Effectiveness of Synthetic Images in Violence Detection

MUHAMMAD SHAHROZ NADEEM*, University of Derby, United Kingdom.
FATIH KURUGOLLU, University of Derby, United Kingdom.
SARA SARAVI, Loughborough University, United Kingdom
HANY F. ATLAM, University of Derby, United Kingdom
VIRGINIA NL FRANQUEIRA, University of Kent, United Kingdom

State-of-the-art deep learning based violence detection methods are mostly based on data corpus taken from either Hollywood movies or videos taken from YouTube. Violence is a subjective and sensitive matter, therefore, visual data containing such scene is often subjected to ethical implications. Therefore, this limits the amount of violence that can be shown, gathered or distributed. To tackle this problem synthetic data for violence is an alternative solution. In this paper, we gauge the effectiveness of synthetic visual data for violence detection. Experimental evaluation shows that synthetic data being less noisy, facilitates easier generation of superior feature representations. To this end, we train a vanilla VGG-16 and VGG-16+LSTM networks on WVD. Afterwards, the performance on real world data is tested by applying two transfer learning strategies. Experimental results show that applying similar approach using only real world visual data results in performance degradation. Instead synthetic images performed exceptionally. Specifically, on Peliculas, Violent Flow, Hockey and SCFD accuracy achieved is 100%, 81%, 97% and 75% respectively.

Additional Key Words and Phrases: Violence detection, Deep Learning, WVD, Hockey, Violent Flow, Peliculas, Synthetic violence

ACM Reference Format:
Muhammad Shahroz Nadeem, Fatih Kurugollu, Sara Saravi, Hany F. Atlam, and Virginia NL Franqueira. 2021. Effectiveness of Synthetic Images in Violence Detection. 1, 1 (October 2021), 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Synthetic data usage has seen a huge rise in the domain of computer vision. Many low-level and high-level tasks such as object classification and detection, semantic segmentation, bioinformatics and autonomous driving use synthetic images to train deep learning models [Nikolenko 2019]. Especially, when training deep networks, lack of real data can be compensated by using synthetic images. This approach makes creation, generation and collection of images easy. Moreover, once the framework for synthetic data has been developed; user have complete control over the parameters and truth label generation. This one time effort, results in generation of data examples which rarely occur naturally. Thus, a well balanced, diverse and automatically labelled dataset can be created. However, models trained using synthetic data still need to bridge the gap between synthetic and real [Tremblay et al. 2018]. Therefore, testing the performance of such models on real world data is mandatory.

Authors' addresses: Muhammad Shahroz Nadeem, M.Nadeem@derby.ac.uk, University of Derby, Derby, United Kingdom., DE223BL; Fatih Kurugollu, F.Kurugollu@derby.ac.uk, University of Derby, Derby, United Kingdom.; Sara Saravi, Loughborough University, Loughborough, United Kingdom; Hany F. Atlam, University of Derby, Derby, United Kingdom; Virginia NL Franqueira, University of Kent, Kent, United Kingdom.

2021. XXXX-XXXX/2021/10-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

As violence is a very subjective matter, and the available datasets are mostly created from corpus taken mostly from movies or YouTube to avoid ethical implications. Such datasets examples include: Hockey dataset [Nievas et al. 2011], Violent Flow [Hassner et al. 2012], Peliculas [Nievas et al. 2011], Real Life Violence Dataset (RLVD) [Soliman et al. 2019] and Violent Scene Detection (VSD) [Demarty et al. 2015]. Keeping these factors in mind, in this paper, firstly, we utilise synthetic images to perform a high level task of violence detection. Secondly, the trained model is tested on real world datasets to see if synthetic images can be effectively used for training deeper networks similar to other vision tasks. To the best of our knowledge, this is the first time synthetic data is been employed for violence detection. To this end, Weapon Violence Dataset (WVD) [Nadeem et al. 2019a] is utilised for violence detection. The WVD contains synthetic images for person-to-person fights using a range of Hot (pistols, shotguns and machine guns) and Cold (Knifes, Hatchet, and Hammer) weapons. Therefore, WVD consists of three classes; cold-violence, hot-violence and non-violence.

Following the footsteps of Soliman et al. [2019] in this paper, a VGG-16 [Simonyan and Zisserman 2014] based Long Short Term Memory (LSTM) network is utilised for violence detection. However, in-contrast to Soliman et al. [2019], instead of using Real Life Violence Dataset (RLVD) [Soliman et al. 2019], containing both person-to-person and crowd based scenarios. Our approach utilises only synthetic images taken from WVD with only person-to-person fights to train the network. Another, notable difference between the two datasets is the presence of weapons in WVDs' violent scenarios. Extensive experimentation shows that not only a synthetic dataset can be used for training deep models for violence detection, further, the learned features are quite superior to feature representations learned from real world images. This can be attributed to the fact that the quality of the synthetic images, with less noise and background clutter, good lighting conditions and high contrast aids in better features extractions [Dodge and Karam 2016]. Our experimentation validate the findings by da Costa et al. [2016], that less noisy data improves classification accuracy. Therefore the initial training is done using high quality synthetic images followed by noisy images from real world data during transfer learning. Specifically, we show that synthetic data been less noisy, facilitates the training process therefore achieving better classification performance. In summary we make the following contributions in this paper:

- We study and empirically evaluate the effect of synthetic images utilised on deep learning models for violence detection. Specifically, we compare the feature representation learned during training for synthetic and real datasets.

- We conduct extensive experimentation using standalone VGG-16 [Simonyan and Zisserman 2014] and additionally combining with LSTM [Hochreiter and Schmidhuber 1997] to test the performance on synthetic and real violence datasets.
- We provide a comparative analysis of the effectiveness of synthetic images for the problem of violence detection. To this end, we show the learned features representations from synthetic and real images by the proposed model.

