2023 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET) | 979-8-3503-0415-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/IICAIET59451.2023.10291749

Spatial Feature Based Violence Detection Using Convolutional Neural Network

Tirthendu Prosad Chakravorty [1], Mobashra Abeer [2], Shaiane Prema Baroi [3], Sristy Roy [4], Dewan Ziaul Karim [5]

[1,2,3,4,5] Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh

Email: tirthendu.prosad.chakravorty@g.bracu.ac.bd [1], mobashra.abeer@g.bracu.ac.bd [2],
shaiane.prema.baroi@g.bracu.ac.bd [3], sristy.roy@g.bracu.ac.bd [4], ziaul.karim@bracu.ac.bd [5]

***Abstract*—In the past decade, surveillance cameras have become a necessary integration for security measures in all types of localities. The omnipresence of these devices has substantially aided in tackling violent criminal activities. However, human error and biased judgment often result in delayed response and erroneous detection. In larger systems, continuous manual monitoring has become a cumbersome task. Therefore, automated recognition of aggressive activities in surveillance systems can enhance the remote monitoring experience and increase the preciseness of response. Previous experiments on various deep-learning techniques and Convolutional Neural Networks (CNN) tackled the challenge by identifying potential violent activities in real-time. The aim of this research is to benefit from reduced computational cost while maintaining optimality for practical implementation in real life. In this study, a lightweight yet highly effective CNN model has been proposed that can classify violent and non-violent behavior in surveillance footage solely based on spatial features. The model has undergone robust tuning and training and is capable of accurately extracting frame-level features. It was then evaluated conclusively on a combination of multiple benchmark datasets to see how well each of them performed. In conclusion, the proposed model has achieved an outstanding test accuracy of 99.6% and outperforms other popular CNN architectures by great margins.**

Keywords*— *violent activity, surveillance systems, activity recognition, deep learning, neural networks, image processing

I. Introduction

Mankind has witnessed numerous significant breakthroughs, wondrous discoveries, and colossal technological advancements throughout the years of its existence on Earth. However, one of the major issues that remains is violence. This research aspires to detect physical violence in real-time from the video footage of surveillance cameras installed throughout a city, with the help of Deep Learning. Once physical violence is detected through the proposed CNN architecture, it will be able to alert the authorities to the crime, allowing them to take appropriate action in response to the incident. This system will undoubtedly assist law enforcement agencies in providing a better and safer city for their citizens while they are on the streets.

The main focus of this research is to find a constructive model to deal with fast and potent automated violence detection from real-time videos. Although this has already been achieved by different deep learning models alongside several techniques that were developed to aid object detection and action recognition, not all models have proven suitable for lightweight devices because of their massive computational requirements. Hence, we analyzed CNN architectures like VGG, MobileNet, ResNet, DenseNet, and Inception, during which it was necessary to modify certain hyper-parameters to achieve the best accuracy. These models were then trained on vast datasets of violent videos to analyze the spatial features from each frame. Therefore, our initial work was to compare the different detection models and determine the best model in terms of accuracy. As a result, we developed an efficient and cost-effective architecture that operates on fewer parameters but achieves superior accuracy for violence detection in real-time videos. To achieve this, it was necessary to:

- Make use of a variety of distinctive and extensive datasets.
- Deeply understand how various machine learning models work and how they are pertinent to the research.
- Apply the deep learning algorithms and analyze the results.
- Build an effective and reliable model for violence detection.

II. Related Work

Elevating the performance of detecting anomalies in videos has been an essential criterion, and such activities have been detected using a semi-supervised learning approach. The study in [1] proposed a reinforcement learning model that used this approach and was based on a hard attention mechanism with collaborative agents that removed needless data from the network's input. Again, the study in [2] applied a weakly supervised approach that annotated videos and looked for anomalies that were classified with the aid of BERT and MIL on CNN's refined snippets, and later on, LSTM was also applied for classifying videos.

For this research, a significant factor is computational power. The study in [3] introduced the MobileNetV2 architecture which outperforms SOTA performance for lightweight devices with limited resources. Along with MobileNet, a modified Single Shot Detector was used, and the authors compared the approach with YOLOv2 and ResNet-101. Similarly, in paper [4], different machine learning methods were used, such as ResNet and VGG-19 models, to distinguish the frames and classify them. The Twostream Multi-dimensional Convolutional Network (2s-MDCN) was used in paper [5] for violence detection from real-time videos that showed remarkable results with lower computational cost and smaller parameter size when compared to other models.

