Abnormal Human Behavior Detection Improvement with an Efficient Attention Block

Anh Dung Ho
Department of Information Technology, East Asia University of Technology, Hanoi, Vietnam
dungha@eaut.edu.vn

Huong Giang Doan
Faculty of Control and Automation, Electric Power University, Hanoi, Vietnam
giangdth@epu.edu.vn

Ngoc Trung Nguyen
Department of Personnel Organization and Administration, Electric Power University, Hanoi, Vietnam
trungnn@epu.edu.vn (corresponding author)

Received: 13 April 2025 | Revised: 30 April 2025 and 6 May 2025 | Accepted: 10 May 2025

 DOI: https://doi.org/10.48084/etasr.11463

ABSTRACT

Convolution Neural Networks (CNNs) have become an attractive method for the detection of anomalous behaviors. However, designing an efficient CNN model in terms of classification accuracy remains a challenging problem. Furthermore, the existing datasets for abnormal behavior detection are limited, with each focusing on a certain context. Therefore, a CNN model trained on a certain dataset will be adaptive for a particular context and not suitable for other contexts. This study proposes a CNN framework with an efficient attention mechanism to capture key information from multiple inputs, namely RGB, optical flow, and heatmap. Experiments were carried out on several benchmark datasets and a self-collected dataset, and the evaluation involved both single- and cross-dataset strategies. The results show the superior performance of the proposed frameworks compared to other SOTA methods in detection accuracy.

Keywords-knowledge distillation; convolutional neural network; transfer learning; deep learning; student-teacher model

I. INTRODUCTION

In visual tasks using CNNs, it can be challenging to process all the input data due to its size and complexity. To address this challenge, attention mechanisms are proposed to help CNNs focus on the most relevant features of the input and ignore irrelevant ones, thus improving the accuracy and efficiency of the learning process. Depending on different CNN architectures and learning goals, different attention mechanism types can be implemented. For the detection of abnormal human behavior, popular attention mechanisms added in CNNs are temporal attention, spatial attention, and their combination.

Spatial attention mechanisms answer the question of where to pay attention to the image, added to CNNs as additional layers for extracting important spatial features from the CNN outputs. In [1], a CNN-LSTM model with attention units was proposed for the detection of abnormal human behavior in videos. First, the input images sampled from the videos were preprocessed by converting them to grayscale, equalizing the histogram, and reshaping them to a smaller size. Then, they were inputted into CNN layers, followed by attention units, and finally directed to the LSTM layer for interpreting the features obtained from the CNN. The attention units work as supplementary feature extraction layers for CNNs. However, random sampling of video frames may skip those that contain unusual behaviors. In [2], a more complex attention structure was proposed using AttM-CNN-AG and AttM-CNN-Porn models for the detection of child sexual abuse content. These models employed Inception and ResNet deep neural networks as basic units and two attention modules were added to help automatically focus on key regions in the input frames. The attention module contained a 1×1 convolution layer, followed by an element-wise dot product with the feature vector of the respective layer. The result was then normalized using softmax. The normalized result was considered to be the attention grid coefficients, which represented the importance of the elements in the feature maps at the chosen layer of the CNN. Although some positive results were achieved on self-collected datasets, there are still some limitations. The detection results of child sexual abuse depend on the age-group classification module, which is based on the human face but not other useful features. This has led to some child sexual abuse images being misclassified due to a 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. Temporal attention modules are normally applied for video processing and relate to the motion patterns that are commonly extracted by an RNN network. In [3], in a combined CNN and LSTM architecture, CNN was used to extract spatial features from the input frame, which were then directed to the LSTM module to generate temporal features. The feature maps of the LSTM component were then fed into an attention module to capture informative features in the video frame. Actions were recognized by informative features using the softmax module. A similar approach was also followed in [4], where a pre-trained lightweight CNN model was used to extract features from video frames. A feature vector was created for each 30 frames of the video. A residual LSTM module added with a self-attention layer used this feature vector to output a context-aware vector and a temporal order representation for sequential features to recognize anomalous actions. Extensive experiments on the UCF-Crime [5], UMN [6], and Avenue [7] datasets showed better results compared to other SOTA models. This work deployed temporal attention mechanisms in the same manner: first spatial feature extraction, then temporal feature extraction, and finally an attention mechanism to weight the temporal features. In [8], spatial and temporal features were extracted from RGB frames and optical flows by two 3D CNN networks and then concatenated to form spatial-temporal feature vectors. These vectors were directed to a temporal fusion layer to obtain distinctive representations. The output of the temporal fusion layer was a temporal fusion feature of a video segment, which was directed to the classification layer to predict its event score. Finally, three constraints, event category, event separation, and temporal smoothness, were considered to learn the anomaly scores and the specific categories of abnormal events. Experiments were carried out on the UCF Crime dataset, showing the outstanding performance of this method compared to others in both the detection and classification of abnormal events.

