Towards Smart City Security: Violence and Weaponized Violence Detection using DCNN

Toluwani Aremu[§], Li Zhiyuan[§], Reem Alameeri[§], Moayad Aloqaily, Mohsen Guizani

Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI), UAE
E-mails: {toluwani.aremu; li.zhiyuan; reem.alameeri; moayad.aloqaily; mohsen.guizani}@mbzuai.ac.ae

***Abstract*—In this ever connected society, CCTVs have had a pivotal role in enforcing safety and security of the citizens by recording unlawful activities for the authorities to take actions. In a smart city context, using Deep Convolutional Neural Networks (DCNN) to detection violence and weaponized violence from CCTV videos will provide an additional layer of security by ensuring real-time detection around the clock. In this work, we introduced a new specialised dataset by gathering real CCTV footage of both weaponized and non-weaponized violence as well as non-violence videos from YouTube. We also proposed a novel approach in merging consecutive video frames into a single salient image which will then be the input to the DCNN. Results from multiple DCNN architectures have proven the effectiveness of our method by having the highest accuracy of 99%. We also take into consideration the efficiency of our methods through several parameter trade-offs to ensure smart city sustainability.**

***Index Terms*—Smart Cities, Violence Detection, DCNN, CCTV, Sustainability**

I. Introduction

Smart Cities development is an emerging innovative research and development field. It brings in perspectives from statistics, architecture, real estate management, urban planning, environmental studies, sustainability, etc. Smart city activities and technologies have mostly had the aim of producing and analyzing data and gaining new knowledge on the complexity and dynamics of a city [1]. Recently, AI has taken this innovative field of research to the next step of utilising the data and knowledge to support decision-making.

One of the modern enabling technologies for Smart Cities development and research is the use of Machine Learning algorithms and Data Analysis tools to perform analysis on data and make future decisions based on the information gotten from the data. Also, the use of Neural Networks make it seem that there is an artificial brain making complex decisions as fast as possible. These data are gotten from another enabling technology which is Internet of Things (IoT). This technology in particular connects any device to the internet and share its sensed data. Devices like temperature sensors, gyroscope, alarm systems, location sensors, CCTVs, and sensors used by autonomous systems and residential houses can be connected with IoT [2].

[§]These authors contributed equally to this work

The use of AI in smart cities can be categorized into may dimensions i.e. Smart Education, Smart Governance, Smart Mobility and Infrastructures, etc. While it is a great thing that AI is being used in smart city applications to make tasks easier for citizens and users, one aspect of smart cities we must not forget is the safe-guarding of lives and properties. Cases of violence and gang related activities in a city could be rampant and serious especially if there is no way the required authorities can get to the scene in time to curb further destruction. Some of these violent activities could result to loss of lives and properties especially when weapons are used. We've witnessed or heard cases of road rage violence, gang related violence, and other random acts of violent crimes [3] [7] [8]. Therefore, focusing on the protection of lives and properties is an imperative task that shouldn't be ignored in the research and development of smart cities.

Due to the growth of CCTV infrastructures, current research in the surveillance field have focused on detecting violence and identifying crimes using DNNs. Ever since the breakthrough of AlexNet in the ImageNet 2012 competition [4] [5], DNNs have been the go-to AI technology for different real world tasks. While some of the current work done focuses on detecting violence in videos leveraging DNNs coupled with different techniques, others focus on weapon detection in videos using object detection DNNs. The datasets currently used for weapon detection purposes are limited to specific well-identified weapons i.e. Knives, Guns, etc. This makes the detection too ambiguous in nature, as it ignores other objects with potentials of being used as weapons. In the real world, any object can be used as a weapon, as long as it can cause a huge damage when applied. Also, no research done in this domain has combined both violence and weaponized violence detection tasks.

Furthermore, while the work that has been done in this domain have worked well at either detecting weapons or detecting violence, none has considered one of the main goals of smart cities, which is **Sustainability**. DNNs i.e. C3D, ConvLSTM, etc, used in this domain have large amounts of parameters and uses a lot of energy due to the large computations involved. Though inference using these algorithms could be fast, we believe inference could be faster when using smaller DNNs, with much lesser parameters

which could in turn improve the energy efficiency of these systems when deployed in a smart city context.