The remaining of the paper is organised as follows. Section 2 gives a brief overview of the related approaches that are directly comparable and relevant to our study and experimentation strategy. Section 3 describes our approach consisting of VGG-16 only and VGG-16+LSTM experimentation strategy, Section 3.2 describes the synthetic and real world datasets utilised, Section 3.1 briefly explains the experimental environment, and in Section 3.3 presents the achieved performance through the use of synthetic images.

2 RELATED WORK

Violence detection is a high level computer vision task. Traditionally, considered a sub-branch of action recognition [Gao et al. 2016]. Due to which, initial action recognition methods were utilised for violence detection. Therefore, similar to action recognition, in addition to visual features, temporal and motion features are required to accurately classify violence. Furthermore, many actions can be performed solo. However,in case of violence usually the participation of at-least two individuals in required unless it is self inflicted violence.

Recently, due to advances in deep learning, many deep learning models for violence have been proposed. Soliman et al. [2019] proposed a ConvLSTM method, composed of pre-trained VGG-16 [Simonyan and Zisserman 2014] followed by LSTM units and a fully connected layers. Furthermore, a new dataset called RLVD has also been proposed containing 1000 violent and nonviolent videos. They trained the proposed network on Hockey [Nievas et al. 2011], Peliculas [Nievas et al. 2011] and Violent Flow [Demarty et al. 2014] datasets separately and reported the performance of their model. Furthermore, they trained the model on RLVD, and applied the learned features directly on the above mentioned datasets. They tried to solve the domain mismatch problem, however, according to them transfer learning is required to get a better performance.

Aktı et al. [2019] proposed a convolutional LSTM with attention for fight detection. In their experimentation, VGG-16 [Simonyan and Zisserman 2014], Xception [Chollet 2017] and FighCNN were utilised for features extraction from the images. FightCNN was trained on the Hockey dataset [Nievas et al. 2011]. However, instead of normal LSTM units they used Bi-directional LSTM units with additional backward flow of information [Schuster and Paliwal 1997]. Afterwards, an attention layer was added to the model to add inportance (weightage) to the learned features in hopes to achieve better performance. Furthermore, they proposed they collected fight sequences from surveillance footage, called Surveillance Camera Fight Dataset (SCFD). They tested their model on Hockey [Nievas et al. 2011], Peliculas [Nievas et al. 2011] and SCFD datasets[Aktı et al. 2019].

Sudhakaran and Lanz [2017] calculated frame difference of the images and passed them to a convolutional LSTM to classify violence on Hockey [Nievas et al. 2011], Peliculas [Nievas et al. 2011], and Violent Flow [Demarty et al. 2014]. According to them using frame difference strategy forces the network to learn the spatio-temporal features which results in better classification performance. Instead of VGG-16 [Simonyan and Zisserman 2014], AlexNet [Krizhevsky et al. 2012] is employed for feature extraction.

Peixoto et al. [2018] train a multi-model CNN network. They train individual CNN model for each violence category. In addition to raw (still) frames they pass three types of frames as input. The three frame types are called central, extremities, and averages combinations detected by Temporal Robust Features (TRoF) [Moreira et al. 2016]. Through these image combinations they remove the need for any recurrent layer to extract temporal features.

From the literature we can observe that most methods employ some form of pre-trained CNN network, VGG-16 been the favourite due to its faster features extraction and being computationally less expensive compared to its counterparts. Followed by LSTM units for extracting temporal features and fully connected layers for classification. Most, Violence detection methods are based on this deep architecture complemented with some changes in order to achieve better performance. Furthermore detailed survey for violence detection methods and datasets [Nadeem et al. 2019b] [Naik and Gopalakrishna 2017] [Yao and Hu 2021]

3 PROPOSED METHOD

The goal of this paper is to investigate that can synthetic images containing violence be utilised for training DL methods for violence detection. Further, what kind of features are learned with synthetic images and can we effectively replace real world violence images with synthetic once during the training stage thus, reducing the need to rely on YouTube or Hollywood movies for data acquisition. To this end, similar to Soliman et al. [2019], we utilised a pre-trained VGG-16 network for visual features extraction followed by LSTM units to compute temporal features and finally a fully connected layers for classification. The VGG-16 is encapsulated in a time distributed framework, so that multiple instances for consecutive images can be trained in parallel. Therefore instead of image-by-image, three consecutive frames are passed through the network in this manner.

Precisely in this paper, two sets of experiments were performed. In the first set, a single VGG-16 network is trained and tested on the WVD. The dataset is split and shuffled during the training process. Once trained the network is tested on real world datasets after applying two transfer learning strategies. We must emphasis here that Soliman et al. [2019] have already established that without transfer learning classification accuracy is very poor, due to this reason direct testing on real world data is not applied. However, still to test the learned features quality two transfer learning techniques were tested. 1) Only the last two convolutional layers are allowed to retrain called the freezed models denoted with “F” 2) The whole network is trainable called the not freezed models denoted by “NF”. The main reason for selecting these two strategies is to investigate and compare the quality of features learned from synthetic images