The spatial attention mechanism only focuses on key regions in the image/frame, and the temporal attention mechanism attends to important frames in the frame sequence of the video. For the problem of human abnormal behavior detection, which is commonly processed in videos presenting human movements, both the spatial and temporal domains should be considered to improve detection performance. The combination of spatial and temporal attention helps to adaptively select both important regions and frames from the video. In [9], the STACNet (Spatio-Temporal Attention ConvNet) framework was proposed, with two spatial and temporal attention modules embedded into a ConvNet. The spatial attention unit was used for pixel-wise weighting of the feature map by combining the value and gradient features of the feature map. The temporal attention module was utilized for keyframe weighting in a frame sequence based on global average and global max pooling. The weighted frames were then directed to ConvNet, which was embedded with a spatial attention module. The experimental results showed that STACNet obtained superior performance on the HMDB51 and UCF101 datasets.

The sequential execution of spatial attention and temporal attention mechanisms was also followed in [10]. However, the input feature maps for the temporal attention unit were extracted from the Inception3D network, and the spatial-temporal attention modules utilized 3D average pooling. A comparison of individual and fusion with attention mechanisms on vehicle thief data obtained from the Internet and the UCF-Crimes dataset showed that the best results were achieved with the combination of temporal and spatial attention mechanisms. The proposed method outperformed other solutions without attention mechanisms. However, this solution pays attention to short-term motions and is not suitable for the recognition of long-term actions. To solve this, in [11] a Temporal Segment Network (TSN) embedded with a Spatial-Temporal Attention (STA) unit was proposed to capture long-term information and skip some unrelated frames or areas in the video. The spatial attention module used a soft attention mechanism with spatial pyramid pooling, and the temporal attention module utilized a soft attention mechanism based on LSTM to learn temporal attention weights. Experimental results on four public datasets (UCF101, HMDB51, JHMDB, and THUMOS14) showed that the proposed framework with temporal-spatial attention was superior to one without attention units.

This study deploys both spatial and temporal attention units for the detection of abnormal behaviors. However, this approach differs from other published methods in the application of attention units to three inputs: RGB, optical flow, and heat map images. The attention feature vectors from these inputs are then optimally combined to provide the final ones for classification. This allows for effective exploitation and focuses on the important image features that need to be detected from many input sources. In addition, this framework also applies the Knowledge Distillation (KD) technique to reduce computation time. Experiments were implemented on several datasets using both single- and cross-dataset evaluation strategies. The results demonstrate that the proposed framework outperformed other SOTA methods in detection accuracy. Furthermore, it is shown that using KD techniques not only reduces computational cost but also maintains high accuracy in the detection of abnormal behaviors.

II. PROPOSED METHOD

This study improves the ROHAC and ROHAC_KD frameworks in [12] with an additional attention block, assigned ROHAC V2 and ROHAC_KD V2. In addition, a continuous learning strategy is also proposed to avoid overfitting and underfitting problems from different human behavior abnormal datasets.

A. Human Abnormal Action Detection Framework

Figure 1 shows the proposed framework for the detection of abnormal behaviors, which was inherited from [12]. Its inputs consist of RGB, optical flow, and heatmap images. This framework is improved by an attention block (pink box in Figure 1).