Our research focuses on detecting two kinds of violence: violence with weapons and violence without weapons. We focus on considering this research in a smart city context to achieve sustainability. To this end, we convert a video classification task into an image recognition task and leverage DCNNs and we also show the trade-offs to be considered to achieve sustainability when choosing a neural network for such tasks. Our main contributions are summarized below:

- We create a new benchmark called the Smart-City CCTV Violence Detection dataset (SCVD). Current datasets [9] [10] for violence detection contain videos recorded from phone cameras which could alter the needed CCTV distributions. Furthermore, this dataset contains weaponized violence class so it could be used by DNNs to learn the distribution of any potential weapons and infer for quicker action to be carried out on such by the authorities. This means that our dataset is tuned to the fact that any handheld object which could be used to harm humans and properties could be regarded as a weapon.
- We propose a new novel technique (Salient Image), based on [11] which converts our task from a video classification task to an image recognition task, to improve its speed and simplicity.
- We leverage multiple DCNN architectures, and compare their results and components. We show the trade-offs to use when selecting what neural network to use when trying to achieve sustainability in a smart city's context. Our results show that DCNNs having small depth and parameters but with optimal combination of the trade-offs can provide good performances, improve energy efficiency and mitigate violence in smart cities.

II. Related Works

There have been several smart approaches and research which have been focused on handling the employment of video surveillance for violence detection, crime detection and weapon detection. In this section, we show some of our most related work and literature.

[17] leveraged a pretrained Faster R-CNN deep learning model on automatically detecting handheld guns in clustered scenes, resulting in a 93% accuracy. Nevertheless, the work is only concerned with detecting handheld guns and therefore, disregarding other forms of violence. Similarly, [18] used Faster R-CNN and Single Shot Detector for guns and other weapons detection. However, due to the huge imbalance in the dataset, and naivety of the classifier, the classifier was quick to identify more handheld objects as a weapon, and majorly as a gun. The average accuracy of the model was 84.6% and therefore, required more work done to distinctively identify the differences between guns and other similarly handheld devices. [19] employed multiple DCNN and object detection DNN architectures to detect weapons in real time CCTV videos. Their experiment with Yolo V4 provided the best result with an accuracy of 91%, but had a shortcoming of still outputting some false positives in obvious cases.

Furthermore, while there are multiple attempts that focused on weapons detection only, there also exists other literature that aim to detect violence which does not include the use of weapons. [20] leveraged a pretrained InceptionNet for which they recorded a 99.28% and 99.97% accuracies on both Hockey and Movies datasets. The distributions however is far from the CCTV domain as they are sports related and movies, and are not recorded with a CCTV, hence defeating the purpose of violence detection through CCTVs. [13] and [21] used the ConvLSTM to detect violence in CCTVs. [21] used a pretrained ResNet50 model to extract spatial features from the video frames and send it to the ConvLSTM model while [13] leveraged the VGG16 architecture. However, the large networks used in these literature are more energy-consuming while the dataset they were trained on hugely lacks the CCTV distributions as most of the videos are recorded with mobiles.

[11] worked on action recognition. They argued that instead of using big and hugely deep networks for video action recognition tasks, a simple image classifier would work. For this, they proposed a technique which extracts frames from videos, resize these frames and combine them into a super image. Their results show an improvement over state-of-the-art deep Conv3D and LSTMs architectures for video action recognition. We build upon this approach, and propose a new technique specifically for Violence and Weaponized Violence detection. To the best of our knowledge, all literature conducted in the domain of security in surveillance systems tackle either weapon detection or violence detection. Current datasets are built to suit those needs. Hence, we created **SCVD**, a new benchmark to detect violence and weaponized violence in CCTVs. This is in addition to the lack of consideration of a Smart City environment/perspective which does not only require smart security, but also other characteristics such as sustainability. For this reason, our research also focus on investigating trade-offs for selecting an efficient DCNN architecture for smart city security while maintaining a good performance.

III. Dataset

In this section, we propose Smart-City CCTV Violence Detection Dataset (SCVD), a new benchmark for purely CCTV recorded violence and weaponized violence events. Table I shows a summary of the comparison with other datasets.

