# Multi-Modality Abnormal Crowd Detection with Self-Attention and Knowledge Distillation

**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

**Thi Thanh Thuy Pham**
Faculty of Information Security, Academy of People Security, Hanoi, Vietnam
thanh-thuy.pham@mica.edu.vn (corresponding author)

*Received: 20 June 2024 | Revised: 24 July 2024 | Accepted: 25 July 2024*

## ABSTRACT

**Deep Neural Networks (DNNs) have become a promising solution for detecting abnormal human behaviors. However, building an efficient DNN model in terms of both computational cost and classification accuracy is still a challenging problem. Furthermore, there are limited existing datasets for abnormal behavior detection, and each focuses on a certain context. Therefore, a DNN model trained on a certain dataset will be adaptive for a particular context and not suitable for others. This study proposes a DNN framework with efficient attention and Knowledge Distillation (KD) mechanisms. Attention units capture key information from multiple RGB, optical flow, and heatmap inputs. KD is applied to scale down model size. Experiments were performed on several benchmark datasets, examining both AUC and accuracy. The results show that the proposed framework outperformed other state-of-the-art methods in detection accuracy. Furthermore, the trade-off between detection performance and computational cost was also addressed by the proposed framework with KD.**

***Keywords-abnormal behavior detection; attention; knowledge distillation***

## I. INTRODUCTION

The proliferation of applications such as Human-Computer Interaction (HCI) [1], camera-based surveillance, and Virtual Reality (VR) has made the problem of detecting human abnormal behaviors increasingly attractive. However, this problem has also emerged as a challenging research domain for image/video understanding and analysis. Abnormal crowding is the behavior of many humans in a public area. It can be the reaction of the crowd when people face danger, fear, violence, robbery, road accidents, or panic, where they chaotically run, fight, chase each other, etc. Thus, the main challenges for human abnormal behavior detection are complex backgrounds, occlusions, low resolution (especially in the case of surveillance videos), variety in camera viewpoints, intra- or inter-class variations, and numerous abnormal behaviors that are defined depending on the different contexts (airport, stadium, supermarket, amusement parks, street, patient room, classroom, hospital, etc.) and applications (intelligent visual monitoring in hospitals or public areas, crime prevention, traffic monitoring, etc.).

Several solutions have been proposed for abnormal human behavior detection, based on two approaches: manual feature extraction with classical classifiers and auto-feature learning by Deep Neural Networks (DNNs). The handcrafted features of SIFT [2], Histogram of Oriented Gradient (HOG) [3], Histogram of Optical Flow (HOF) [4], and optical flow [5] are mainly exploited for the recognition of human abnormal actions. The SIFT feature is less dependent on light intensity, noise, rotation, and motion variation. The HOG contains the shape and appearance information of the targets. Optical flow and HOF are efficient in encoding human motion. Classical classifiers, such as Hidden Markov Models (HMM) [6], Gaussian Mixture Models (GMM) [7], Support Vector Machines (SVM) [8], Random Forest (RF), K-Nearest Neighbor (KNN) [9], and Decision Trees (DT) [10] are commonly utilized to identify human abnormalities from handcrafted features.

In DNN-based approaches, features can be extracted automatically from raw data. Some common DNN architectures for auto-feature learning and classification of human abnormalities are Convolutional Neural Networks

(CNNs) [11] and Recurrent Neural Networks (RNN) [12]. In [13], a CNN-RNN combination was proposed for abnormal human behavior detection, where spatial features were extracted by CNNs and temporal features were generated by RNNs [13]. Recently, transformer-based neural networks have been employed for human abnormal action recognition [14]. In general, a DNN with a large-scale architecture and trained on large enough datasets offers better detection accuracy than traditional classifiers. However, the complex architecture of DNNs with the high-dimensional data of visual tasks leads to high computational costs and is difficult to apply in practice. Moreover, building an appropriate deep learning model for the problem of abnormal human behavior detection is a challenging task, due to the dynamic nature and variations in real-world problems compared to the data used for training DNNs. The existing datasets for human abnormal behavior detection are limited, and each of them focuses on a certain context. Therefore, a DNN trained on a specific dataset will be adaptive for a particular context and not suitable for others.