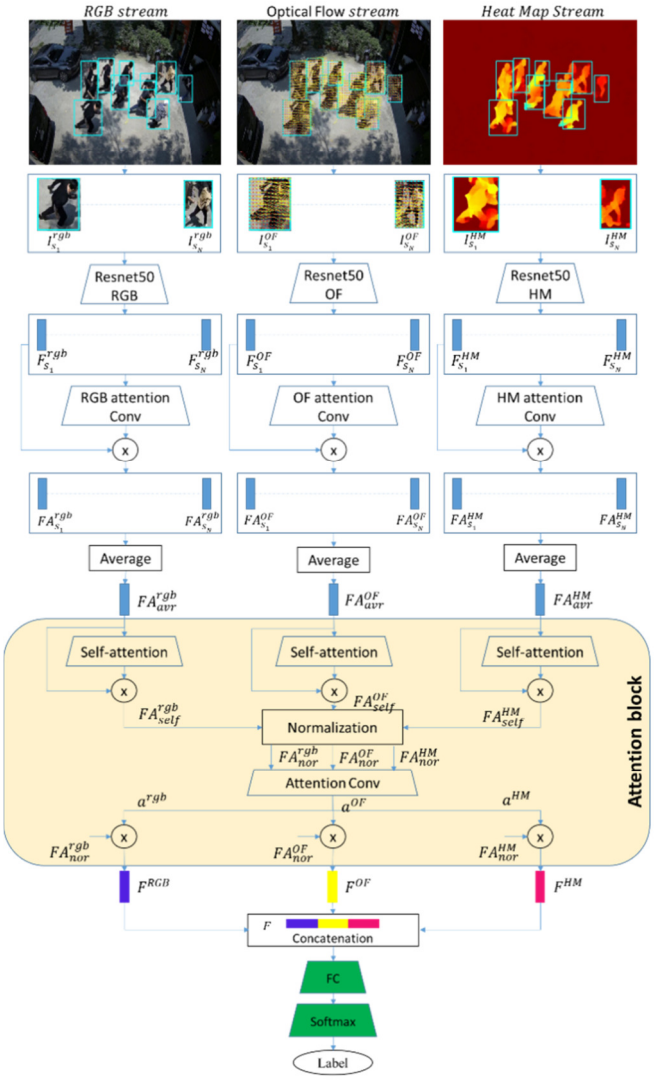

Fig. 1. The proposed framework for human anomalous action detection.

RGB, optical flow, and heat map streams are processed to obtain three context vectors $FA_{s_i}^{j} = \{FA_{s_i}^{rgb}, FA_{s_i}^{OF}, FA_{s_i}^{rgb}\}$, $j = (1,2,3)$. Each of these vectors is a weighted sum of the value vectors $F_{s_i}^{j} \in R^{1\,x\,M} = \{F_{s_i}^{rgb}, F_{s_i}^{OF}, F_{s_i}^{rgb}\}$, $M = 2048$. In the next step, the context vectors of each data type are passed over a self-attention convolution. Self-attention allows attendance to other elements, learning to assign attention weights to other elements, and enabling the model to focus on the most important parts. Self-attention processes three modalities (RGB, Optical Flow, and Heatmap) in parallel, making it efficient for anomaly detection models. Attention values are calculated as:

$$Attention_j(Q_j, K_j, V_j) = softmax\left(\frac{Q_j K_j^T}{\sqrt{d_k^j}}\right) x\, V_j \quad (1)$$

The output of the j^{th} self-attention module ($FA_{self}^{j} \in R^{1xM}$) on j^{th} channel is calculated as:

$$FA_{self}^{j} = Attention_j(Q_j, K_j, V_j) \otimes FA_{s_i}^{j} \quad (2)$$

where $j = (1,2,3)$. Three feature vectors from these self-attentions (FA_{self}^{j}) are then normalized and put into an attention conv module. Three attention convs are applied to generate attention scores for each stream. The attention convs use F_{self}^{rgb}, F_{self}^{OF}, and FA_{self}^{HM} as inputs and output the attention scores $a_j = \{a^{rgb}, a^{OF}, a^{HM}\}$ for RGB, optical flow, and heat map streams, respectively. The attention scores are calculated using the sigmoid and L_1 normalization functions [13] as:

$$a^j = \frac{\sigma^{x_i^j}}{\sum_{n=1}^{M} \sigma^{x_i^j}} = \frac{\frac{1}{1-e^{x_i^j}}}{\sum_{n=1}^{M} \frac{1}{1-e^{x_i^j}}} \quad (3)$$

The attention score of each stream is a weighted sum of the value vectors F^j, $j = (1,2,3)$ in $\{F^{rgb}, F^{OF}, F^{HM}\}$, respectively, which are then used to compute the context vectors F_{self}^{j}, $(j = (1,2,3))$ in $\{F_{nor}^{rgb}, F_{nor}^{OF}, FA_{nor}^{HM}\}$ as:

$$F^j = a^j * FA_{nor}^{j} \quad (4)$$

The three average feature vectors of RGB, optical flow, and heat map streams $F^j \in R^{1\times M}$, $j = (1,2,3)$ are then concatenated into a feature vector $F \in R^{1\times(3*M)}$ as:

$$F \in R^{1x(3*M)} = [F^{rgb}, F^{OF}, F^{HM}] \quad (5)$$

The final feature vector F passes then through the softmax layer. The softmax cross-entropy loss function is used to train the attention networks and classify abnormal actions. Given the predicted results of abnormal gestures $\overline{p_i}$, ground truth values are p_i, $i = (1, \ldots, K)$, and the loss function is calculated as:

$$L_{softmax} = \frac{1}{K}\sum_{i=1}^{K} p_i log\overline{p_i} \quad (6)$$

III. DATASETS AND SCENARIOS

A. *Published Experimental Datasets*

- The UMN dataset [6] contains videos captured in three different indoor and outdoor scenes with a total of 4 minutes and 17 seconds of 30 fps video. The frame size is 320×240 pixels. Each video starts with video segments of normal behaviors and ends with sequences of abnormal behaviors. This dataset focuses only on a single behavior of panic movement.
- The Crow-11 dataset [14] defines 11 crowd motion patterns that appear in 6000 video sequences with an average length of 100 frames per sequence. Most of the videos were manually selected and extracted from the web. In addition, some of the existing datasets are also partly used, such as WWW [15], CUHK [16], Violent-Flows [17], World Expo 10 [18], Agoraset [19], PETS [20], UMN [6], Hockey Fight and Movies [21].
- The UCF CC 50 dataset [5] contains 50 frames of 50 extremely dense crowd scenes with a total of 63,974 pedestrians annotated by heads. Each frame has from 94 to 4,543 heads. The images were mainly collected from FLICKR and used for crowd counting. Although it is a

small dataset, it is very challenging due to the large variance.

- The UCSD Ped 2 dataset [22] contains 2,000 images captured from one scene. The number of people in a frame ranges from 11 to 46, and the total number of labeled pedestrians is 49,885.
- The UBNormal dataset [23] includes a total of 236,902 frames of scenes created using 29 natural 2D background images of street scenes, train stations, and office rooms, among others. A virtual 3D scene was created from each natural image. On average, 19 videos per scene were generated. Both normal and abnormal videos were generated in the same proportion for each scene.
- The ShanghaiTech dataset [24] contains 317,398 frames from 437 videos captured from 13 different scenes. It has 158 anomalies that belong to 11 categories.
- The CUHK Avenue dataset [7] contains 35,240 frames of 15 video sequences that are about 2 minutes long for each. There are 14 unusual events, including running, throwing objects, and loitering.

These datasets either have too few or too many people in the scenes. In addition, the context appearances in these datasets are quite specific. Some datasets lack diversity in scenarios, with certain situations, such as protests that rarely appear in some countries, or they do not contain abnormal actions such as using a machete, appearing sick, etc.

B. Self-Built Dataset

This research built a new dataset on abnormal behavior in crowd scenes, called EPUAbN. Abnormal behaviors were captured outdoors at an amusement park and on the street. As shown in Table I, 11 abnormal actions in crowd scenes are defined in the dataset. These abnormal actions are captured from four HiK-Vision cameras that are fixed in positions as shown in Figure 2. The cameras are DS-2CD2643G2-IZS, and capture RGB videos at a resolution of 2688×1520 pixels at 30 fps.

TABLE I. ABNORMAL ACTIONS DEFINED IN THE EPUABN DATASET

No	Abnormal action	Description
Ab1	Fighting event 1	Appearance of fight event of two people
Ab2	Fighting event 2	Appearance of a fight event of more than two people
Ab3	Falling object(s)	Appearance of the falling objects
Ab4	Weapon(s) wearing	A person is wearing a weapon (knife, gun, sword,...)
Ab5	Sick person	A sick person suddenly falls
Ab6	Smoke	Appearance of smoke
Ab7	Fire	Appearance of fire
Ab8	Robbery action	A person is suddenly robbed
Ab9	Smash objects	A person is suddenly smashed
Ab10	Moving motorbike	A moving motorbike suddenly enters
Ab11	Moving car	A moving car suddenly enters

Each abnormal action is implemented from 10 to 20 times and stored in an RGB video with an average length of about 12 minutes. There is a total of 300 videos in the EPUAbN dataset. Each RGB video contains carefully annotated normal and abnormal actions. The number of participants in the crowd scenes ranges from five to twenty-five people.

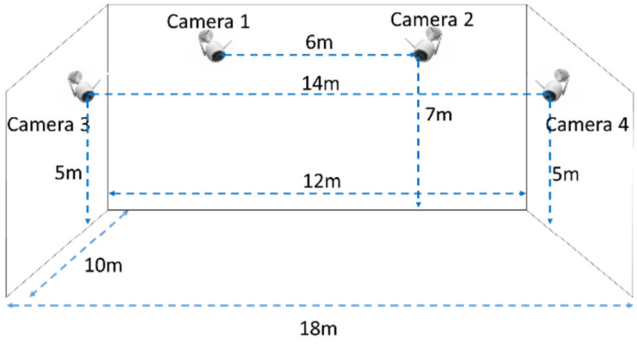

Fig. 2. The camera setup in the EPUAbN dataset.