In [6], a Separable Convolutional LSTM (SepConvLSTM) was implemented where the pre-trained MobileNet took an input of frames with a suppressed background and another stream determined changes in consecutive frames. The combination of CNN and LSTM was once again explored in paper [7] where a CNN architecture was particularly applied to acquire frames from videos and was combined with the aid of an LSTM variant using convolutional gates for perceiving localized spatiotemporal features. Similar to this, in [8] a novel method was presented to identify violent behavior utilizing a combination of two CNN architectures- AlexNet and SqueezeNet where each network was then followed by a separate Convolution Long Short Term Memory (ConvLSTM) to extract deeper and stronger features from a video in its final concealed state.

With the goal of designing an object detector with fast and optimal speed, the authors in paper [9], [10], [11], [12] analyzed the YOLO architecture. Many CNN-based object detectors failed to process live footage and depended on many GPUs. The paper [9] used YOLOv4 and also analyzed the impact of the SOTA Bag-of-Freebies and Bag-of-Specials methods for object detection and established the new Mosaic and Self-Adversarial Training for augmenting input images. On the other hand, the work in [10] investigated the You Only Look Once (YOLO) architecture for human action recognition with the aim to reduce computation time and training overhead by recognizing a scene through instances of visual data. The research in [11] introduced a method for the detection of fast objects, which was derived from the YOLO-v4-tiny architecture ideal for embedded devices. For the field of multiple object tracking (MOT), paper [12] suggested a proposed method that added YOLOv7 as an object detection network to DeepSORT to obtain the YOLOv7-DeepSORT model that enhanced object detection accuracy and speed, more than its former variants.

III. Dataset

A. Dataset Description

For the research purposes of this paper, a compilation of four different datasets was used: (a) the Real Life Violence Situations Dataset(RLVD) [13] with 2000 violent and non-violent videos, (b) the Smart City Violence Detection dataset(SCVD) [14] consisting of 248 videos categorized into non-violence, 112 videos as violence and 124 specifically as weapon violence (c) the Hockey Fights Dataset [15] with 1000 videos in total and (d) Bus Violence dataset [16] comprised of 700 videos in each class. This was done to get a wider range of data with the goal of training the model better and achieving higher accuracy.

These four datasets were used as the primary reference material, incorporating them into one cohesive dataset. In total, 4815 videos were collected for the final dataset for the research. However, due to memory restrictions of the computing device, only 2506 videos were selected and distributed into two categories: Non-Violence (1258 videos) and Violence (1248 videos). After these videos underwent certain preprocessing, the dataset was further divided into three parts: training (65%), validation (25%), and testing (10%). For the testing phase, the remaining videos from the compiled dataset that were not initially included in the original dataset were used. The figures 1 and 2 depict the class distribution of the four individual datasets, the compiled dataset's class distribution, and the splitting of this compiled dataset.

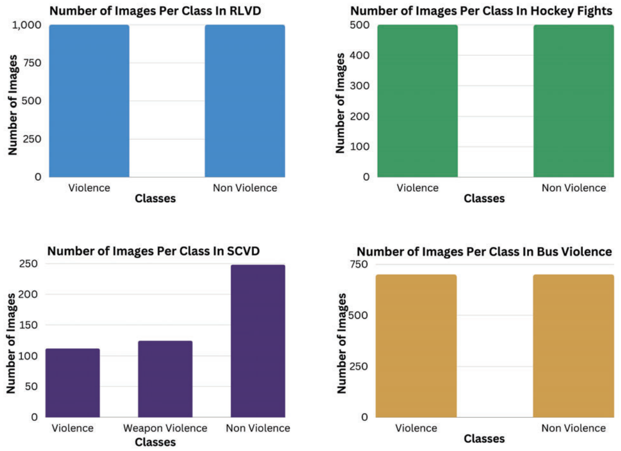

Fig. 1. Class Distribution of the Datasets

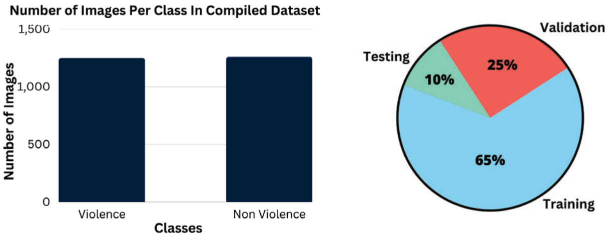

Fig. 2. Class Distribution of the Compiled Datset

B. Data Preprocessing

To begin with, it was necessary to convert the video files into frames to utilize them in the models. Videos from different sources had obvious differences in dimensions. Furthermore, modification of raw data could contribute to enhancing the performance of the neural network. Hence, the pre-processing was done in 3 stages; i) Frame extraction, ii) Frame resize, and iii) Frame Augmentation. Firstly, frame extraction was performed by iterating through each video file and extracting frames from each video.