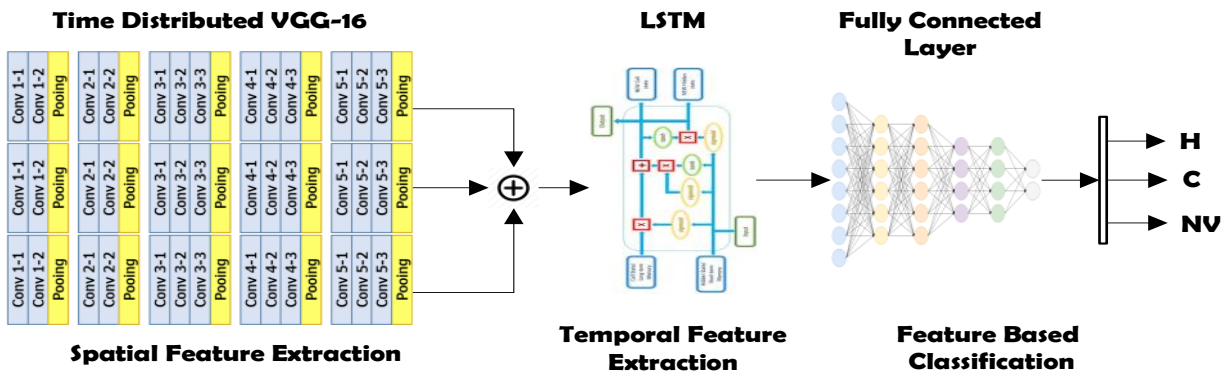


Fig. 1. The proposed architecture for the VGG-16 [Simonyan and Zisserman 2014] based convolutional LSTM [Hochreiter and Schmidhuber 1997] model with Time Distributed layer

against real world images. Furthermore, the generalisation ability of these features can also be tested.

In the second experimental set, LSTM units are added after VGG-16 to extract temporal features in addition to visual features. This DL model architecture is widely utilised in literature for violence detection [Aktı et al. 2019; Soliman et al. 2019; Sudhakaran and Lanz 2017]. Similarly, this model is also trained on the WVD, followed by transfer learning on the real world violence datasets. Figure 1 shows the model architecture used in the second set of experimentation process.

One major difference between both experimental sets is in regards to how images were passed to both the models. For the first set (VGG-16 only network), images were split into testing and training sets. The training images were shuffled and passed in a similar fashion as to a object classification model. Thus violence detection problem is intentionally treated as a object classification problem. The first reason behind this aspect was to test how synthetic images will perform on a simple CNN model for violence if no additional temporal or motion based information is passed. The second reason is to test either LSTM as claimed in literature does make a significant impact on classification performance specifically for violence detection.

For the second set (VGG-16+LSTM model), the images were batched together to a size of 16, with each batch containing 3 consecutive images. Depending upon the total number of frames extracted from the videos total number of batches differ for each video against each dataset used in the experimentation process. This enabled us to employ a time distributed VGG-16 followed by 64 LSTM units and fully connect layers. This experimental setup enabled us to gauge how LSTMs' would react to synthetic images and how significant is the increase in performance compared to VGG-16 only model. Therefore, providing us empirical prove if using synthetic images for violence detection is a feasible path or not. For both the experimental sets.

3.1 Experimental setup

The utilised model is created using Keras [Gulli and Pal 2017] with Tensorflow [Abadi et al. 2016] at backend, furthermore, OpenCV [Bradski and Kaehler 2008], Pandas [McKinney et al. 2011] and Matplotlib [Hunter 2007] are other libraries used in the experimentation process. In our experimentation many iterations of the VGG-16 only and VGG-16+LSTM models are trained on one synthetic and six real world datasets. Due to this reason the learning rate ranges from 1e-2 to 1e-6 for each dataset during transfer learning. The Adam [Kingma and Ba 2014] optimiser is selected as gradient decent for learning parameters and ReLU activation function is utilised (as shown in equation 1). Categorical cross entropy loss function (as shown in equation 2) is utilised, due to fact that WVD contains three class labels. For regularisation purposes, a range of different dropout [Srivastava et al. 2014] rates are selected, ranging from 0.3 to 0.8. Maximum training epochs were set to be 100.

The experimentation are performed on the Nvidia P4000 GPU with 8 GB RAM. Due to memory constraints, during the preprocessing stage the images are reshaped to a dimension of $(100, 160, 3)$ for VGG-16 only model. While for VGG-16+LSTM model the images are reshaped to $(100, 63, 3)$ from the original dimensions of $(800, 500, 3)$. Afterwards the images are normalised to value of zero to one. The dataset was divided into 70% training. 10% validation, 20% testing sets across all the datasets utilised in this work. Due to class imbalance for VGG-16 only and VGG-16+LSTM models, F1-score (shown in equation 3) is utilised to during training to for the six datasets. However, during testing, images extracted from the videos are passed to the models and based on voting mechanism the video is classified, therefore accuracy (shown in equation 4) is reported for testsets.

$$Relu(z) = max(0, z) \tag{1}$$

$$-\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}) \tag{2}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (4)$$

3.2 Violence Datasets

Violence is a sensitive topic and is often subjected to ethical issues while conducting research. Furthermore, its subjective nature makes it difficult to define [Demarty et al. 2015]. Due to these reason, only available datasets that contain violence, their data corpus is mainly taken mostly from movies, and YouTube videos. However, still there are limitations to the extent to which violence can be portrayed in these datasets. The reasons for this limitation include is associated with human psychology, as this can be negatively effected by continued longer exposure to such videos. Furthermore, consent is also required and facial identities (privacy) also needs to be protected of the people involved in violent videos. Therefore alot of such issues limit the volume and diversity of the violence detection datasets.

These factors have hampered the growth potential of violence detection method. As only a hand full of datasets are available in literature [Nadeem et al. 2019b]. Hockey [Nievas et al. 2011] and Peliculas [Nievas et al. 2011] datasets were compiled from videos taken from national hockey league and movies respectively. Hockey contains 1000 and Peliculas consists of 201 videos in total. Both of these datasets contain person-to-person fight sequences.

UNITN Social Interaction (USI) dataset [Rota et al. 2012] is another dataset which contains social interactions between two individuals. This relatively small dataset contains four human interactions, three are nonviolent "Talking", "Shaking" and "Hugging", while one interaction "Fighting" is violent. The videos are captured in a frontal/dash cam view in USI.

