA-Net in modeling global-local motion correlations and its superior ability to transfer knowledge between scenes with different densities and motion distributions.

Table 2. The micro-AUC and macro-AUC (%) results on cross dataset evaluation of ShanghaiTech dataset

ROHAC-KD	ShanghaiTech		CUHK Avenue		UCSD Ped2	
	Micro-AUC	Macro-AUC	Micro-AUC	Macro-AUC	Micro-AUC	Macro-AUC
Georgescu *et al.* [28]	82.7	89.3	76.3	86.3	-	-
ROHAC [12]	92.4	94.8	91.9	90.1	92.3	89.9
ROHAC V2 [13]	95.1	96.2	93.2	92.6	95.1	94.7
Our (TMA-NET)	96.3	97.8	94.6	94.8	96.8	95.9

Table 3 further shows that when trained on other datasets and tested on UCSD Ped2, TMA-Net continues to yield near-saturated results, achieving 96.9% micro-AUC and 97.1% macro-AUC. This consistency highlights the model's capacity to generalize even in simpler surveillance environments with lower scene variability. Moreover, the marginal difference between ROHAC V2 and TMA-Net indicates that both frameworks maintain stable detection accuracy while improving inter-dataset adaptability.

Table 3. The micro-AUC and macro-AUC (%) results on cross dataset evaluation of UCSD Ped2 dataset

ROHAC-KD	UCSD Ped2		CUHK Avenue		ShanghaiTech	
	Micro-AUC	Macro-AUC	Micro-AUC	Macro-AUC	Micro-AUC	Macro-AUC
Georgescu *et al.* [28]	98.7	99.7	87	97.2	90.6	95.7
ROHAC [12]	99.6	99.9	94.8	97.5	95.8	97.8
ROHAC V2 [13]	99.6	99.9	97.2	99.1	97.9	98.7
Our (TMA-NET)	99.7	99.9	98.6	99.4	98.5	99.1

The cross-dataset evaluations demonstrate that the proposed TMA-Net exhibits remarkable robustness and generalization compared with both previous ROHAC variants and the SOTA method [28]. The improvements range from 5% to 17% in AUC scores across all dataset pairs. Such stability under different training and testing domains indicates that the multi-modal attention fusion mechanism enhances feature transferability and reduces overfitting to dataset-specific patterns. Therefore, TMA-Net not only performs well under intra-dataset evaluations but also maintains superior accuracy in cross-domain scenarios, a key requirement for real-world abnormal behavior detection systems deployed in diverse surveillance contexts.

4. CONCLUSION

This paper proposed TMA-Net, a multi-modal attention-based framework for abnormal behavior detection. By integrating RGB, OF, and HM modalities through a dual-stage attention fusion mechanism, TMA-Net effectively captures both spatial-temporal and motion-energy dependencies. Experimental results on six benchmark datasets (UBNormal, ShanghaiTech, CUHK Avenue, UMN, UCSD Ped2, and EPUAbN) demonstrate that TMA-Net achieves up to 97–100% AUC and accuracy, the outperforming all previous ROHAC-based and SOTA methods. These results highlight its strong generalization ability, robustness, and practical potential for "hand-in-wild" intelligent surveillance and abnormal behavior detection systems.

ACKNOWLEDGMENTS

This research was supported by the implementation of the scientific research project at Electric Power University in 2025 for staff and employees, project code DTKHCN.09/2025 with title: "Research and design of a monitoring and management system for computer-based multiple-choice examination rooms".

FUNDING INFORMATION

This research was funded by Electric Power University under the scientific research project for staff and employees in 2025, project code DTKHCN.09/2025.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Huong-Giang Doan	✓	✓	✓	✓	✓	✓			✓	✓		✓	✓	✓
Ngoc-Trung Nguyen		✓		✓		✓	✓	✓		✓	✓			

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflict of interest.

INFORMED CONSENT

Informed consent is not applicable for this study as the datasets used are publicly available and contain no identifiable personal information.

ETHICAL APPROVAL

Ethical approval is not applicable for this study since no human or animal subjects were involved

DATA AVAILABILITY

Public datasets used in this study (UMN, Crowd-11, UBNormal, ShanghaiTech, CUHK Avenue, and UCSD Ped2) are available from their original sources. The EPUAbN dataset generated during this study is available from the corresponding author upon reasonable request.

REFERENCES

[1] R. T. Ionescu, F. S. Khan, M.-I. Georgescu, and L. Shao, "Object-centric auto-encoders and dummy anomalies for abnormal event detection in video," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 7834–7843, doi: 10.1109/CVPR.2019.00803.

[2] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection - a new baseline," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 6536–6545, doi: 10.1109/CVPR.2018.00684.

[3] M. Hasan, J. Choi, J. Neumann, A. K. R. -Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 733–742, doi: 10.1109/CVPR.2016.86.

[4] Y. S. Chong and Y. H. Tay, "Abnormal event detection in videos using spatiotemporal autoencoder," *Advances in Neural Networks - ISNN 2017*, 2017, pp. 189–196, doi: 10.1007/978-3-319-59081-3_23.

[5] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2013, pp. 2547–2554, doi: 10.1109/CVPR.2013.329.

[6] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2009, pp. 935–942, doi: 10.1109/CVPR.2009.5206641.

[7] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 FPS in MATLAB," in *IEEE International Conference on Computer Vision*, Sydney, Australia, 2013, pp. 2720–2727, doi: 10.1109/ICCV.2013.338.