There are mainly two kinds of datasets we considered in the table, which contains different data types. The first kind are the datasets used for violence detection, of which all contains videos. The second kind are datasets used for weapons detection, of which one contains images. The NTU CCTV-Fights dataset [9] has the largest collection of recorded

TABLE I
COMPARISONS BETWEEN THE SCVD AND THE PREVIOUS DATASETS

Dataset	Type	Size	Length/video (sec)	Annotation	Violence	Weapons	Characteristics	Scenario
NTU CCTV-Fights [9]	Video	1000 videos	5-720	Frame	Yes	No	CCTV + Mobile	Natural
Hockey Fight [12]	Video	1000 videos	1-2	Video	Yes	No	Aerial Camera	Hockey Games
RLVS [13]	Video	2000 videos	5-15	Video	Yes	No	CCTV + Mobile	Natural
RWF-2000 [14]	Video	2000 videos	5	Video	Yes	No	CCTV	Surveillance
Sohas [15]	Image	3255 images	N/A	Image	No	Yes	Captured Images	Demonstrations
WVD [16]	Video	168 videos	10-72	Video	No	Yes	Synthetic	Computer Games
Ours	Video	500 videos	5-10	Video	Yes	Yes	CCTV	Surveillance

violent events. It contains 1000 untrimmed and unprocessed videos which are annotated at the frame level. However, most of the videos are recorded with mobile cameras, which defeats the need for the CCTV distribution. Data Distribution is very important in these kind of tasks. Also, the data annotation proves to be very tricky and inconsistent, which could be a problem during training. Another problem is that it classifies all sort of violence under the general form. The Hockey Fight dataset [12] also contains 1000 videos gotten from aerial cameras in the hockey stadium. While the videos are annotated are video level and are classified into either fights or no fights, the dataset distribution isn't usable for violence detection in a real world.

The Real Life Violence Situation dataset [13] were annotated at the video level as they were trimmed and also classified into Violence and Non Violence videos. Both classes contain 1000 videos of lengths 5 to 15 seconds. Some of the Violent class videos however were acted, and it is possible for deep neural networks to identify acted out events to real events using techniques used in anomaly detection. Also, the dataset contains a lot of non CCTV recorded events. RWF-2000 [14] was created as a solution to these issues. They contained fully preprocessed video-level annotated videos that were classified into violent and non violent situations. Each video has a length of 5 seconds and 30 frames could be extracted in each second. While they tried to improve the data quality and limit the dataset to pure CCTV recordings, they classified all forms of violence under the Violent class, which defeats our purpose of detecting weapons too asides violence detection.

Furthermore, while there are multiple datasets that were created for violence detection, there exists other datasets used to detect weapons in video surveillance systems. Sohas [15] proposed multiple benchmarks for weapons detection, but since other benchmarks were focused on guns and knives alone, they created the Sohas Weapons and similar handled objects dataset which had 3255 images showing captured demonstrations of someone holding different weapons and other objects that aren't weapons. WVD [16] created synthetic animated videos using the GTA-V computer game. These videos show different scenarios in which weapons could be used. They also classified weapons into hot weapons which contains different kinds of guns and cold weapons like knives, batons, etc. However, the distribution of WVD's data is very much different from the distribution one can get from CCTV recordings.

Out of all the discussed datasets, the most similar to ours is the RWF-2000. We got our dataset by getting real life CCTV recorded violence, non-violence and weapon violence videos, and trimmed all videos to get a length ranging from 5 to 10 seconds. We classified these videos into Violent, Non-Violent and Weaponized-Violent events with the aim that our deep learning models are able to learn and distinctively identify the difference between violent events involving weapons, and violent events without weapons.

IV. METHOD

In a classic CNN architecture, the input to the convolutional layer is usually a single image with the object of interest mostly at the center of the image. In this way, the CNN learns the context image by image without relations to the previous input. But this may not be the best option for video input, since each frame is related to the frame before and after. Therefore, the temporal context is also important in order for the CNN to make a decision.

With this in mind, the authors of [11] explored the possibility of merging multiple input frames from a video to form a super image before sending this input to the CNN. They experimented with different spatial combinations and found that a square formation for merging the images has the best performance as compared to linear or rectangular formations.