Surveillance Camera Fight Dataset (SCFD) [Aktı et al. 2019] is another fight dataset, captured with CCTV view of fights. Most fights are person-to-person with an exception of few group fights. The non violent videos are also taken from CCTV footage, resulting in total of 300 videos.

Violent Flow [Hassner et al. 2012] however is a crowd violence dataset containing a total of 246 videos. In contrast to above mentioned datasets huge number of participants is present in this dataset.

RVLD [Soliman et al. 2019] is a recently proposed violence dataset, containing both person-to-person and crowd based violence scenarios taken from YouTube. The dataset is diverse in terms of background, capturing view and person participation with a total of 2000 videos.

All of the above mentioned datasets contains videos of real world violence taken from Hollywood movies, or YouTube. Moreover, only hand-to-hand non lethal combat is shown with a limited threshold of blood and gore in these datasets. It is a fact, that human life is precious and real life weapon based violence is brutal and disturbing for viewers. To tackle these ethical and moral issues. The synthetic data route seems logical and necessary. Keeping these factors in mind, Nadeem et al. [2019a] proposed the WVD based on the photo realistic game Grand Theft Auto (GTA)-V. It contains two violent and one non violent classes. The violent category contains person-to-person fight using cold (bat, hammer, knuckleduster etc) and hot (piston, sub-machine and machine guns) weapons. While the non violent category contain normal action present in the games framework (exercise, arguments, dancing etc).

In our experimentation process, training, validation and testing is first performed on WVD, afterwards testing is performed on all the above mentioned datasets. However, as proved by Soliman et al. [2019] without transfer learning no meaning full accuracy can be achieved. So before testing on real world violence datasets transfer learning is applied. Figure 2 shows the frames taken from each dataset. Furthermore, the camera angle of capture, background scene, number of participants, image quality (i.e noise and blur) and data corpus utilised for building the dataset can be viewed.

3.3 Experimental Results

As mentioned in Section 3, the first set of experiments are conducted using only the VGG-16 network. Table 1 shows the achieved accuracy of the model on both the transfer learning strategies over the six real world violence datasets. Here, it can be seen that both the "F" and "NF" transfer learning strategies have very similar performance. This is better elaborated in the confusion matrices as shown in Figure 3. Each dataset has two confusion matrix for each transfer learning strategy i.e. "F" and "NF". The confusion matrices show the performance of the model against each dataset utilised in the experimentation. The achieved normalised accuracy per class can be observed. In particular, it can be seen that USINF model has overfitted on the non violent class. Overall, for the smaller datasets (Peliculas, USI) the "F" strategy performs better, as it is less prone to overfitting. However, the overall performance difference is not significant.

It can be observed that even without the LSTM units i.e. the temporal component; the accuracy achieved overall for majority of the violence datasets is quite high. Especially, on the Peliculas, USI, Hockey and RLVD with maximum accuracies of 97.5%, 100%, 95%, and 91%. However it must be noted that, USI and Peliculas are relatively smaller datasets. This performance is surprising as most violence detection methods employ some form of recurrent unit to include the temporal aspect which is proved to facilitate better performance by learning video representation [Srivastava et al. 2015]. It must be noted that the basic parameters for the model were fixed in each iteration except dropout and learning rate. Overall the performance margin is not quite significant between all "F" models to its counterpart "NF" models. Moreover, for SCFD the freezed models achieved accuracy of 68% higher then 66%. Peliculas and RLVD had similar performance with both transfer learning strategies with maximum accuracies of 97.5% and 91% for both "F" and "NF". This shows that the weights learned during the WVD

Datasets	Peliculas		USI		Violent Flow	
TL Strategy	F	NF	F	NF	F	NF
F1-Score	97.5	97.5	100	80	80	82
Datasets	Hockey		SCFD		RLVD	
TL Strategy	F	NF	F	NF	F	NF
F1-Score	93	95	68	66	91	91

Table 1. The Table shows the accuracy achieved by VGG-16 on the violence datasets.

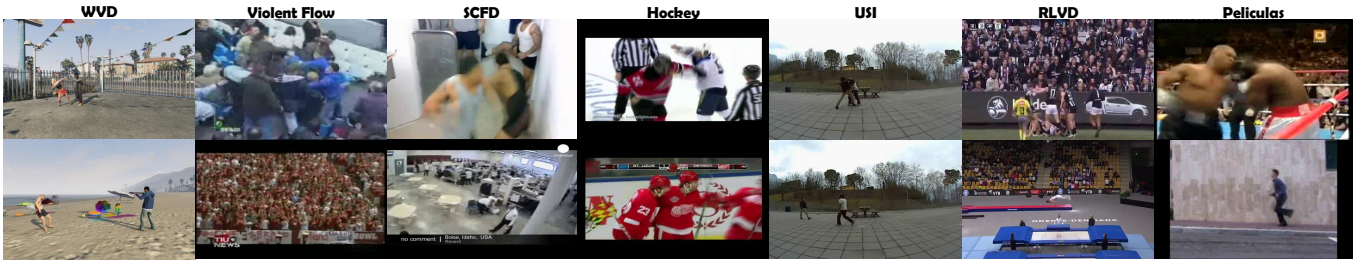


Fig. 2. The figure shows the video frames of the seven datasets utilised in the experimentation.