(a) Violent

(b) Non-Violent

Fig. 3. Random Frames from the Dataset

Then the BGR color channels of each were converted to RGB color channels. Resizing each frame was done to conform to the expected input shape of the model. So, each frame was represented as a 3-dimensional array along with its color channels. The frames were reshaped to 112x112 pixels. Frames were augmented by zooming and adjusting brightness. Finally, each resulting frame was normalized before storing it for further use.

IV. Methodology

Throughout the research, the focus was to maintain a reasonable model size for a lightweight computation for spatial feature extraction and obtain the maximum accuracy while utilizing the minimum number of parameters with a balanced tradeoff between the two. First of all, the videos in the dataset were iterated, extracting the frames from each video along with the corresponding label for each frame. Since it was a compiled dataset, the dimensions of the frames were unequal and needed to be resized accordingly. The resolution of the frames was reduced to only 112x112 pixels and 3 color channels for each frame and were normalized subsequently. The proposed model contained blocks of layers. Each block contained a Conv2D layer, a Batch Normalization layer, and a MaxPooling2D layer sequentially. There were a total of seven such blocks of such layers. To standardize the inputs to subsequent layers and stabilize the learning process, the Batch Normalization layer was utilized. MaxPooling2D reduced the dimensions of the hidden layer and minimized computation. To tackle the overfitting problem, one Dropout layer with a rate of 50% was used in the hidden layer. Each of the seven Conv2D layers contained the RELU activation function. RELU was the desired choice because, unlike other activation functions, it had proven to speed up the stochastic gradient descent. Afterward, a Flatten layer was used to obtain a one-dimensional array which was then sequentially fed into the output layer. The output layer used the Sigmoid activation function for the binary classification.

It is to be noted that the choices of parameters and hyperparameters were not entirely random. The layers were finely tuned for the best hyperparameter combinations using the Bayesian Optimization Tuner. Each convolution layer was tuned with the kernel sizes of 3x3, 5x5, and 7x7. By setting up boolean values, it was also determined whether a convolution block would be followed by a dropout layer or not with each dropout layer, if present, having a combination of dropout rates of 20%, 30%, and 50%. The number of fully connected layers was also decided using rigorous tuning. However, after obtaining the best result, it was identified that no Dense layers were required apart from the output layer.

Finally, the tuned model contained two Conv2D layers with 64 filters and a kernel size of 3x3 followed by two Conv2D layers with 128 filters and a 5x5 kernel. It then sequentially had just one Conv2D with 512 filters and 3x3 kernel size followed by a Dropout layer with a 50% dropout rate. Finally, another two Conv2D layers followed with 1024 filters and a 3x3 kernel size. As mentioned above, each layer was followed by a batch normalization and a pooling layer. Figure 4 illustrates the architecture of the model and a summary of the layers can be found in Table I.

During the compilation of the model, the Adam optimizer was utilized with an initial learning rate of 1×10^{-4}. However, using the same learning rate throughout the learning process led to local minima for the validation loss. To tackle this problem, the ReduceLROnPlateau function was used to decrease the learning rate if local minima were reached. It was also observed that the model converged early in the training process even though 80 epochs were set for training. Hence, to avoid repetition, the Early Stopping technique was utilized with a patience level of 12.

To visualize the outputs, a completely unseen portion of the dataset was used. Similar to the preliminary preprocessing, the videos were once again converted to frames and were passed to the trained model. This time the model predicted a label, 0 or 1, just using the spatial features of the frames and outputted the decision on the video frame. Figure 5 to figure 8 show some sample outputs for the predicted labels in the visualization phase.