In our method, for each class in the SCVD dataset, all videos are converted to images on a per-frame basis, without skipping frames. With the frames converted from the videos, they are merged together in a 5×3 grid or a 3×2 grid in successive frames. These merged frames form the larger salient images. Figure 1 shows the process of how the salient image is created. Since the frames extracted from CCTV videos are usually rectangular in shape, mostly in the 720p (1280×720) dimensions, in the creation of the salient images for this task, original resolution of the frames are used instead of resizing them to smaller square images of 224×224 which is the approach taken by [11].

Fig. 1. Salient Image: A sequence of video frames gotten from a CCTV surveillance system are first rearranged into a salient image based on a 5 × 3 or 3 × 2 spatial arrangement, and then fed into a CNN architecture for training and recognition.

Furthermore, by keeping the original resolution, the spatial information contained within the frames are preserved (resizing to a square shape reduces the information which the models can learn from the spatial width). We chose the 5 × 3 grid and a 3 × 2 as the frames have longer width than height. Combining them using this grid structure forms a square shape and this helps in preserving the spatial information especially in the width when the overall image is resized later for input into our chosen architectures. The salient image will also contain the temporal information as the frames are merged in successive order according to video timeline.

Therefore, one salient image contains fifteen or six sequence of actions captured in a video within a specific time period.

V. Experiments and Results

In this section, we show and discuss results of the experiments carried out, the relationship between the model architectures and sustainability and how to ensure pro-activity in violence detection systems.

For the training process, we used the SGD optimizer with a learning rate of 0.001 and a momentum of 0.9. We could have gone for the Adam optimizer due to it's speed and ability to converge fast, but the SGD optimizer ensures that the model reaches the global minima. The SGD optimizer is stated as:

$$w_{t+1} = w_t - \alpha \frac{\partial L}{\partial w_t} \tag{1}$$

where w_t is the weight, w_{t+1} is the parameter being updated, α is the learning rate and $\frac{\partial L}{\partial w_t}$ is the partial derivative of the gradient.

We also leverage the categorical crossentropy loss function as it is the most suitable for our task, as it is a hyperparameter which is tuned and used to tune the optimizer. We denote this loss function L_{CE} as:

$$L_{CE} = -\sum_{i=1}^{N} y_i . \log\left(\hat{y}_i\right) \tag{2}$$

where y_i is the true probability label distribution of of the input and $log\left(\hat{y}_i\right)$ is the predicted probability distribution.

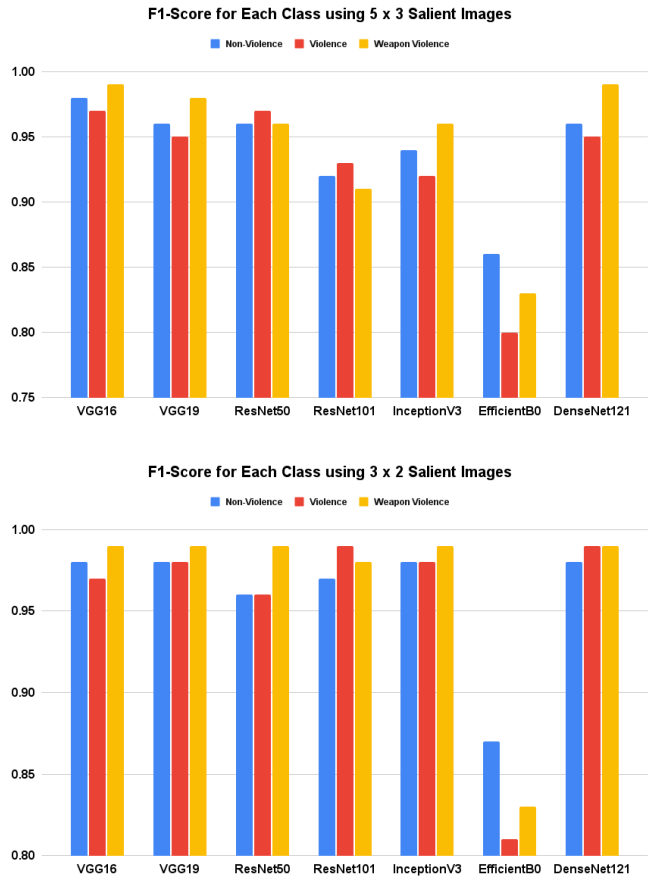

Fig. 2. The f1-score is computed to compare the performances of each model architecture across the classes (Top: 5 × 3; Bottom: 3 × 2)