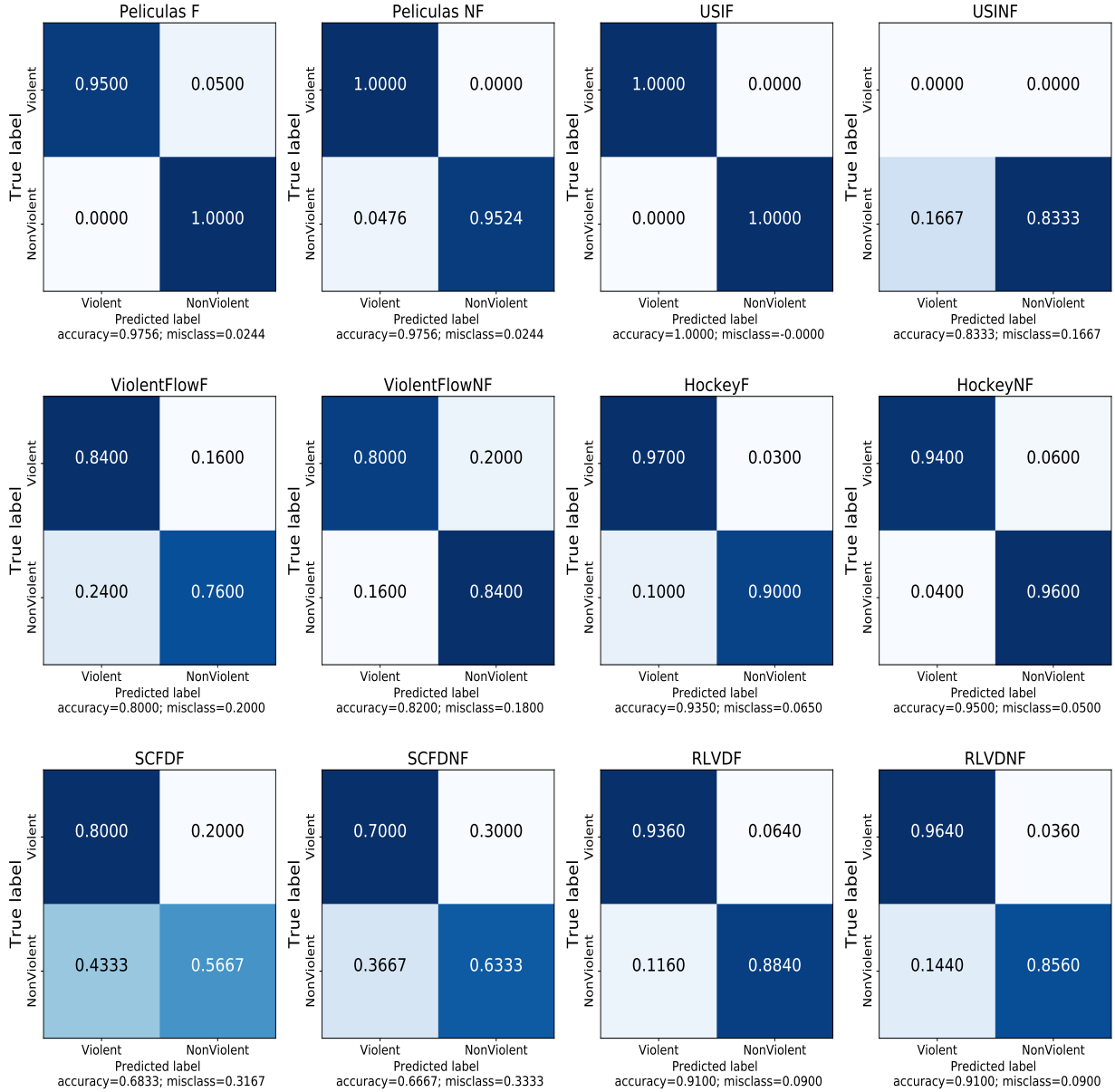


Fig. 3. The confusion matrix shows per class performance against each transfer learning strategy for every dataset used in the experimentation process.

training phase are capable to extract meaningful features without any form of retraining. However, the fully connected layers needed to be retrained.

Direct comparison with literature for VGG-16 only experiments is not possible, as majority of the violence methods include LSTM units to compute the temporal and spatial features of the videos in the datasets. Therefore, for fair comparison of the performance of synthetic images on the model LSTM units were added to VGG-16. This initiated the second set of experiments, and the performance of the model with synthetic images is compared with the models proposed by Soliman et al. [2019], [Sudhakaran and Lanz 2017] [Aktı et al. 2019] as shown in Table 2.

However, it must be noted that the methods of Aktı et al. [2019], Sudhakaran and Lanz [2017] and Soliman et al. [2019] were directly trained and tested on the mentioned datasets. Furthermore, Soliman et al. [2019] trained their network on their proposed RLVD dataset and applied transfer learning to test the models generalisation power. Following this example, our model was trained first on WVD and afterwards tested on the datasets through transfer learning.

The first method in Table 2 by Aktı et al. [2019] trained two model configurations. First is the vanilla VGG-16+LSTM, second is their own designed model which comprises of Xception model, for features extraction, Bi-directional LSTMs for temporal feature extraction followed by attention layer. According to them, number of frames has no significant impact on the performance of the model. They experimented with five and ten frames per video. Their models were trained directly on Peliculas, Hockey and SCFD datasets. Achieving max accuracies of 100%, 92%, 62% for VGG-16+LSTM and 100%, 98%, 72% for their proposed model as shown in Table 2. Aktı et al. [2019] achieved better performance by using less number of frames per video.

Similarly, Sudhakaran and Lanz [2017] trained a custom convolutional LSTM instead of VGG-16 based method and directly trained their model on the Peliculas, Violent Flow and Hockey datasets shown in Table 2. They achieved very high classification accuracies of 100%, 94.57% and 97.1% respectively. Both Aktı et al. [2019] and Sudhakaran and Lanz [2017] trained the models directly over the datasets of their choice no transfer learning was applied.

Soliman et al. [2019] first trained their model directly on Peliculas, Violent Flow and Hockey datasets. Afterwards, the model was trained on their own proposed dataset RLVD. To test the generalisation of their method, they applied transfer learning. Table 2 shows the performance of both the strategies with 99%, 90.01% and 95.1% on peliculas, violent flow and hockey datasets. However, when they train the network on RLVD and apply transfer learning on these datasets. The resulted performance degrades to 88.2%, 84% and 86.16%.