An attention mechanism was recently proposed for the detection of abnormal human behavior, along with DNNs, to reduce computational costs and memory usage while improving accuracy and robustness. Attention mechanisms can be categorized into four main types: channel attention, spatial attention, temporal attention, and branch attention. In addition, combinations of channel and spatial attention, or spatial and temporal attention, have also been proposed. In an attention mechanism, attention masks are generated across the channel domains (channel attention), spatial domains (spatial attention), time domain (temporal attention), and in different branches (branch attention) to select important channels, spatial regions or locations, keyframes, and branches, respectively. Studies on abnormal human behavior detection by DNNs with attention units focus mainly on spatial attention [15], temporal attention [16], or a combination of them [17].

This study proposes an end-to-end framework for detecting abnormal human behaviors. Attention units are deployed to help the model focus on important information from the input image sequences. The proposed approach is different from other common approaches that utilize RGB and optical flow inputs for DNN models, as it exploits three types of input image sequences: RGB, optical flow, and heatmap. Two consecutive RGB frames are used to calculate an optical flow, and then a heat map image is generated from the optical flow. RGB, optical flow, and heatmap attention units are then used to extract the key spatial and temporal information. Experimental results on standard datasets showed that the proposed framework outperformed other state-of-the-art methods. In addition, a Knowledge Distillation (KD) solution was implemented to reduce computation costs while maintaining the high recognition accuracy of abnormal behaviors. This will be helpful for practical and real-time implementations.

## II. PROPOSED METHOD

An end-to-end framework, named ROHAC, was implemented with spatial and temporal attention units on the inputs of three image types, RGB, optical flow, and heatmap, for abnormal human behavior detection. These input images are prepared by a data preprocessing step. Then, they are input into CNN networks for feature extraction. The important information from the feature vectors extracted by CNNs is then selected by attention layers to provide the final attention feature vectors for classification. In addition, KD is used to reduce the computational cost for abnormal human behavior detection.

### A. *Data Prepossessing*

RGB video frames are commonly used in visual analysis in general and in abnormal human behavior detection in particular. This study explores two consecutive RGB frames ($F_t$ and $F_{t+1}$) from input video at a time. At each considered frame, person detection is performed using the YOLO v5 model to obtain the human bounding boxes of $B^t_{Si}$ (red boxes in Figure 1) and $B^{t+1}_{Si}$ (cyan boxes in Figure 1). The final bounding box of person *si* is calculated from these two frames using:

$$B^{rgb}_{Si} = B^t_{Si} \cup B^{t+1}_{Si}, \quad i = (1, \ldots, N) \qquad (1)$$

where $N$ is the number of people in a frame. Instead of calculating time-consuming optical flow over the entire pixels of the image, optical flows (middle image at below part of Figure 1) are computed on axes of bounding boxes $I^{rgb}_{si}$ from two consecutive RGB frames ($F_t$ and $F_{t+1}$) by RAFT networks [18]. The optical flow images $I^{si}_{OF}$ of humans are then used to transfer to heat map images $I^{HM}_{si}$. Then, three modality streams (RGB stream, optical flow stream, and heat map stream) are exploited.

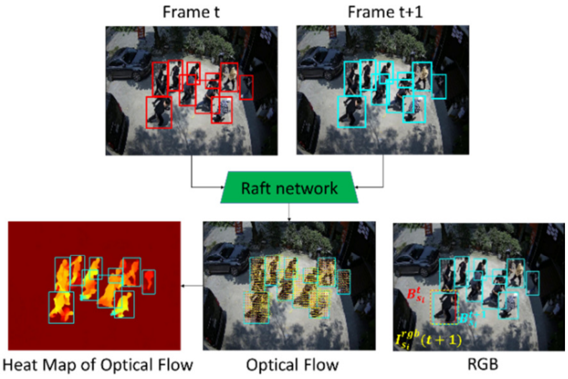

Fig. 1. Preprocessing for optical flow and jetmap image.