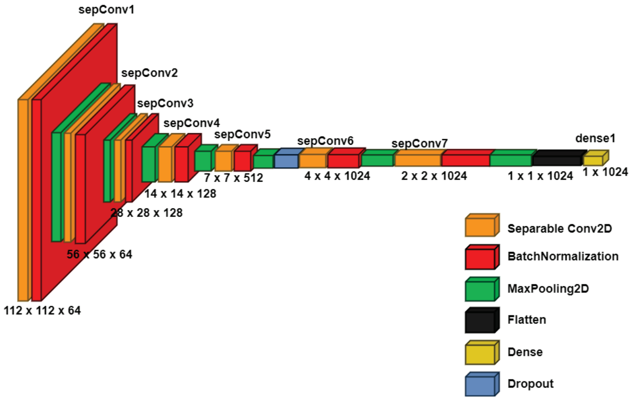

Fig. 4. Visualization of Proposed CNN Architecture

TABLE I. Summary of Layers

Layers	Output Shape	Parameters
Separable Conv2D	(None, 112, 112, 64)	283
BatchNormalization	(None, 112, 112, 64)	256
MaxPooling2D	(None, 56, 56, 64)	0
Separable Conv2D	(None, 56, 56, 64)	4736
BatchNormalization	(None, 56, 56, 64)	256
MaxPooling2D	(None, 28, 28, 64)	0
Separable Conv2D	(None, 28, 28, 128)	9920
BatchNormalization	(None, 28, 28, 128)	512
MaxPooling2D	(None, 14, 14, 128)	0
Separate Conv2D	(None, 14, 14, 128)	19712
BatchNormalization	(None, 14, 14, 128)	512
MaxPooling2D	(None, 7, 7, 128)	0
Separate Conv2D	(None, 7, 7, 512)	67200
BatchNormalization	(None, 7, 7, 512)	2048
MaxPooling2D	(None, 4, 4, 512)	0
Dropout	(None, 4, 4, 512)	0
Separate Conv2D	(None, 4, 4, 1024)	538112
BatchNormalization	(None, 4, 4, 1024)	4096
MaxPooling2D	(None, 2, 2, 1024)	0
Separate Conv2D	(None, 2, 2, 1024)	538112
BatchNormalization	(None, 2, 2, 1024)	4096
MaxPooling2D	(None, 1, 1, 1024)	0
Flatten (Flatten)	(None, 1024)	0
Dense (Dense)	(None, 2)	2050
Total params: 1,728,989	Trainable params: 1,723,101	Non-trainable params: 5,888

Fig. 5. Sample output for label non-violence

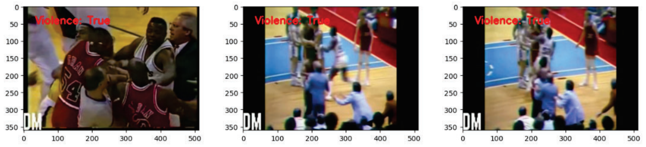

Fig. 6. Sample output for label violence

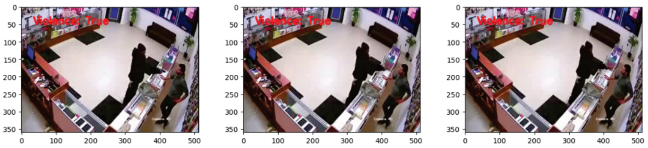

Fig. 7. Sample output for label violence

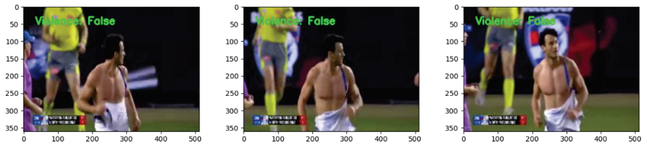

Fig. 8. Sample output for label non-violence

V. Results

The proposed CNN model results were compared with the above-mentioned five other pre-trained CNN architectures, which include: VGG-19, DenseNet-201, Inception V3, ResNet 50, and MobileNet V2. Among these deep learning models, Inception V3 had the lowest accuracy level with an accuracy score of only 0.896. The provided version of the CNN model gave the highest accuracy score with the least amount of parameters among all the models and maintained a high classification accuracy between violence and non-violence predictions on its confusion matrix. This made the model reliable, lightweight, and more accurate than other pre-trained architectures. Evaluation metrics such as accuracy, precision, recall, and f1-score were used to compare the results.

The following evaluation metrics were used, and the corresponding formulae of these metrics are given below:

$$Precision = \frac{TP}{TP+FP} \tag{1}$$

$$Recall = \frac{TP}{TP+FN} \tag{2}$$

$$F1-Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3}$$

A. Comparison With Pre-Trained Architectures

The proposed model, when compared to other pre-trained architectures using the compiled dataset, showed a convincingly higher accuracy value and higher precision, recall, and F1-score with the lowest number of parameters (1.72 million). Hence, it demonstrated the most reliable prediction capability among other models. Table II shows the comparison of different metrics attained by all the respective models.