Fig. 3. An example of an abnormal action captured from four viewpoints in the EPUAbN dataset.

Figure 3 shows an example of an abnormal action in the EPUAbN dataset. Compared to other available datasets for abnormal behavior detection, target objects in this dataset are captured at different viewpoints by different cameras at the same time.

C. Evaluation Criteria

Four metrics, micro AUC, macro AUC [23], micro accuracy, and macro accuracy [25, 26] are used for experimental evaluation. The final output of the proposed system (p score) is taken from the softmax layer and is used to obtain the final label that could belong to normal or abnormal action. An ROC curve is computed based on the True Positive Rate (TPR) and False Positive Rate (FPR) that is obtained at $\alpha = p$ changing from 0.1 to 1.

For the micro AUC score, at each α value, test videos $\{V_i, i = (1, \ldots, N)\}$ are input into the end-to-end trained model. The final labels of frames in V_i are combined with the ground

truth of V_i at frame level to obtain True Positive (TP_i) and True Negative (TN_i) predictions, along with False Positive (FP_i) and False Negative (FN_i) predictions. The ROC curve has two axes, TPR and FPR, at a certain $\alpha = \{\alpha_j,\ j = (1, \dots, M = 10)\} = (0.1, 0.2, \dots, 1)$, as:

$$TPR_{micro}^{\alpha_j} = \frac{\sum_{i=1}^{i=N} TP_i^{\alpha_j}}{\sum_{i=1}^{i=N} TP_i^{\alpha_j} + \sum_{i=1}^{i=N} FN_i^{\alpha_j}} \quad (7)$$

$$FPR_{micro}^{\alpha_j} = \frac{\sum_{i=1}^{i=N} FP_i^{\alpha_j}}{\sum_{i=1}^{i=N} TN_i^{\alpha_j} + \sum_{i=1}^{i=N} FP_i^{\alpha_j}} \quad (8)$$

Micro AUC is then calculated as:

$$AUC_{micro} = \int_{\alpha_j=0.1}^{1} f\left(TPR_{micro}^{\alpha_j}, FPR_{micro}^{\alpha_j}\right) \quad (9)$$

For macro AUC, a test video $\{V_i,\ i = (1, \dots, N)\}$ provides TPR_i and FPR_i at an α_i value:

$$TPR_i^{\alpha_j} = \frac{TP_i^{\alpha_j}}{TP_i^{\alpha_j} + FN_i^{\alpha_j}} \quad (10)$$

$$FPR_i^{\alpha_j} = \frac{FP_i^{\alpha_j}}{TN_i^{\alpha_j} + FP_i^{\alpha_j}} \quad (11)$$

where $TPR_i^{\alpha_j}$ and $FPR_i^{\alpha_j}$ are calculated on the entire videos of the test set. Then, the TPR and FPR axes of the ROC curve at a certain α_j are defined as:

$$TPR_{macro}^{\alpha_j} = \frac{\sum_{i=1}^{i=N} TPR_i^{\alpha_j}}{N} \quad (12)$$

$$FPR_{macro}^{\alpha_j} = \frac{\sum_{i=1}^{i=N} FPP_i^{\alpha_j}}{N} \quad (13)$$

The macro AUC value is then calculated as:

$$AUC_{macro} = \int_{\alpha_j=0.1}^{1} f\left(TPR_{macro}^{\alpha_j}, FPR_{macro}^{\alpha_j}\right) \quad (14)$$

For accuracy metrics, based on the ROC curve, $\alpha = 0.5$ is chosen, which obtains a balance between the TPR and FPR. This value is then applied to evaluate the accuracy score. In this study, the accuracy score is also computed using both micro and macro criteria:

$$Acc_{micro} = \frac{\sum_{i=1}^{i=N} (TP_i + TN_i)}{\sum_{i=1}^{i=N} (TP_i + TN_i + FP_i + FN_i)} \quad (15)$$

$$Acc_{macro}^{i} = \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \quad (16)$$

$$Acc_{macro} = \frac{\sum_{i=1}^{i=N} Acc_{macro}^{i}}{N} \quad (17)$$

IV. EXPERIMENTAL RESULTS

The proposed method was evaluated on challenging benchmark datasets, such as UMN [11], Crow-11 [12], UBNormal [22], and UCSD Ped2 [21]. UMN and Crow-11 are crowded scene datasets, unlike the others. In addition, a new dataset, called EPUAbN, was built for experimental evaluation of the abnormal behavior detection process. The experiments were implemented using two strategies: single-dataset and cross-dataset evaluation. In the first case, training and testing data were separated from the original dataset. In the other case, an entire dataset was used for training the model and another was used for testing. The proposed frameworks of ROHAC V2 and ROHAC-KD V2 were evaluated using both strategies.