### B. *Abnormal Detection with Spatial-Temporal Attention Units*

Figure 2 shows the proposed framework for abnormal behavior detection. It embeds Resnet50 networks and attention units for extracting key spatial and temporal features from RGB, optical flow, and heatmap input images. Given three image streams, RGB, optical flow, and heatmap, each image consists of $N$ subjects. Thus, each image provides $N$ blobs that are bounding boxes of the subjects in the image, such as $(I^{rgb}_{s1}, \ldots, I^{rgb}_{sN})$, $(I^{OF}_{s1}, \ldots, I^{OF}_{sN})$, and $(I^{HM}_{s1}, \ldots, I^{HM}_{sN})$ taken from the RGB, optical flow, and heatmap streams, respectively. Then, these human bounding boxes are utilized as inputs for the Resnet50_RGB, the Resnet50_OF, and the Resnet50_HM models. The outputs are feature vectors taken at FC layers as $(F^{rgb}_{s1}, \ldots, F^{rgb}_{sN})$ for RGB inputs, $(F^{OF}_{s1}, \ldots, F^{OF}_{sN})$ for optical flow inputs, and $(F^{HM}_{s1}, \ldots, F^{HM}_{sN})$ for heatmap inputs. These feature vectors belong to $R^{1 \times M}$, $M$ =2048.

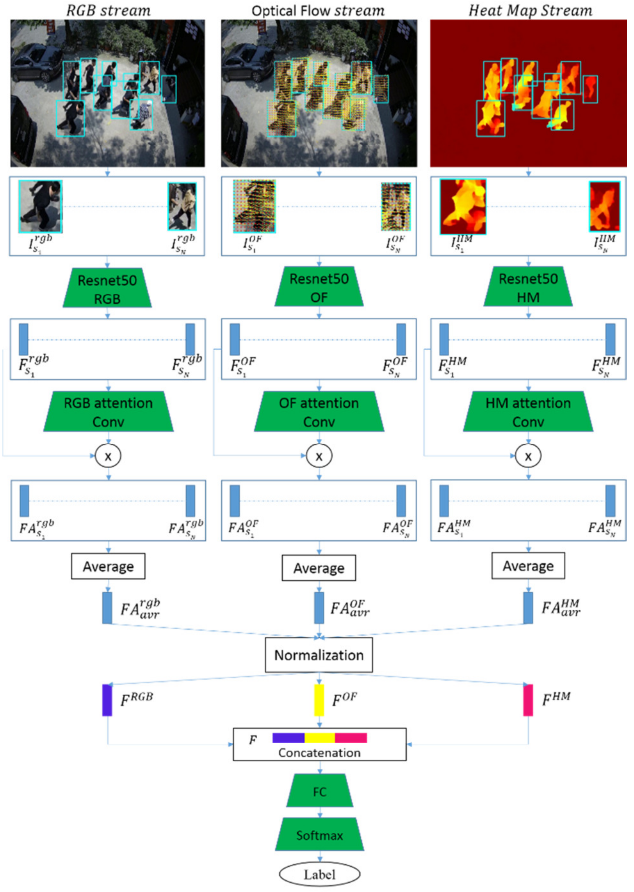

Fig. 2. The proposed abnormal behavior detection framework

In the next step, three attention convs (2, N, 1), RGB attention conv, OF attention conv, and HM attention conv are applied for image-level features to generate attention scores for each stream. The attention convs utilize $F_{si}^{rgb}$,$F_{si}^{OF}$, and $F_{si}^{HM}$ as inputs and output the attention scores $a_{si}^{j} = \{a_{si}^{rgb}, a_{si}^{OF}, a_{si}^{HM}\}$ for the RGB, optical flow, and heatmap streams, respectively. The attention scores are calculated using sigmoid and L1 normalization functions [19] as:

$$a_{s_i}^{j} = \frac{\sigma^{x_i^j}}{\sum_{n=1}^{M} \sigma^{x_i^j}} = \frac{\frac{1}{1-e^{x_i^j}}}{\sum_{n=1}^{M} \frac{1}{1-e^{x_i^j}}} \qquad (2)$$

The attention scores of each stream are then used to compute the context vectors $FA_{si}^{j} = \{FA_{si}^{rgb}, FA_{si}^{OF}, FA_{si}^{HM}\}$, $j = (1,\ldots,3)$ as shown in (3). Each of these vectors is a weighted sum of the value vectors $F_{si}^{j}$ =$\{F_{si}^{rgb}, F_{si}^{OF}, F_{si}^{HM}\}$, $j = (1, \ldots, 3)$, respectively.

$$FA_{s_i}^{j} = a_{s_i}^{j} * F_{s_i}^{j}, i = (1, \ldots, N) \qquad (3)$$

Then, the context vectors of each data type are averaged:

$$FA_{avr}^{j} \in R^{1xM} = \frac{1}{N} \sum_{i=1}^{N} FA_{s_i}^{j} \qquad (4)$$