It was made sure that the pre-trained architectures and the proposed model had certain factors like frame dimension, learning rate, and epoch number consistent in order to maintain a fair comparison among the models. The results are solely based on the models' performance, and no other factors altered the outcomes.

TABLE II. Comparison of metrics using compiled dataset

Deep Learning Models	Accuracy	Precision	Recall	F1-Score	Params
VGG-19	0.948	0.95	0.95	0.95	20.02M
DenseNet-201	0.966	0.97	0.97	0.97	18.32M
Inception-V3	0.896	0.90	0.90	0.90	21.80M
ResNet-50	0.908	0.91	0.91	0.91	23.58M
MobileNet-v2	0.974	0.97	0.97	0.97	2.25M
Proposed Model	0.996	1.00	1.00	1.00	1.72M

B. Comparison Among Different Datasets

To further evaluate the model's performance and the significance of compiling the datasets, training was done using the benchmark datasets, i.e., Hockey Fights and RLVD, individually from which the videos were originally taken. Both the compiled and Hockey Fights datasets performed equally while RLVD slightly fell behind yet showed a promising outcome. Table III shows the comparison of different metrics attained using all three respective datasets.

TABLE III. Comparison of metrics using different datasets

Datasets	Accuracy	Precision	Recall	F1-Score
Hockey Fights	0.996	1.00	1.00	1.00
RLVD	0.987	0.99	0.99	0.99
Compiled	0.996	1.00	1.00	1.00

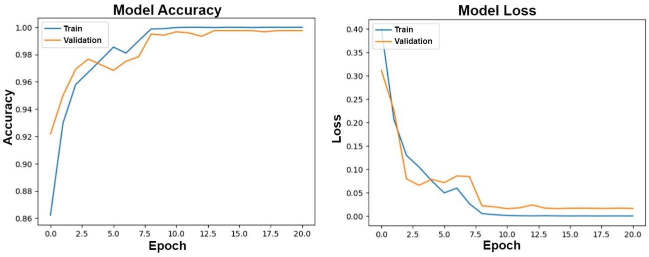

Fig. 9. Accuracy and Loss curve for Hockey Fights Dataset

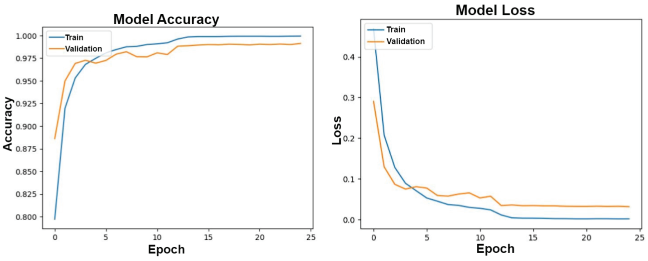

Fig. 10. Accuracy and Loss curve for RLVD Dataset

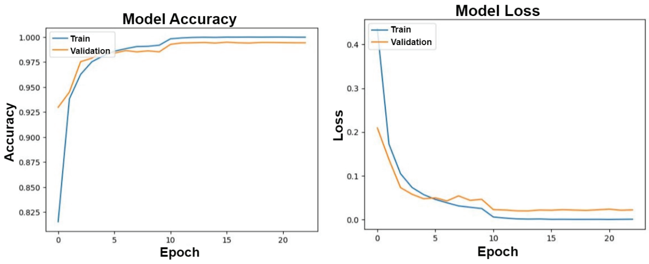

Fig. 11. Accuracy and Loss curve for Compiled Dataset

Even though the Hockey Fights and compiled datasets yielded the same result, the training process was more stable with the compiled datasets. Comparing figure 11 with figure 9, it can be seen that the accuracy and loss curves of the former contain fewer fluctuations and stabilize quicker.

VI. Limitations of the Paper

The proposed spatial feature-based model proved to be quite efficient in labeling action states from a single frame. However, without the entirety of a sequence, it might lead to strictly logic-driven conclusions which, in real-life situations, might not be practical to the human eye. Action sequencing plays an important role in depicting such situations. For instance, a hard pat on the back of a person might be labeled as a violent action frame, which in real life is just a zestful gesture. In such scenarios, single-shot action patterns might be incorrectly classified.

Since the goal of this research was to classify violence in a video in general, all the collected datasets were compiled in accordance with a binary classification problem, and the model was trained accordingly. Therefore, it cannot be fully claimed that the model will perform equally well if the types of violence are further classified. However, the compiled dataset contains multiple different scenarios. With the current performance, the model holds the potential to classify multiple types of violence without adding further complexities.