A. Environment

For the training and implementation during our research, we used the following:

- System used: NVIDIA Quadro RTX6000
- GPU Memory: 24GB GDDR6
- Operating System: Ubuntu 21.04
- Deep Learning Library: Tensorflow 2.0, Keras
- Models Used:
  - VGG16 [22]

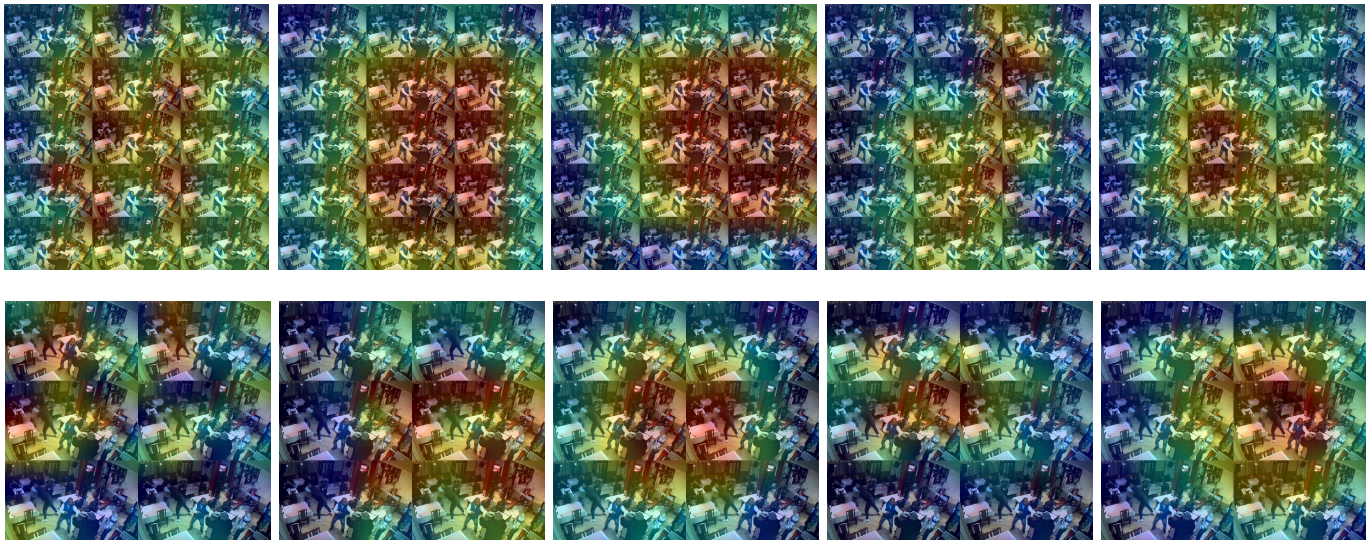

Fig. 3. GradCAM outputs showing the activated regions for each model on a 5 × 3 and 3 × 2 salient image; From L-R: VGG16, ResNet50, DenseNet121, EfficientNetB0, InceptionV3

TABLE II
SUMMARY OF RESULTS GOTTEN FROM EXPERIMENTS ACROSS DIFFERENT ARCHITECTURES AND NUMBER OF INPUT FRAMES

Model Architecture	Input Frames	Train Loss	Test Accuracy (%)
VGG16	15	0.0551	**97.98**
VGG19	15	0.0956	96.42
ResNet50	15	0.0641	96.51
ResNet101	15	0.0751	91.91
DenseNet121	15	**0.0405**	96.60
EfficientNetB0	15	0.6973	82.26
InceptionV3	15	0.0449	94.12
VGG16	6	**0.0121**	98.14
VGG19	6	0.0306	98.31
ResNet50	6	0.0429	96.79
ResNet101	6	0.0429	97.89
DenseNet121	6	0.0272	**98.65**
EfficientNetB0	6	0.5583	84.04
InceptionV3	6	0.0311	**98.65**

- VGG19 [22]
- ResNet50 [23]
- ResNet101 [23]
- DenseNet121 [24]
- EfficientNetB0 [25]
- InceptionV3 [26] [27]

B. Results

We start by extracting 5 × 3 salient images (about 5500 images) from our SCVD dataset and passing them into the seven selected CNN architectures for training. We evaluate the training loss and test accuracy for each of these architectures as they provide information as to which architecture is performing well. We also compute the F1-Score for each class (Violence, Non Violence and Weapon Violence) to make sure our selected architectures are doing well across each class (see Figure 2). We also extract 3 × 2 salient images (6000 images) from the SCVD dataset and pass them into the selected architectures to further evaluate the best grid size to choose for the different architectures. The F1-Score for each class is given as:

$$F_{1_k} = \frac{2(P_k \cdot R_k)}{P_k + R_k} \tag{3}$$

where k is the class we are evaluating the F_1 score for, P_k is the precision for the class and R_k is the recall.