The three average feature vectors of RGB, optical flow, and heatmap streams, $FA_{j}^{avr}$, $j = (1, \ldots, 3)$, are normalized to $F_j \in R^{1xM}$. They are then concatenated into a feature vector $F \in \Re^{1x(3M)}$. The final feature vector $F$ is fully connected and passes 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 $\overline{p_i}$ and ground truth values $p_i$, $i = (1, \ldots, K)$, the loss function is calculated by:

$$L_{softmax} = \frac{1}{K} \sum_{i=1}^{K} p_i log \overline{p_i} \qquad (5)$$

### C. Knowledge Distillation (KD) Framework for Abnormal Behavior Detection

This study uses KD, as shown in Figure 3, to compress the CNN models used in the above-mentioned ROHAC framework. Scaling down the CNN model size using KD aims at reducing computational costs while maintaining the model's performance in abnormal behavior detection. The proposed framework includes two main parts. The first part aims to compress from the YOLO v5 model to the YOLO v3-tiny model (pink boxes in Figure 3). The second part compresses the Resnet50 model to the Resnet18 model (blue boxes in Figure 3).

In the first KD component, the ground truth is the human bounding box in the dataset $y$, the predicted label of YOLO v5 is $\widehat{y^t}$, and the predicted label of YOLO v3-tiny is $\widehat{y^s}$. The loss function of the KD YOLO model is computed by:

$$L^{(1)}\left(y, \widehat{y^s}, \widehat{y^t}\right) = (1 - \lambda) L_{YOLOV3}^{(1)}(y, \widehat{y^s}) + \lambda L_{KD}^{(1)}(\widehat{y^t}, \widehat{y^s})$$

$$= -(1 - \lambda) \sum_{i=1}^{C} y_i log \widehat{y_i^s} - \lambda T^2 \sum_{i=1}^{C} \frac{\widehat{y_i^t}}{T} log \frac{\widehat{y_i^s}}{T} \qquad (7)$$

In the second KD component, the ground truth is an abnormal label in the abnormal action dataset $z$, the predicted abnormal label by the Resnet50 model is $\widehat{z^t}$, and the abnormal predicted label by the Resnet18 model is $\widehat{z^s}$ . The loss function of the KD Resnet model is computed as:

$$L^{(2)}\left(z, \widehat{z^s}, \widehat{z^t}\right) = (1 - \lambda) L_{Resnet18}^{(2)}(z, \widehat{z^s}) + \lambda L_{KD}^{(2)}(\widehat{z^t}, \widehat{z^s})$$

$$= -(1 - \lambda) \sum_{i=1}^{C} z_i log \widehat{z_i^s} - \lambda T^2 \sum_{i=1}^{C} \frac{\widehat{z_i^t}}{T} log \frac{\widehat{z_i^s}}{T} \qquad (8)$$

where $T$ is softmax temperature. Prediction distribution is softened by the distillation temperature $T$. This study chose $T = 6$. The efficiency of the proposed KD framework is evaluated on both time cost and performance.

## III. EXPERIMENTS AND RESULTS

The proposed method was evaluated on challenging benchmark datasets: UMN [20], Crow-11 [21], UCF CC 50 [22], UBNormal [23], and UCSD [24]. Four metrics, micro- and macro AUC [23] and micro- and macro accuracy [25] were used for experimental evaluations. The final output of the proposed system ($p$ score) was taken from the softmax layer. This $p$ score is changed to obtain the final label that could belong to a normal or abnormal action. Thus, a ROC curve is computed based on the True Positive Rate (TPR) and the False Positive Rate (FPR) at $\alpha = p$ changing from 0.1 to 1.