A. Single-Dataset Evaluation

The Single Dataset Evaluation (SDE) used AUC [22] and Accuracy [25, 26] on several benchmark datasets. Figure. 4 shows micro and macro AUC scores to compare the results of the proposed ROHAC V2 and ROHAC-KD V2 with SOTA methods. The proposed methods achieved significantly better results, demonstrating their efficiency in detecting subtle patterns.

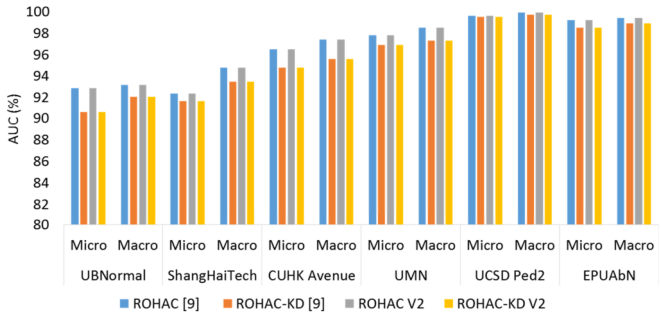

Fig. 4. Micro AUC (%) and macro AUC (%) results on SDE.

Figure 5 shows micro and macro accuracy values using SDE. These results are slightly higher than the previous methods [9] for detecting abnormal behaviors in benchmark datasets.

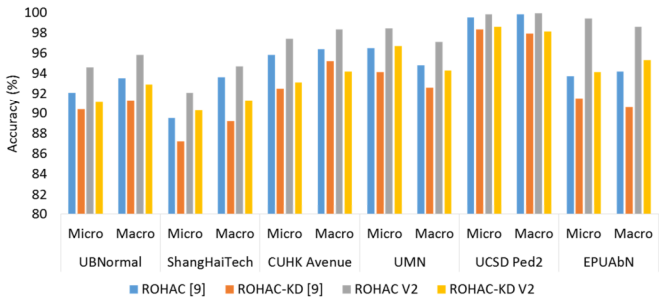

Fig. 5. Micro and macro accuracy (%) results on SDE.

B. Cross-Dataset Evaluation

In these evaluations, one dataset was used for training and another for testing. The results of the proposed methods (ROHAC V2 and ROHAC-KD V2) were compared with the SOTA method in [27], ROHAC, and ROHAC-KD [12] using micro and macro AUC scores. The experimental settings were the same as in [27]. The experiments were carried out five times and then averaged to obtain the final results, as shown in Tables II, III, and IV.

TABLE II. MICRO AUC AND MACRO AUC (%) RESULTS ON CROSS-DATASET EVALUATION ON CUHK AVENUE

Method	CUHK Avenue [7]		ShanghaiTech [24]		UCSD Ped2 [22]	
	Micro AUC	Macro AUC	Micro AUC	Macro AUC	Micro AUC	Macro AUC
[27]	92.3	90.4	83.6	81	-	-
ROHAC	93.7	94.8	93.8	94.5	92.5	95.2
ROHAC-KD	91.8	92.6	92.1	93.9	91.1	93.8
ROHAC V2	96.1	95.7	95.4	97.2	95.3	96.7
ROHAC-KD V2	94.8	95.6	92.9	94.3	92.6	94.5

The experimental results of ROHAC and ROHAC-KD methods for single-dataset evaluation on CUHK Avenue are from 2.5% to 4.2% micro AUC and from 5.2% to 7% macro AUC higher than the method in [27]. However, for cross-dataset evaluation with the ShangHaiTech dataset for training and CUHK Avenue for testing, the results of ROHAC for micro and macro AUC were 10.2% and 13.5% higher than those of [27], respectively. Similar results were achieved with ROHAC-KD compared to [27], with an increase of 8.5% and 12.9% for micro and macro AUC scores, respectively.

Cross-dataset evaluation was also performed using the UCSD Ped2 dataset for training and CUHK Avenue for testing, which was not performed in [27]. For SDE on the ShangHaiTech dataset, the model in [27] had 82.7% micro AUC and 89.3% macro AUC, which are much higher than in the case of cross-dataset evaluation with CUHK Avenue for training and ShangHaiTech for testing, which showed 76.3% for micro AUC and 86.3% for macro AUC. In this case, the ROHAC and ROAHC-KD methods achieved 91.9% and 89.6% micro AUC, and 90.1% and 88.2% macro AUC, respectively.