From the results in Table II, it is seen that the results of the architectures using the 3 × 2 salient images outperform the same architectures using 5 × 3 salient images. While the test accuracy of the VGG16 seems to have no much difference across the two grid arrangements, the train loss on the 3 × 2 arrangement is lower furthermore proving that using six frames at a time improves the architecture's inference performance. Comparing the train loss and test accuracy for the selected architectures on the two grid arrangements, we see a huge improvement which leads to a conclusion that learning from lesser frames at an instance could be better for the any Convolutional Neural Network architecture.

For the 5 × 3 arrangement, the DenseNet121 had the least train loss out of all the selected architectures with a train loss of 0.0405, while the VGG16 had the best test accuracy, getting its prediction right 97.98% of the time. Meanwhile, for the 3 × 2 salient arrangement, the VGG16 had the lowest train loss with 0.0121, while DenseNet121 and InceptionV3 had the best test accuracy with 98.65% each. EfficientNetB0 gave the worst performance for both salient arrangements as it had train losses of 0.6973 and 0.5583 respectively, and a test accuracy of 82% and 84% across both arrangements.

TABLE III
PARAMETER TRADE-OFFS BETWEEN DIFFERENT MODELS FOR SUSTAINABILITY

Model Architecture	Input Frames	Param (M)	# Layers	Time (ms)	Val Loss	Accuracy (%)
VGG16	15	134.27	16	154.3	0.0667	98
VGG19	15	138.79	19	179.6	0.1292	96
ResNet50	15	23.5	107	137.5	0.0831	97
ResNet101	15	42.6	209	232.6	0.2441	92
DenseNet121	15	7.04	242	177.4	0.1465	97
EfficientNetB0	15	4.05	132	144.1	0.9327	82
InceptionV3	15	21.8	189	167.6	0.1746	94
VGG16	6	134.27	16	157.3	0.0569	98
VGG19	6	138.79	19	171.9	0.0632	98
ResNet50	6	23.5	107	126.8	0.1067	97
ResNet101	6	42.6	209	226.6	0.0576	98
DenseNet121	6	7.04	242	167.9	0.0426	99
EfficientNetB0	6	4.05	132	117	0.4487	84
InceptionV3	6	21.8	189	151.7	0.0757	99

To furthermore explain the performances of the selected architectures for this research, we also leverage the Grad-CAM [28] architecture for explainable AI, to explain and show what neurons are activated by each architecture and why they perform well (or not, based on the architecture). Figures 3 show the outputs for VGG16, ResNet50, DenseNet121, EfficientNetB0 and InceptionV3. Looking at the plots, we see that the VGG16 on the 5 × 3 arrangement is able to identify the actions from each specific frame within the salient image. It is also able to uniquely identify the different regions in the 3 × 2 arrangement which enables it to make a right decision. This could be said for the other architectures as the heat-map shows the regions being focused on before inference.

C. Sustainability in a Smart City Context

The use of AI in Smart Cities has improved the living conditions and ease of getting things done, but at the same time increased the carbon footprint. In our use case of smart surveillance systems for violence and weaponized violence detection, we hope to improve the effectiveness of these systems and at the same time improve the efficiency. Here, we compare our trained models based on certain trade-offs and choose the best model to be deployed for our task.

The trade-offs we consider for sustainability are: Inference time per input $[Time]$, Average Accuracy $[Accuracy]$, and Average Validation loss $[Loss]$, on the assumption that all models consume equal amounts of energy per time. We tackled this problem using a reinforcement learning approach of rewards and punishments for each considered trade-off.