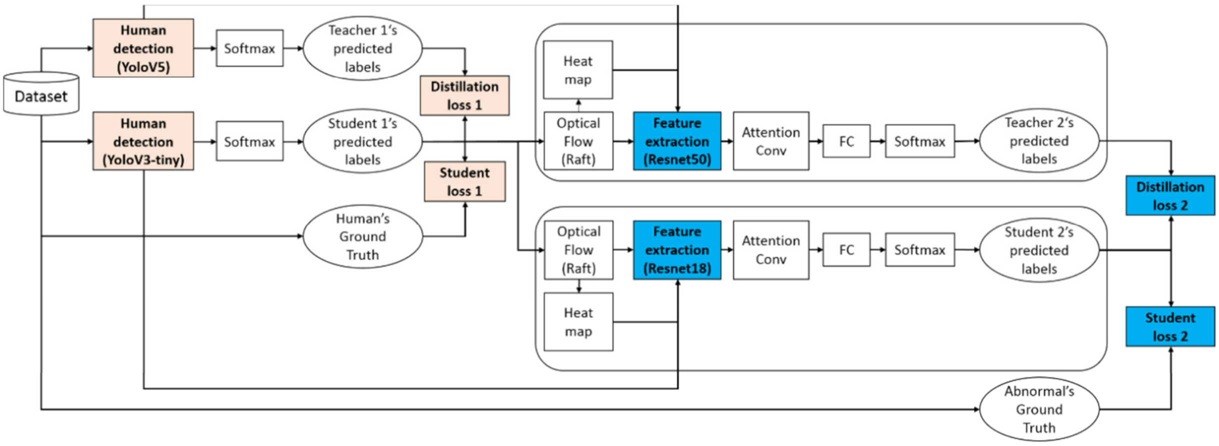

Fig. 3. The knowledge distillation framework for anomaly behavior detection.

TABLE I. MICRO AND MACRO AUC (%) OF ROHAC AND ROHAC-KD IN COMPARISON WITH OTHER METHODS

| Method | UBNormal | | ShangHaiTech | | CUHK Avenue | | UMN | | UCSD Ped2 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | AUC(%) | | | | | | | | | |
| | Micro | Macro | Micro | Macro | Micro | Macro | Micro | Macro | Micro | Macro |
| [20] | - | - | - | - | - | - | 96 | - | - | - |
| [23] | 59.3 | 84.9 | 82.7 | 89.3 | 92.3 | 90.4 | 85.5 | 94.4 | 98.7 | 99.7 |
| [26] | 74.8 | - | - | - | 90.2 | - | - | - | 97.3 | - |
| [27] | 56 | 85.9 | 83.8 | 90.5 | 91.6 | 92.5 | - | - | - | - |
| [28] | 61.3 | 85.6 | 83.7 | 90.5 | 93 | 93.2 | - | - | 91.5 | 90.6 |
| [29] | - | - | 68 | - | 81.7 | - | - | - | 92.2 | - |
| [30] | - | - | 74.5 | 82.9 | 87.3 | 84.5 | - | - | - | - |
| [31] | - | - | 75.5 | 83.7 | 90.9 | 92.2 | - | - | - | - |
| [32] | - | - | 78.7 | 84.9 | 87.4 | 90.4 | - | - | - | - |
| [33] | - | - | 72.1 | - | 83.3 | - | - | - | - | - |
| [34] | - | - | - | - | 87.2 | - | - | - | 94 | - |
| [35] | - | - | 70.5 | - | 88.5 | - | - | - | - | - |
| [36] | - | - | - | 76.2 | - | 92.1 | - | - | - | 99.3 |
| ROHAC | 92.9 | 93.2 | 92.4 | 94.8 | 96.5 | 97.4 | 97.8 | 98.5 | 99.6 | 99.9 |
| ROHAC KD | 90.6 | 92.1 | 91.7 | 93.5 | 94.8 | 95.6 | 96.9 | 97.3 | 99.5 | 99.7 |

The experiments were implemented for AUC accuracy. At first, training and test data were separated. The proposed frameworks of ROHAC and ROHAC with KD, named ROHAC-KD, were evaluated on the benchmark datasets.

## *A. AUC Score Evaluation*

For AUC evaluation, micro and macro AUC scores were calculated for various benchmark datasets, such as UBNormal, ShangHaiTech, CUHK Avenue, UMN, and UCSD Ped2. Table I shows the comparative results of the proposed with other methods. For the UBNormal dataset, the method in [26] achieved a micro AUC of 74.8%, which is much lower those achieved by ROHAC (92.9%) and ROHAC-KD (90.6%). The macro AUC of ROHAC and ROHAC-KD was also better than that of [28], with 93.2% and 92.1% compared to 85.9%, respectively [28]. The experimental results on the ShangHaiTech dataset also show the higher performance of the proposed methods. The method in [27] had micro and macro AUC scores of 83.8% and 90.5%, respectively, while ROHAC and ROHAC-KD achieved 92.4% and 94.8% for micro AUC and 91.7% and 93.5% for macro AUC. In evaluations on the CUHK Avenue, UMN, and UCSD Ped2 datasets, ROHAC and ROHAC-KD achieved also the highest results, which were about 1-3.5% and 0.2-4.2% higher than those of the other best methods at micro and macro AUC, respectively. The experimental results in Table I show that the same method on different datasets gave significantly different results. For example, the evaluation of the method in [28] had micro and macro AUC of only 61.3% and 85.6%, respectively, on the UBNormal dataset. However, the relative results for the CUHK Avenue, UMN, and UCSD Ped2 datasets were all above 90%. The same happens with the methods in [28] and [27]. However, considering the proposed ROHAC and ROHAC-KD methods, the micro and macro AUC values are almost steady above 90% for all evaluation datasets. This demonstrates the efficiency of the proposed frameworks in picking up subtle patterns in datasets.