[8] Y. Chang, Z. Tu, W. Xie, and J. Yuan, "Clustering driven deep autoencoder for video anomaly detection," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 329–345, doi: 10.1007/978-3-030-58555-6_20.

[9] J. Li, X. Liu, W. Zhang, M. Zhang, J. Song, and N. Sebe, "Spatio-temporal attention networks for action recognition and detection," *IEEE Transactions on Multimedia*, vol. 22, no. 11, pp. 2990–3001, Nov. 2020, doi: 10.1109/TMM.2020.2965434.

[10] H. Chen, X. Mei, Z. Ma, X. Wu, and Y. Wei, "Spatial–temporal graph attention network for video anomaly detection," *Image and Vision Computing*, vol. 131, Mar. 2023, doi: 10.1016/j.imavis.2023.104629.

[11] H. C. Liu, J. H. Chuah, A. S. M. Khairuddin, X. M. Zhao, and X. D. Wang, "Campus abnormal behavior recognition with temporal segment transformers," *IEEE Access*, vol. 11, pp. 38471–38484, 2023, doi: 10.1109/ACCESS.2023.3266440.

[12] A. D. Ho, H. G. Doan, and T. T. Thuy, "Multi-modality abnormal crowd detection with self-attention and knowledge distillation," *Engineering, Technology and Applied Science Research*, vol. 14, no. 5, pp. 16674–16679, 2024, doi: 10.48084/etasr.8194.

[13] A. D. Ho, H. G. Doan, and N. T. Nguyen, "Abnormal human behavior detection improvement with an efficient attention block," *Engineering, Technology and Applied Science Research*, vol. 15, no. 4, pp. 25048–25054, 2025, doi: 10.48084/etasr.11463.

[14] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: transformers for image recognition at scale," 2021, arXiv:2010.11929.

[15] C. Dupont, L. Tobias, and B. Luvison, "Crowd-11: a dataset for fine grained crowd behaviour analysis," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jul. 2017, pp. 2184–2191, doi: 10.1109/CVPRW.2017.271.

[16] J. Shao, C. C. Loy, K. Kang, and X. Wang, "Crowded scene understanding by deeply learned volumetric slices," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 613–623, Mar. 2017, doi: 10.1109/TCSVT.2016.2593647.
[17] J. Shao, C. C. Loy, and X. Wang, "Scene-independent group profiling in crowd," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp. 2227–2234, doi: 10.1109/CVPR.2014.285.
[18] T. Hassner, Y. Itcher, and O. K. -Gross, "Violent flows: real-time detection of violent crowd behavior," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2012, pp. 1–6, doi: 10.1109/CVPRW.2012.6239348.
[19] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 833–841, doi: 10.1109/CVPR.2015.7298684.
[20] P. Allain, N. Courty, and T. Corpetti, "AGORASET: a dataset for crowd video analysis," in *1st ICPR International Workshop on Pattern Recognition and Crowd Analysis*, 2012, pp. 1–6.
[21] T. Ellis, "Performance metrics and methods for tracking in surveillance," in *3rd IEEE Workshop on Performance Evaluation of Tracking and Surveillance*, Oct. 2002, pp. 26–31.
[22] E. B. Nievas, O. D. Suarez, G. B. García, and R. Sukthankar, "Violence detection in video using computer vision techniques," *Computer Analysis of Images and Patterns*, Berlin, Heidelberg: Springer, 2011, pp. 332–339, doi: 10.1007/978-3-642-23678-5_39.
[23] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2010, pp. 1975–1981, doi: 10.1109/CVPR.2010.5539872.
[24] A. Acsintoae *et al.*, "UBnormal: new benchmark for supervised open-set video anomaly detection," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 20111–20121, doi: 10.1109/CVPR52688.2022.01951.
[25] W. Luo, W. Liu, and S. Gao, "A revisit of sparse coding based anomaly detection in stacked RNN framework," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 341–349, doi: 10.1109/ICCV.2017.45.
[26] B. Y. -Meng, W. Yang, and W. S. -Shen, "Detection of abnormal human behavior in video images based on a hybrid approach," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 11, 2022, doi: 10.14569/IJACSA.2022.0131138.
[27] H. Bagherinezhad and S. Y. Soltani, "Abnormal human behavior detection system in video surveillance systems," *SSRN Electronic Journal*, 2022, doi: 10.2139/ssrn.4106323.
[28] M. I. Georgescu, R. Ionescu, F. S. Khan, M. Popescu, and M. Shah, "A background-agnostic framework with adversarial training for abnormal event detection in video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 4505–4523, 2022, doi: 10.1109/TPAMI.2021.3074805.

BIOGRAPHIES OF AUTHORS

Huong-Giang Doan received B.E. degree in Instrumentation and Industrial Informatics in 2003, M.E. in Instrumentation and Automatic Control System in 2006 and Ph.D. in Control Engineering and Automation in 2017, all from Hanoi University of Science and Technology, Hanoi, Vietnam. She can be contacted at email: giangdth@epu.edu.vn.

Ngoc-Trung Nguyen received B.E. degree in Power System in 2003, M.E in Electrical Engineering in 2006, all from Hanoi University of Science and Technology, Hanoi, Vietnam; received Ph.D. in Electrical Engineering from University of Palermo, Palermo, Italy, in 2014. He can be contacted at email: trungnn@epu.edu.vn.