Time: The inference time should be as short as possible to maximize the number of frames processed at a given time. For example, a model $\mathbb{A}$ processing at a rate of 30 frames per second is considered slower and more energy consuming than a model $\mathbb{B}$ processing at a rate of 45 frames per second, given that both models use the same amount of energy per second. The reward/punishment for our trained models $\mathbb{M}$ with respect to time T is given by:

$$\mathbb{M}_T = \begin{cases} 1, & \text{if } T < \frac{1}{N}\sum_{i=1}^{N} T_i \\ 0, & otherwise \end{cases} \tag{4}$$

Loss: The loss is also an important trade-off to be considered when selecting a model for violence detection in sustainable smart city surveillance systems. The validation loss serves as an indicator for how much a model has learnt from the training samples, and shows if a model has overfitted the training data. The lower the loss, the higher the probability of predicting True positives. The reward/punishment for our trained models $\mathbb{M}$ with respect to loss L is given by:

$$\mathbb{M}_L = \begin{cases} 1, & \text{if } L < 0.1 \\ 0, & \text{if } 0.1 < L < 0.2 \\ -1, & otherwise \end{cases} \tag{5}$$

Accuracy: While the major requirement is the speed of inference, the inference accuracy is also very important in deciding what model to use in these kinds of systems. In this work, the threshold required for us to consider that the model's performance is good enough is set at 95%. The reward/punishment for our trained models $\mathbb{M}$ with respect to accuracy P is given by:

$$\mathbb{M}_P = \begin{cases} 1, & \text{if } P > 95\% \\ 0, & \text{if } 90\% < P < 95\% \\ -1, & otherwise \end{cases} \tag{6}$$

VI. CONCLUSION AND FUTURE WORK

In this paper, we introduced the Smart City Violence Detection (SCVD) dataset, which strictly contains CCTV recorded violence and weaponized violence videos. We also introduced a novel method which allows Deep Convolutional Neural Networks detect violence and weaponized violence in videos. This method converts a video frames into a single merged image while preserving the spatial information obtained from the original size of the video frames. We employed several DCNN architectures for training and inference using our novel method on the SCVD dataset. The

Grad-CAM was used to show activated regions for different architectures during evaluation. We also considered the use of these models for violence and weaponized violence detection in a smart city context and show what trade-offs are to be considered for sustainability and reduction of carbon footprints in smart cities. Though it had the smallest number of layers, the VGG16 stood out to be the best choice for both proposed salient arrangements as it maintained a validation accuracy of 98% at an average time of 0.155 seconds. Given the right operating system and processor to be deployed on, it can predict up to 90 fps while maintaining an accuracy of 98%.

While we have considered and improved upon violence and weaponized violence detection in a smart city context, it is however imperative to work on the communications between smart surveillance systems in the future. This would shift the independent violence detection on each CCTV system to a centralized violence detection on an edge system or the cloud, further reducing the carbon footprint in a smart society. Proactivity in these surveillance systems, such as voice warning systems, should also be looked into. This would help mitigate the severity of the violence in sustainable smart cities before the arrival of necessary authorities, since it will deter the perpetrator from carrying out further harm.

REFERENCES

[1] D. Diran, A. F. Veenstra, P. Testa, and M. Kirova, "Artificial Intelligence in smart cities and urban mobility," Europarl, 12-Jul-2021.

[2] C. Englund, E. E. Aksoy, F. Alonso-Fernandez, M. D. Cooney, S. Pashami, and B. Åstrand, "AI perspectives in Smart Cities and communities to Enable Road Vehicle Automation and smart traffic control," MDPI, 18-May-2021.

[3] A. Kraft, "Woman pulls gun on man over pro-vaccine bumper sticker, police say," WFLA, 02-Mar-2022. [Online]. Available: https://www.wfla.com/mobile/woman-accused-of-road-rage-over-a-mans-pro-vaccine-bumper-sticker/. [Accessed: 12-Apr-2022].

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional Neural Networks," Communications of the ACM, 01-Jun-2017.

[5] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[6] A. Romero, "GPT-4 will have 100 trillion parameters-500x the size of GPT-3," Medium, 11-Sep-2021. [Online]. Available: https://towardsdatascience.com/gpt-4-will-have-100-trillion-parameters-500x-the-size-of-gpt-3-582b98d82253. [Accessed: 24-Mar-2022].