## *B. Accuracy Score Evaluation*

Accuracy evaluation metrics were also examined on the same datasets. Figure 4 shows the promising results in both micro and macro accuracy metrics. The experimental results of the ROHAC method on the UBNormal, ShangHaiTech, CUHK Avenue, UMN, and UCSD Ped2 datasets were 92.1%, 89.5%, 95.8%, 96.5%, and 99.5% for micro accuracy, and 93.5%, 93.6%, 96.4%, 94.8%, and 99.8% for macro accuracy, respectively. These results are slightly higher than those of the ROHAC-KD method on all datasets.

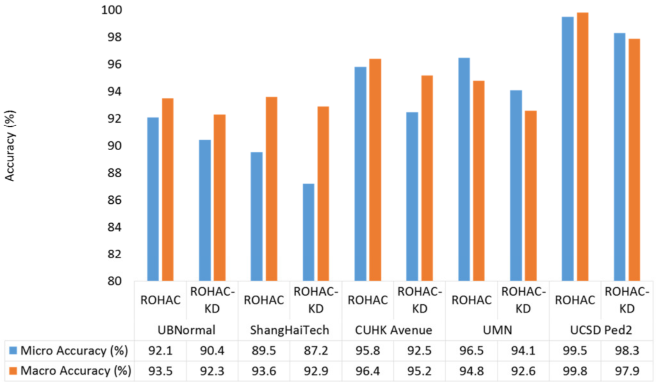

Fig. 4. Micro and macro accuracy (%) of ROHAC and ROHAC-KD on different benchmark datasets.

### C. *Computational Cost Evaluation*

Table II shows the computational costs for the phases of human detection, optical flow calculation, and classification, and the total time for abnormal behavior detection by ROHAC and ROHAC-KD. The time for running experiments was evaluated on Jeston Xavier NX and Tesla T4.

TABLE II. TIME COST FOR HUMAN DETECTION, OPTICAL FLOW, ABNORMAL CLASSIFICATION, AND TOTAL.

| Method | Human detection (ms) | Optical flow + classification (ms) | Frames per second (fps) |
|---|---|---|---|
| **Run on Jeston Xavier NX** | | | |
| ROHAC | 0.325 | 0.101 | 2.3 |
| ROHAC-KD | 0.028 | 0.014 | 23.8 |
| **Run on Tesla T4** | | | |
| ROHAC | 0.048 | 0.024 | 13.7 |
| ROHAC-KD | 0.016 | 0.008 | 41.7 |

For human detection on Jeston Xavier NX, the computational time was 0.325 ms for ROHAC and 0.028 ms for ROHAC-KD. The respective times on Tesla T4 were 0.048 ms and 0.016 ms. On Jeston Xavier NX, the computational costs for optical flow calculation and classification were 0.101 ms and 0.014 ms for ROHAC and ROHAC-KD, whereas on Tesla T4 were 0.024 ms and 0.008 ms, respectively. When running on Jeston Xavier NX, the total fps processed by ROHAC-KD was about ten times higher than the ROHAC method. Running on a Tesla T4, ROHAC-KD achieved 41.7 fps, whereas ROHAC achieved only 13.7 fps.

The results in Table II demonstrate the effectiveness of using KD in the proposed framework for abnormal human behavior detection, as it greatly reduces computational costs while almost preserving recognition accuracy. This study also compared the computational costs of the proposed with other methods when running on Jeston Xavier NX, and the results are shown in Table III. ROHAC can process 13.7 fps, which is only 2.9 fps higher than the lowest value of the method in [37]. However, compared to other methods, ROHAC achieved a lower number of fps. ROHAC-KD achieved 41.7 fps, which is higher than other methods, except for MemAE [34] (42 fps) and MNAD [36] (56 fps). However, the recognition performance of both ROHAC and ROHAC-KD was higher than that of these methods.