Vanilla VGG-16+LSTM trained by Aktı et al. [2019] and Sudhakaran and Lanz [2017] achieved highest accuracy of 100% and 99% on the Peliculas. Whereas 92% and 95.1% on the hockey dataset respectively. Furthermore, Sudhakaran and Lanz [2017] custom Convolutional LSTM model also achieved great performance. Therefore, we can conclude that training VGG-16+LSTM directly on these datasets would give similar performance. Hence, direct training was skipped in our experimentation process. Instead the focus is brought to how effective is synthetic data for violence and how good is their generalisation power.

Keeping, this in mind, first our method is trained on the WVD. The training accuracy achieved on the WVD is 94%. Afterwards transfer learning is applied, Table 2 shows that our approach performed better then Soliman et al. [2019] on Peliculas and Hockey datasets with accuracies of 100%, and 96.50%. However, on Violent Flow dataset performance achieved is 81% slightly below 84%. However, it must be noted that WVD is a person-to-person fight dataset, however, Violet Flow is crowd based. Keeping this in mind, the performance is quite good. On the SCFD the transfer learned model performed even better then directly trained model model of Aktı et al. [2019]. In contrast to 72%, with synthetic images we achieved 75% accuracy.

Overall, this shows that using synthetic images for violence gives better performance using the vanilla VGG-16+LSTM network configuration , however, if the network is changed the performance with real world data is slightly better as shown in 2. Our experiments prove that synthetic images are as good as real world violence images and thus they can be effectively used to train larger and diverse deep learning networks. Furthermore, as the synthetic images generally been of higher quality, sharpness captured under better lighting condition, facilitate quick and better features representation in comparison to real world images that are often of poor quality contain different degrees of blur. Therefore, initial training with synthetic data followed by transfer learning on real world images can effectively solve the data shortage problem in violence detection domain. Moreover, synthetic images such as in WVD are not subjected to ethical implications, which makes it promising for future research potential.

3.4 Result Discussion

To better understand the increase in achieved performance using a vanilla VGG-16+LSTM model with synthetic images. In this section, we will dissect and observe the learned feature representations for all the violence datasets utilised in this work. It is a known fact that DL models are black box [Nadeem et al. 2019b; Rudin 2019] in nature. DL methods perform millions of mathematical linear and non linear transformations during the training process. Making sense of such transformations and fundamentally asking why a particular transformation was applied is very difficult for humans, due to which DL methods are considered black-box. Due to this methods are now being proposed for "Explainable AI" [Arief et al. 2021; Gaur et al. 2021]. However, such methods have not gained enough confidence and often such methods are unreliable.

However, observing the activation of feature maps is one way to look at what the network has learned during the training process [Zeiler and Fergus 2014]. Therefore, we extracted the output after the activation feature maps of the last convolution layer from the VGG-16+LSTM model against each dataset. Figure 4-6 shows activation maps for the WVD dataset. At an initial glance it appears that CNN learned the motion representation for the video. Upon visual inspection of these activation maps, it is certain that for the "No Violence" the images contain singular patterns that are monotonic, and less noisy (shown in Figure 4). This can be due to the fact