[7] E. Tucker, "At least 6 NYC subway stabbings reported since the mayor unveiled new Safety Plan Friday," CNN, 21-Feb-2022. [Online]. Available: https://edition.cnn.com/2022/02/20/us/nyc-five-subway-stabbings/index.html. [Accessed: 12-Apr-2022].

[8] K. Zraick, A. Southall, N. Gavrielov, and Z. Small, "Police ID suspect in stabbing of moma employees," The New York Times, 12-Mar-2022. [Online]. Available: https://www.nytimes.com/2022/03/12/nyregion/moma-stabbing.html. [Accessed: 12-Apr-2022].

[9] M. Perez, A. C. Kot and A. Rocha, "Detection of Real-world Fights in Surveillance Videos," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 2662-2666, doi: 10.1109/ICASSP.2019.8683676.

[10] M. M. Soliman, M. H. Kamal, M. A. El-Massih Nashed, Y. M. Mostafa, B. S. Chawky and D. Khattab, "Violence Recognition from Videos using Deep Learning Techniques," 2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS), 2019, pp. 80-85, doi: 10.1109/ICICIS46948.2019.9014714.

[11] Q. Fan, C.-F. Chen, and R. Panda, "An image classifier can suffice for video understanding," arXiv.org, 30-Jun-2021. [Online]. Available: https://arxiv.org/abs/2106.14104. [Accessed: 13-Apr-2022].

[12] E.R. Nievas, D.S. Oscar, B.G. Gloria, and S. Rahul. "Hockey fight detection dataset." In Computer Analysis of Images and Patterns, pp. 332-339. Springer, 2011.

[13] M. M. Soliman, M. H. Kamal, M. A. El-Massih Nashed, Y. M. Mostafa, B. S. Chawky and D. Khattab, "Violence Recognition from Videos using Deep Learning Techniques," 2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS), 2019, pp. 80-85, doi: 10.1109/ICICIS46948.2019.9014714.

[14] M. Cheng, K. Cai and M. Li, "RWF-2000: An Open Large Scale Video Database for Violence Detection," 2020 25th International Conference on Pattern Recognition (ICPR), 2021, pp. 4183-4190, doi: 10.1109/ICPR48806.2021.9412502.

[15] F. Pérez-Hernández, S. Tabik, A. Lamas, R. Olmos, H. Fujita, and F. Herrera, "Object detection binary classifiers methodology based on deep learning to identify small objects handled similarly: Application in video surveillance," Knowledge-Based Systems, vol. 194, p. 105590, 2020.

[16] M. S. Nadeem, V. N. Franqueira, F. Kurugollu, and X. Zhai, "WVD: A new synthetic dataset for video-based violence detection," Lecture Notes in Computer Science, pp. 158–164, 2019.

[17] G. K. Verma and A. Dhillon, "A handheld gun detection using faster R-CNN deep learning," Proceedings of the 7th International Conference on Computer and Communication Technology - ICCCT-2017, 2017.

[18] H. Jain, A. Vikram, Mohana, A. Kashyap, and A. Jain, "Weapon detection using artificial intelligence and deep learning for security applications," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020.

[19] M. T. Bhatti, M. G. Khan, M. Aslam, and M. J. Fiaz, "Weapon detection in real-time CCTV videos using Deep Learning," IEEE Access, vol. 9, pp. 34366–34382, 2021.

[20] A. Mumtaz, A. B. Sargano and Z. Habib, "Violence Detection in Surveillance Videos with Deep Network Using Transfer Learning," 2018 2nd European Conference on Electrical Engineering and Computer Science (EECS), 2018, pp. 558-563, doi: 10.1109/EECS.2018.00109.

[21] M. Sharma and R. Baghel, "Video surveillance for violence detection using Deep Learning," Advances in Data Science and Management, pp. 411–420, 2020.

[22] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," CoRR, vol. abs/1409.1556, 2014.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[24] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected Convolutional Networks," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[25] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for Convolutional Neural Networks," arXiv.org, 11-Sep-2020. [Online]. Available: https://arxiv.org/abs/1905.11946v5. [Accessed: 26-Apr-2022].

[26] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," arXiv.org, 17-Sep-2014. [Online]. Available: https://doi.org/10.48550/arXiv.1409.4842. [Accessed: 26-Apr-2022].

[27] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[28] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618-626, doi: 10.1109/ICCV.2017.74.