TABLE III. COMPARISON OF COMPUTATIONAL COST BETWEEN THE PROPOSED AND OTHER METHODS

| No | Method | fps |
|---|---|---|
| 1 | [34] | 42 |
| 2 | [36] | 56 |
| 3 | [37] | 10.8 |
| 4 | [29] | 18 |
| 5 | SSMTL++v1 [28] | 20.2 |
| 6 | SSMTL++v2 [28] | 18.8 |
| 7 | ROHAC | 13.7 |
| 8 | ROHAC-KD | 41.7 |

## IV. CONCLUSION

This study proposed an efficient framework for abnormal human behavior detection by exploring three input frames: RGB, optical flow, and heatmap. Attention units were used to exploit the important information from these three input images. The experimental results on both micro and macro accuracy and AUC metrics on multiple datasets showed that the proposed method outperformed other state-of-the-art methods. In addition, KD was used to reduce the computational cost for abnormal behavior detection. The experimental results showed that with KD, the proposed framework reduced significantly the processing time but still had high detection accuracy compared to without using KD. Future work will consider deploying multimodal learning of RGB and depth images in the proposed system and the evaluation of the models on more diverse datasets of abnormal behaviors. Moreover, the scale-down models using KD will be expanded to include multiple teacher models for several contexts from different abnormal behavior datasets.

## REFERENCES

[1] H. G. Doan and N. T. Nguyen, "Fusion Machine Learning Strategies for Multi-modal Sensor-based Hand Gesture Recognition," *Engineering, Technology & Applied Science Research*, vol. 12, no. 3, pp. 8628–8633, Jun. 2022, https://doi.org/10.48084/etasr.4913.

[2] I. P. Febin, K. Jayasree, and P. T. Joy, "Violence detection in videos for an intelligent surveillance system using MoBSIFT and movement filtering algorithm," *Pattern Analysis and Applications*, vol. 23, no. 2, pp. 611–623, May 2020, https://doi.org/10.1007/s10044-019-00821-3.

[3] S. P. Sahoo and S. Ari, "On an algorithm for human action recognition," *Expert Systems with Applications*, vol. 115, pp. 524–534, Jan. 2019, https://doi.org/10.1016/j.eswa.2018.08.014.

[4] H. Lin, J. D. Deng, B. J. Woodford, and A. Shahi, "Online Weighted Clustering for Real-time Abnormal Event Detection in Video Surveillance," in *Proceedings of the 24th ACM International Conference on Multimedia*, Amsterdam, Netherlands, Jul. 2016, pp. 536–540, https://doi.org/10.1145/2964284.2967279.

[5] X. Zhang, S. Yang, J. Zhang, and W. Zhang, "Video anomaly detection and localization using motion-field shape description and homogeneity testing," *Pattern Recognition*, vol. 105, Sep. 2020, Art. no. 107394, https://doi.org/10.1016/j.patcog.2020.107394.

[6] V. G. Sánchez, O. M. Lysaker, and N.-O. Skeie, "Human behaviour modelling for welfare technology using hidden Markov models," *Pattern Recognition Letters*, vol. 137, pp. 71–79, Sep. 2020, https://doi.org/10.1016/j.patrec.2019.09.022.

[7] T. Huang, Q. Han, W. Min, X. Li, Y. Yu, and Y. Zhang, "Loitering Detection Based on Pedestrian Activity Area Classification," *Applied Sciences*, vol. 9, no. 9, Jan. 2019, Art. no. 1866, https://doi.org/10.3390/app9091866.

[8] D. Gao and H. Yu, "The use of optimised SVM method in human abnormal behaviour detection," *International Journal of Grid and Utility*

*Computing*, vol. 13, no. 2–3, pp. 164–172, Jan. 2022, https://doi.org/10.1504/IJGUC.2022.124408.

[9] S. Samudra, M. Barbosh, and A. Sadhu, "Machine Learning-Assisted Improved Anomaly Detection for Structural Health Monitoring," *Sensors*, vol. 23, no. 7, Jan. 2023, Art. no. 3365, https://doi.org/10.3390/s23073365.