TABLE III. MICRO AUC AND MACRO AUC (%) RESULTS ON CROSS-DATASET EVALUATION ON SHANGHAITECH

Method	CUHK Avenue [7]		ShanghaiTech [24]		UCSD Ped2 [22]	
	Micro AUC	Macro AUC	Micro AUC	Macro AUC	Micro AUC	Macro AUC
[27]	82.7	89.3	76.3	86.3	-	-
ROHAC	92.4	94.8	91.9	90.1	92.3	89.9
ROHAC-KD	91.7	93.5	89.6	88.2	87.6	90.1
ROHAC V2	95.1	96.2	93.2	92.6	95.1	94.7
ROHAC-KD V2	93.4	94.6	91.7	90.3	90.2	91.9

TABLE IV. MICRO AUC (%) AND MACRO AUC (%) RESULTS ON CROSS DATASET EVALUATION ON UCSD PED2

Method	CUHK Avenue [7]		ShanghaiTech [24]		UCSD Ped2 [22]	
	Micro AUC	Macro AUC	Micro AUC	Macro AUC	Micro AUC	Macro AUC
[27]	98.7	99.7	87	97.2	90.6	95.7
ROHAC	99.6	99.9	94.8	97.5	95.8	97.8
ROHAC-KD	99.5	99.7	91.5	94.4	94.7	95.9
ROHAC V2	99.6	99.9	97.2	99.1	97.9	98.7
ROHAC-KD V2	99.5	99.7	93.4	97.6	95.4	97.1

Table IV presents the cross-dataset evaluation on UCSD Ped2 using ShangHaiTech and CUHK Avenue for training. In general, the results of the proposed frameworks were better than those of [27]. The decrease in micro AUC accuracy of the method in [27] is significant from SDE to cross-dataset evaluation, with 11.7% and 8.1% for CUHK Avenue-UCSD Ped2 and ShanghaiTech-UCSD Ped2 scenarios, respectively. However, the decreases of ROHAC and ROHAC-KD were only 4.8% and 3.8%, respectively.

In summary, the experimental results of the cross-dataset evaluation show the superior performance of the proposed frameworks compared to the method in [27]. Although it is a very challenging problem, cross-dataset evaluation should be implemented to estimate the suitability of datasets for diverse-context assessments in practice, since each available dataset for human abnormal behavior detection is set for a certain evaluation context. This work, based on experiments in different cross-dataset evaluation settings, demonstrates the effectiveness of the proposed frameworks in exploiting data features from available datasets for human abnormal behavior detection. The proposed frameworks are effective not only for single-dataset evaluation but also for cross-dataset evaluation, showing high prospects for practical implementations.

V. CONCLUSION

This work presents an efficient framework for the detection of human abnormal behaviors by exploring three input modalities: RGB, optical flow, and heatmap. Attention units are used to exploit the important information from these three input modalities. The experimental results in single- and cross-dataset evaluation scenarios show the outperformance of the proposed methods compared to others in the accuracy of human abnormal behavior detection. In addition, using KD can reduce the computational time required to detect abnormal behaviors. The experimental results show that with KD, the proposed framework significantly reduces the processing time, but still has a high detection accuracy. Future work will consider the deployment of multimodal learning of RGB and depth images for the proposed system and evaluate the models on more diverse datasets with abnormal behaviors. Moreover, scaled-down models with KD will be expanded to include multiple learning models for several contexts from different abnormal behavior datasets. In addition, single-modality frameworks will be evaluated and compared with multimodal ones.

REFERENCES

[1] N. C. Tay, C. Tee, T. S. Ong, and P. S. Teh, "Abnormal Behavior Recognition using CNN-LSTM with Attention Mechanism," in *2019 1st International Conference on Electrical, Control and Instrumentation Engineering (ICECIE)*, Kuala Lumpur, Malaysia, Nov. 2019, pp. 1–5, https://doi.org/10.1109/ICECIE47765.2019.8974824.

[2] A. Gangwar, V. González-Castro, E. Alegre, and E. Fidalgo, "AttM-CNN: Attention and metric learning based CNN for pornography, age and Child Sexual Abuse (CSA) Detection in images," *Neurocomputing*, vol. 445, pp. 81–104, Jul. 2021, https://doi.org/10.1016/j.neucom.2021.02.056.

[3] P. Kuppusamy and C. Harika, "Human Action Recognition using CNN and LSTM-RNN with Attention Model," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 8, pp. 1639–1643, 2019.