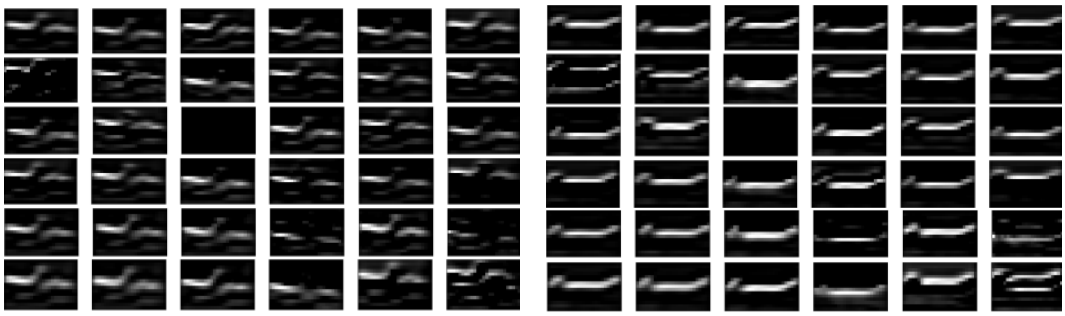


Fig. 4. Activation map outputs for No Violence class for WVD

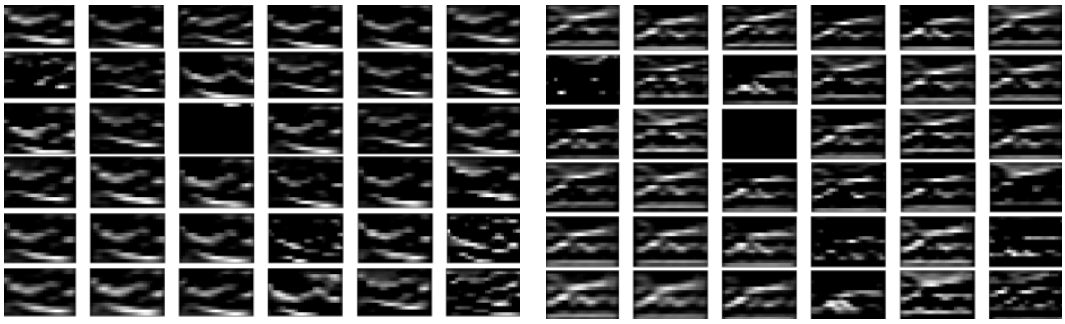


Fig. 5. Activation map outputs for hot Violence class for WVD

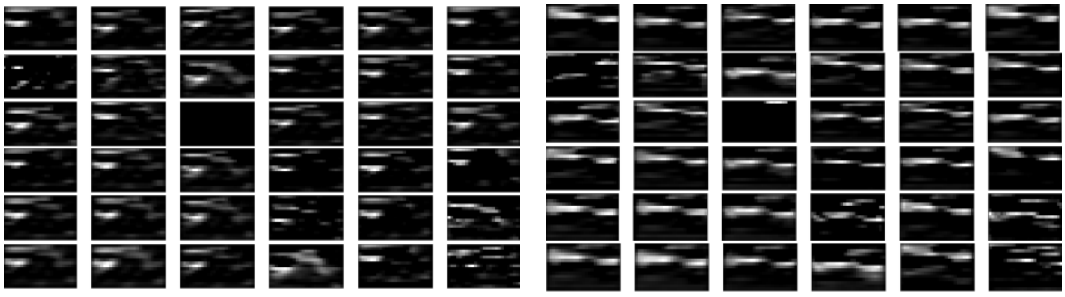
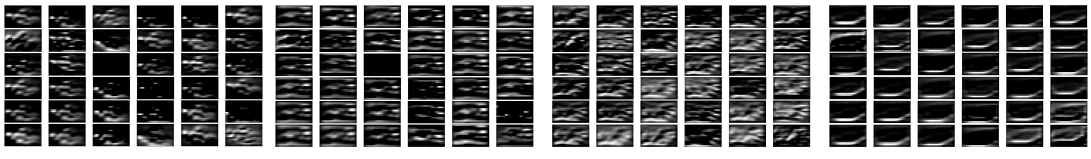
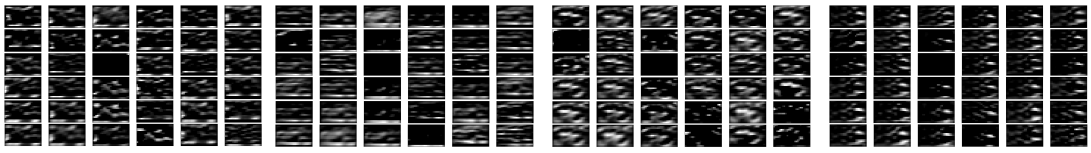


Fig. 6. Activation map outputs for cold Violence class for WVD



(a) Peliculas Violence (b) Peliculas No Violence (c) Hockey Violence (d) Hockey No Violence



(e) Violent Flow Violence (f) Violent Flow No Violence (g) SCFD Violence (h) SCFD No Violence

Fig. 7. Activation maps output for the real world datasets.

Table 2. The table shows performance of our approach against convolutional LSTM methods for violence detection. It is to be noted that the performance of our transfer learned model is compared against methods which are directly trained on real world datasets and VGG-16+LSTM method transfer learned on real world dataset.

Authors	Model Configuration	Peliculas		Violent flow	Hockey		SCFD	
Aktı et al. [2019]		10 frames	5 frames		10 frames	5 frames	10 frames	5 frames
	VGG-16+LSTM (Direct Training)	95	100		87.05	92	62	61.67
	Xception+bi-LSTM+Attention (Direct Training)	100	100		97.5	98	71	72
Sudhakaran and Lanz [2017]	Custom CNN+LSTM (Direct Training)	100		94.57	97.1			
Soliman et al. [2019]	VGG-16+LSTM (Direct Training)	99		90.01	95.1			
	VGG-16+LSTM (transfer learning RLVD)	88.2		84	86.16			
Our model	VGG-16+LSTM (transfer learning WVD)	100		81	97.00		75	

that for this class videos of WVD have less change in motion. In contrast to this, "Hot violence" activation maps appear very noisy and slightly less monotonic. The features are cluttered allover the images (shown in Figure 5). However, "Cold violence" activation maps seems to be a hybrid between the "No" and "Hot" violence class. These images show that WVD based activation maps are visually distinct to some aspect. However, this hypothesis is based on visual inspection.

To test if this hypothesis hold true for real world datasets, activation maps outputs were extracted for Peliculas, Hockey, SCFD and RLVD datasets after transfer learning as shown in Figure 7. However, it must be noted that these datasets contain two classes: violent and non-violent. Hockey and SCFD have similar activation map outputs to WVD. The no violence feature maps shown in Figure 7d and 7h are monotonic and less cluttered. While the violent class feature maps are more cluttered (shown in Figure 7c and 7g). However, it must be noted that Hockey dataset has higher test accuracy then SCFD still the activation feature maps are similar in nature. Peliculas and Violent Flow datasets however, go against this hypothesis. This can be observed in Figure 7a, 7b, 7e and 7f, the activation have produced evenly cluttered outputs, Peliculas has 100% accuracy however, this is a comparatively smaller dataset, while for Violent Flow the accuracy achieved is slightly low. In comparison to synthetic images, activation map output of the real world images are more cluttered. This shows that synthetic images contain less noise, therefore, better activation maps are achieved. Due to this reason, better classification accuracy is achieved in this work by using synthetic images.

4 CONCLUSION

In this paper, we show empirically that synthetic images can be effectively utilised during training to replace real world data for violence detection. Using synthetic images have inherent advantages which include: less noise, high sharpness, contrast and brightness. In this paper our experimental results indicate that synthetic images produce better feature representations. The activation maps show the superiority of synthetic images in contrast to output activation maps of real world violence images. To this end, initially we trained a VGG-16 network on WVD, followed by two transfer learning strategies on real world violence datasets which include: Hockey, Peliculas, USID, Violent Flow, SCFD and RLVD. Afterwards, to integrate the temporal component LSTM units were added to VGG-16 network. This new network configuration is then trained on WVD, followed by transfer learning on real world datasets.