[10] V. G. Sánchez and N.-O. Skeie, "Decision Trees for Human Activity Recognition in Smart House Environments," in *The 59th Conference on Imulation and Modelling (SIMS 59)*, Oslo, Norway, Sep. 2018, pp. 222–229, https://doi.org/10.3384/ecp18153222.

[11] P. Kuppusamy and V. C. Bharathi, "Human abnormal behavior detection using CNNs in crowded and uncrowded surveillance – A survey," *Measurement: Sensors*, vol. 24, Dec. 2022, Art. no. 100510, https://doi.org/10.1016/j.measen.2022.100510.

[12] M. Zerkouk and B. Chikhaoui, "Long Short Term Memory Based Model for Abnormal Behavior Prediction in Elderly Persons," in *How AI Impacts Urban Living and Public Health*, New York, NY, USA, 2019, pp. 36–45, https://doi.org/10.1007/978-3-030-32785-9_4.

[13] C. W. Chang, C. Y. Chang, and Y. Y. Lin, "A hybrid CNN and LSTM-based deep learning model for abnormal behavior detection," *Multimedia Tools and Applications*, vol. 81, no. 9, pp. 11825–11843, Apr. 2022, https://doi.org/10.1007/s11042-021-11887-9.

[14] 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, https://doi.org/10.1109/ACCESS.2023.3266440.

[15] 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.

[16] 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.

[17] 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.

[18] Z. Teed and J. Deng, "RAFT: Recurrent All-Pairs Field Transforms for Optical Flow," in *Computer Vision – ECCV 2020*, Glasgow, UK, 2020, pp. 402–419, https://doi.org/10.1007/978-3-030-58536-5_24.

[19] 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.

[20] 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, USA, Jun. 2009, pp. 935–942, https://doi.org/10.1109/CVPR.2009.5206641.

[21] C. Dupont, L. Tobías, 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.

[22] 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.

[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] 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, Jun. 2005, vol. 1, pp. 878–885 vol. 1, https://doi.org/10.1109/CVPR.2005.272.

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

[26] G. Yu *et al.*, "Cloze Test Helps: Effective Video Anomaly Detection via Learning to Complete Video Events," in *Proceedings of the 28th ACM International Conference on Multimedia*, Seattle, WA, USA, Oct. 2020, pp. 583–591, https://doi.org/10.1145/3394171.3413973.

[27] A. Barbalau *et al.*, "SSMTL++: Revisiting self-supervised multi-task learning for video anomaly detection," *Computer Vision and Image Understanding*, vol. 229, Mar. 2023, Art. no. 103656, https://doi.org/10.1016/j.cviu.2023.103656.

[28] 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.

[29] 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, https://doi.org/10.1109/ICCV.2017.45.

[30] 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*, Salt Lake City, UT, USA, Dec. 2018, pp. 6536–6545, https://doi.org/10.1109/CVPR.2018.00684.

[31] N. C. Ristea *et al.*, "Self-Supervised Predictive Convolutional Attentive Block for Anomaly Detection," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 13566–13576, https://doi.org/10.1109/CVPR52688.2022.01321.

[32] R. T. Ionescu, S. Smeureanu, M. Popescu, and B. Alexe, "Detecting Abnormal Events in Video Using Narrowed Normality Clusters," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA, Jan. 2019, pp. 1951–1960, https://doi.org/10.1109/WACV.2019.00212.

[33] D. Gong *et al.*, "Memorizing Normality to Detect Anomaly: Memory-Augmented Deep Autoencoder for Unsupervised Anomaly Detection," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), Oct. 2019, pp. 1705–1714, https://doi.org/10.1109/ICCV.2019.00179.

[34] B. Ramachandra and M. J. Jones, "Street Scene: A new dataset and evaluation protocol for video anomaly detection," in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Snowmass, CO, USA, Mar. 2020, pp. 2558–2567, https://doi.org/10.1109/WACV45572.2020.9093457.

[35] H. Park, J. Noh, and B. Ham, "Learning Memory-Guided Normality for Anomaly Detection," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 14360–14369, https://doi.org/10.1109/CVPR42600.2020.01438.

[36] Z. Liu, Y. Nie, C. Long, Q. Zhang, and G. Li, "A Hybrid Video Anomaly Detection Framework via Memory-Augmented Flow Reconstruction and Flow-Guided Frame Prediction," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 13568–13577, https://doi.org/10.1109/ICCV48922.2021.01333.