VII. Conclusion and Future Work

If modern-day surveillance cameras are equipped with a violence detection algorithm, they can lead to significant advancements in security systems for law enforcement and security agencies. Furthermore, public safety can be ensured in public spaces, schools, transportation, etc., if security personnel can intervene promptly. Incorporating an effective and lightweight model with surveillance cameras can tackle this task with great efficacy. With the suggested CNN model, this becomes possible since it uses fewer parameters than many other pre-trained deep-learning models.

The future work of this research paper will focus on collecting and annotating large-scale datasets specifically designed for violent activity detection. Using this wide range of data, different machine-learning models will be trained for better accuracy. Currently, this research focuses on detecting violent and non-violent activities, but the aim will be to further classify the type of violence such as physical violence, use of weapons and other tools, vandalism, mugging, and many more. In the near future, consideration will be given to incorporating methods like attention mechanisms into the suggested model to capture temporal dependencies. Moreover, integrating Vision Transformers and comparing them with state-of-the-art models will provide a clearer picture of the performance of transformers in violent activity detection. On the other hand, understanding the context in which violent activities occur, such as scene context, object context, and social context, can significantly improve detection accuracy.

References

[1] Hamid Mohammadi and Ehsan Nazerfard. Video violence recognition and localization using a semi-supervised hard-attention model. 02 2022.

[2] Weijun Tan, Qi Yao, and Jingfeng Liu. Overlooked video classification in weakly supervised video anomaly detection. 10 2022.

[3] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. pages 4510–4520, 06 2018.

[4] Nadia Mumtaz, · Ejaz, Shabana Habib, Syed Muhammad Mohsin, Muhammad Mohsin, Prayag Tiwari, Shahab S. Band, and Neeraj Kumar. An overview of violence detection techniques: Current challenges and future directions. *Artificial Intelligence Review*, 09 2022.

[5] Dipon Ghosh and Amitabha Chakrabarty. Two-stream multi-dimensional convolutional network for real-time violence detection. 11 2022.

[6] Zahidul Islam, Mohammad Rukonuzzaman, Raiyan Ahmed, Md Kabir, and Moshiur Farazi. Efficient two-stream network for violence detection using separable convolutional lstm. pages 1–8, 07 2021.

[7] Swathikiran Sudhakaran and Oswald Lanz. Learning to detect violent videos using convolutional long short-term memory. 09 2017.

[8] Heyam Mohammed and Lamiaa Elrefaei. Detecting violence in video based on deep features fusion technique. 04 2022.

[9] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-yuan Liao. Yolov4: Optimal speed and accuracy of object detection. 04 2020.

[10] Shubham Shinde, Ashwin Kothari, and Vikram Gupta. Yolo based human action recognition and localization. *Procedia Computer Science*, 133:831–838, 01 2018.

[11] Zicong Jiang, Liquan Zhao, Shuaiyang Li, and Yanfei Jia. Real-time object detection method based on improved yolov4-tiny. 11 2020.

[12] Feng Yang, Xingle Zhang, and Bo Liu. Video object tracking based on yolov7 and deepsort. 07 2022.

[13] Mohamed Mostafa Soliman, Mohamed Hussein Kamal, Mina Abd El-Massih Nashed, Youssef Mohamed Mostafa, Bassel Safwat Chawky, and Dina Khattab. Violence recognition from videos using deep learning techniques. In *2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS)*, pages 80–85. IEEE, 2019.

[14] Toluwani Aremu, Li Zhiyuan, Reem Alameeri, and Abdulmotaleb El Saddik. Sividet: Salient image for efficient weaponized violence detection, 2023.

[15] Enrique Bermejo Nievas, Oscar Deniz Suarez, Gloria Bueno García, and Rahul Sukthankar. Violence detection in video using computer vision techniques. In *Computer Analysis of Images and Patterns: 14th International Conference, CAIP 2011, Seville, Spain, August 29-31, 2011, Proceedings, Part II 14*, pages 332–339. Springer, 2011.

[16] Luca Ciampi, Paweł Foszner, Nicola Messina, Michał Staniszewski, Claudio Gennaro, Fabrizio Falchi, Gianluca Serao, Michał Cogiel, Dominik Golba, Agnieszka Szczesna, et al. Bus violence: an open benchmark for video violence detection on public transport. *Sensors*, 22(21):8345, 2022.