[4] W. Ullah, A. Ullah, T. Hussain, Z. A. Khan, and S. W. Baik, "An Efficient Anomaly Recognition Framework Using an Attention Residual

LSTM in Surveillance Videos," *Sensors*, vol. 21, no. 8, Jan. 2021, Art. no. 2811, https://doi.org/10.3390/s21082811.

[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*, Portland, OR, USA, Jun. 2013, pp. 2547–2554, https://doi.org/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*, Miami, FL, Jun. 2009, pp. 935–942, https://doi.org/10.1109/CVPR.2009.5206641.

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

[8] X. Zheng, Y. Zhang, Y. Zheng, F. Luo, and X. Lu, "Abnormal event detection by a weakly supervised temporal attention network," *CAAI Transactions on Intelligence Technology*, vol. 7, no. 3, pp. 419–431, 2022, https://doi.org/10.1049/cit2.12068.

[9] S. Liu, X. Ma, H. Wu, and Y. Li, "An End to End Framework With Adaptive Spatio-Temporal Attention Module for Human Action Recognition," *IEEE Access*, vol. 8, pp. 47220–47231, 2020, https://doi.org/10.1109/ACCESS.2020.2979549.

[10] L. He, S. Wen, L. Wang, and F. Li, "Vehicle theft recognition from surveillance video based on spatiotemporal attention," *Applied Intelligence*, vol. 51, no. 4, pp. 2128–2143, Apr. 2021, https://doi.org/10.1007/s10489-020-01933-8.

[11] G. Yang *et al.*, "STA-TSN: Spatial-Temporal Attention Temporal Segment Network for action recognition in video," *PLOS ONE*, vol. 17, no. 3, 2022, Art. no. e0265115, https://doi.org/10.1371/journal.pone.0265115.

[12] A. D. Ho, H. G. Doan, and T. T. T. Pham, "Multi-Modality Abnormal Crowd Detection with Self-Attention and Knowledge Distillation," *Engineering, Technology & Applied Science Research*, vol. 14, no. 5, pp. 16674–16679, Oct. 2024, https://doi.org/10.48084/etasr.8194.

[13] Y. Liu, J. Yan, and W. Ouyang, "Quality Aware Network for Set to Set Recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 4694–4703, https://doi.org/10.1109/CVPR.2017.499.

[14] 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)*, Honolulu, HI, USA, Jul. 2017, pp. 2184–2191, https://doi.org/10.1109/CVPRW.2017.271.

[15] J. Shao, K. Kang, C. C. Loy, and X. Wang, "Deeply learned attributes for crowded scene understanding," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 4657–4666, https://doi.org/10.1109/CVPR.2015.7299097.

[16] J. Shao, C. C. Loy, and X. Wang, "Scene-Independent Group Profiling in Crowd," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, Jun. 2014, pp. 2227–2234, https://doi.org/10.1109/CVPR.2014.285.

[17] T. Hassner, Y. Itcher, and O. Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Jun. 2012, pp. 1–6, https://doi.org/10.1109/CVPRW.2012.6239348.

[18] 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, https://doi.org/10.1109/CVPR.2015.7298684.

[19] 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*, Tsukuba, Japan, Aug. 2012.

[20] T. Ellis, "Performance Metrics and Methods for Tracking in Surveillance," in *Proceedings 3rd IEEE International Workshop on PETS*, Copenhagen, Denmark, 2002.

[21] E. Bermejo Nievas, O. Deniz Suarez, G. Bueno García, and R. Sukthankar, "Violence Detection in Video Using Computer Vision Techniques," in *Computer Analysis of Images and Patterns*, 2011, pp. 332–339, https://doi.org/10.1007/978-3-642-23678-5_39.

[22] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian Detection in Crowded Scenes," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA, 2005, vol. 1, pp. 878–885, https://doi.org/10.1109/CVPR.2005.272.

[23] 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, https://doi.org/10.1109/CVPR52688.2022.01951.

[24] 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)*, Venice, Italy, Oct. 2017, pp. 341–349, https://doi.org/10.1109/ICCV.2017.45.

[25] Y. M. Bai, Y. Wang, and S. S. Wu, "Detection of Abnormal Human Behavior in Video Images based on a Hybrid Approach," *nternational Journal of Advanced Computer Science and Applications*, vol. 13, no. 11, pp. 346–356, 2022.

[26] H. Bagherinezhad and S. Y. Soltani, "Abnormal Human Behavior Detection System in Video Surveillance Systems." Social Science Research Network, May 11, 2022, https://doi.org/10.2139/ssrn.4106323.

[27] M. I. Georgescu, R. T. 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, Sep. 2022, https://doi.org/10.1109/TPAMI.2021.3074805.