The transfer learned models show that using the vanilla VGG-16 and VGG-16+LSTM models outperform similar models that use real images instead of synthetic ones on Peliculas, Hockey and SCFD datasets. However, for Violent Flow the achieved accuracy is lower. Precisely, accuracy achieved on Peliculas, Violent Flow, Hockey and SCFD is 100%, 81%, 97% and 75% respectively. More broadly, our paper proves that synthetically generated images are effective replacement against real violence datasets. Furthermore, such data is not subjected to ethical implications and therefore, can compliment data shortage problem.

REFERENCES

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*. 265–283.

Şeymanur Aktı, Gözde Ayşe Tataroğlu, and Hazım Kemal Ekenel. 2019. Vision-based fight detection from surveillance cameras. In *2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA)*. IEEE, 1–6.

Mansur Arief, Zhiyuan Huang, Guru Koushik Senthil Kumar, Yuanlu Bai, Shengyi He, Wenhao Ding, Henry Lam, and Ding Zhao. 2021. Deep Probabilistic Accelerated Evaluation: A Robust Certifiable Rare-Event Simulation Methodology for Black-Box Safety-Critical Systems. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 595–603.

Gary Bradski and Adrian Kaehler. 2008. *Learning OpenCV: Computer vision with the OpenCV library*. " O'Reilly Media, Inc.".

François Chollet. 2017. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1251–1258.

Gabriel B Paranhos da Costa, Welinton A Contato, Tiago S Nazare, João ES Neto, and Moacir Ponti. 2016. An empirical study on the effects of different types of noise in image classification tasks. *arXiv preprint arXiv:1609.02781* (2016).

C. Demarty, B. Ionescu, Y. Jiang, V. L. Quang, M. Schedl, and C. Penet. 2014. Benchmarking Violent Scenes Detection in movies. In *2014 12th International Workshop on Content-Based Multimedia Indexing (CBMI)*. 1–6.

Claire-Hélène Demarty, Cédric Penet, Mohammad Soleymani, and Guillaume Gravier. 2015. VSD, a public dataset for the detection of violent scenes in movies: design, annotation, analysis and evaluation. *Multimedia Tools and Applications* 74, 17 (2015), 7379–7404.

Samuel Dodge and Lina Karam. 2016. Understanding how image quality affects deep neural networks. In *2016 eighth international conference on quality of multimedia experience (QoMEX)*. IEEE, 1–6.

Yuan Gao, Hong Liu, Xiaohu Sun, Can Wang, and Yi Liu. 2016. Violence detection using oriented violent flows. *Image and vision computing* 48 (2016), 37–41.

Manas Gaur, Keyur Faldu, and Amit Sheth. 2021. Semantics of the Black-Box: Can knowledge graphs help make deep learning systems more interpretable and explainable? *IEEE Internet Computing* 25, 1 (2021), 51–59.

Antonio Gulli and Sujit Pal. 2017. *Deep learning with Keras*. Packt Publishing Ltd.

Tal Hassner, Yossi Itcher, and Orit Kliper-Gross. 2012. Violent flows: Real-time detection of violent crowd behavior. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 1–6.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

John D Hunter. 2007. Matplotlib: A 2D graphics environment. *Computing in science & engineering* 9, 03 (2007), 90–95.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25 (2012), 1097–1105.

Wes McKinney et al. 2011. pandas: a foundational Python library for data analysis and statistics. *Python for high performance and scientific computing* 14, 9 (2011), 1–9.

Daniel Moreira, Sandra Avila, Mauricio Perez, Daniel Moraes, Vanessa Testoni, Eduardo Valle, Siome Goldenstein, and Anderson Rocha. 2016. Pornography classification: The hidden clues in video space–time. *Forensic science international* 268 (2016), 46–61.

Muhammad Shahroz Nadeem, Virginia NL Franqueira, Fatih Kurugollu, and Xiaojun Zhai. 2019a. WVD: A New Synthetic Dataset for Video-Based Violence Detection. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence*. Springer, 158–164.

Muhammad Shahroz Nadeem, Virginia NL Franqueira, Xiaojun Zhai, and Fatih Kurugollu. 2019b. A survey of deep learning solutions for multimedia visual content analysis. *IEEE Access* 7 (2019), 84003–84019.

Anuja Jana Naik and MT Gopalakrishna. 2017. Violence detection in surveillance video a survey. *International Journal of Latest Research in Engineering and Technology (IJLRET)* (2017), 11–17.

Enrique Bermejo Nievas, Oscar Deniz Suarez, Gloria Bueno García, and Rahul Sukthankar. 2011. Violence detection in video using computer vision techniques. In *International conference on Computer analysis of images and patterns*. Springer, 332–339.

Sergey I Nikolenko. 2019. Synthetic data for deep learning.

Bruno Malveira Peixoto, Sandra Avila, Zanoni Dias, and Anderson Rocha. 2018. Breaking down violence: A deep-learning strategy to model and classify violence in videos. In *Proceedings of the 13th International Conference on Availability, Reliability and Security*. 1–7.

Paolo Rota, Nicola Conci, and Nicu Sebe. 2012. Real time detection of social interactions in surveillance video. In *Computer vision–ECCV 2012. Workshops and demonstrations*. Springer, 111–120.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing* 45, 11 (1997), 2673–2681.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

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

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15, 1 (2014), 1929–1958.

Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. 2015. Unsupervised learning of video representations using lstms. In *International conference on machine learning*. PMLR, 843–852.

Swathikiran Sudhakaran and Oswald Lanz. 2017. Learning to detect violent videos using convolutional long short-term memory. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 1–6.

Jonathan Tremblay, Aayush Prakash, David Acuna, Mark Brophy, Varun Jampani, Cem Anil, Thang To, Eric Cameracci, Shaad Boochoon, and Stan Birchfield. 2018. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 969–977.

Huiling Yao and Xing Hu. 2021. A survey of video violence detection. *Cyber-Physical Systems* (2021), 1–24.

Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*. Springer, 818–833.