electronics

MDPI

Article

VioleNet: Dense Multi-Head Self-Attention with Bidirectional Convolutional LSTM for Detecting Violence

Fernando J. Rendón-Segador [1,*], Juan A. Álvarez-García [1], Fernando Enríquez [1] and Oscar Deniz [2]

[1] Departamento de Lenguajes y Sistemas Informáticos, Universidad de Sevilla, 41012 Sevilla, Spain; jaalvarez@us.es (J.A.Á.-G.); fenros@us.es (F.E.)
[2] VISILAB, E.T.S.I.I, University of Castilla-La Mancha, 13071 Ciudad Real, Spain; oscar.deniz@uclm.es
* Correspondence: frendon@us.es; Tel.: +34-628-555-881

Abstract: Introducing efficient automatic violence detection in video surveillance or audiovisual content monitoring systems would greatly facilitate the work of closed-circuit television (CCTV) operators, rating agencies or those in charge of monitoring social network content. In this paper we present a new deep learning architecture, using an adapted version of DenseNet for three dimensions, a multi-head self-attention layer and a bidirectional convolutional long short-term memory (LSTM) module, that allows encoding relevant spatio-temporal features, to determine whether a video is violent or not. Furthermore, an ablation study of the input frames, comparing dense optical flow and adjacent frames subtraction and the influence of the attention layer is carried out, showing that the combination of optical flow and the attention mechanism improves results up to 4.4%. The conducted experiments using four of the most widely used datasets for this problem, matching or exceeding in some cases the results of the state of the art, reducing the number of network parameters needed (4.5 millions), and increasing its efficiency in test accuracy (from 95.6% on the most complex dataset to 100% on the simplest one) and inference time (less than 0.3 s for the longest clips). Finally, to check if the generated model is able to generalize violence, a cross-dataset analysis is performed, which shows the complexity of this approach: using three datasets to train and testing on the remaining one the accuracy drops in the worst case to 70.08% and in the best case to 81.51%, which points to future work oriented towards anomaly detection in new datasets.

Keywords: violence detection; fight detection; deep learning; dense net; bidirectional ConvLSTM

Citation: Rendón-Segador, F.J.; Álvarez-García, J.A.; Enríquez, F.; Deniz, O. VioleNet: Dense Multi-Head Self-Attention with Bidirectional Convolutional LSTM for Detecting Violence. *Electronics* **2021**, *10*, 1601. https://doi.org/10.3390/electronics10131601

Academic Editor: Juan M. Corchado

Received: 24 May 2021
Accepted: 30 June 2021
Published: 3 July 2021

1. Introduction

In recent years, the problem of recognizing human action in videos has gained importance in the field of computer vision [1–3], but the detection of violent behavior has been comparatively less studied than other human actions. However, the detection of violence has great applicability in both public and private security. Today there are surveillance cameras nearly everywhere, especially in schools, prisons, hospitals, shopping centers, etc. The availability of this growing number of cameras requires sufficient human resources to monitor the large volume of images they generate. This is normally not possible, losing much of the potential they offer.

Problems arise such as lack of personnel and overlooked threats, given that after 20 min of monitoring a CCTV system, operators fail to detect objects or activities [4,5]. This calls for innovation in automated systems, optionally combined with human attention, for the detection of violent actions [6–8] or gun detection [9–11]. Safety in all areas of daily life has always been a general concern over time and in all parts of the world. The current socio-economic differences and the world economic crisis have led to an increase in violent [12] and the recording and dissemination of such crimes. That is why it is imperative to develop automated methods to detect these actions and improve the responsiveness of security teams, the review of audiovisual content for age rating and content control in social networks.

This work is focused on intentional violence, especially in person-to-person violence, not considering unintentional actions such as traffic accidents or playing rough in sports. To visually recognize violence in videos, it is necessary to perform a spatio-temporal analysis since there are sudden movements associated with blows and punches. Violent actions can be confused with other types of actions causing false positives. For instance a video with little motion, sudden movements such as those produced during cardiopulmonary resuscitation can be misinterpreted as punching, causing a false positive. To avoid cases like that, it is necessary to analyze the temporal context of the entire video, before the action and after the action.

Although there is an extensive body of research dealing with the problem, as far as we know, there is currently no commercial system that applies artificial intelligence and human operators jointly to detect violence. This is necessary from a job quality perspective. In this way, the stress of the operators who have to watch this type of video can be avoided so that in some cases they can dedicate themselves to more productive tasks. Above all however, it is a matter of inability to carry out the work 100% correctly and efficiently, due to a simple limitation regarding the number of videos that can be watched simultaneously and with the necessary attention. The main obstacles to achieving a video monitoring system capable of automatically detecting violence effectively are the number of false positives that lead to the deactivation of the system and the occurrence of false negatives that imply failures in the functionality. This is why a more accurate system, tested in multiple datasets, is proposed. In addition, violence can be of many types and it is difficult to obtain a generalization from a single dataset. In this sense, we consider that it is essential to perform a cross-dataset analysis to verify if a model trained on one dataset can obtain good results on a different one, to determine whether it is feasible to put the model into production or it is necessary to refine it for each case.

This study presents the next contributions:

1. An architecture based on existing blocks such as 3-dimension DenseNet, multi-head self-attention mechanism, and bidirectional convolutional LSTM, trained to detect violent actions in videos.
2. An analysis of the input format (optical flow and adjacent frames subtraction) and their influence in the results.
3. An experimentation with four datasets in which the state of the art for violence detection is improved.
4. A cross-dataset analysis to check the generalization of the concept of violent actions.

This paper is organized as follows. In Section 2 a brief study on the state of the art of the problem is carried out. Section 3 provides in-depth details on the proposed model architecture. Section 4 describes in detail each of the datasets used in this work to evaluate the model. Section 5 summarizes the experiments carried out and the methodology used. In Section 6 experimental results are presented. Section 7 shows the importance and relevance of the obtained results and summarizes the strengths and weaknesses of the proposed model. Finally in Section 8 the main conclusions and possible lines of progress for future development are exposed.

2. Related Work

In this section, a review of the state of the art on the detection of violent actions based on visual features is carried out. Of all the references, special emphasis is placed on those using deep learning techniques which in general have achieved better results, something that has made them the most widely used. However, the automatic feature extraction provided by deep learning methods also means less control and higher difficulty in obtaining explainable models, something simpler using more traditional methods that involved manual and generally complex feature extraction.

2.1. Non-Deep Learning Methods

State of the art based only in visual features has evolved from local methods based on Kohonen self-organizing map to distinguish blood [13], to motion descriptors [14–16]. For instance, [17] considered space-time interest points (STIPs) and the SIFT extension (MoSIFT) algorithms as spatio-temporal descriptors, representing each video using a bag-of-features and classifying them using support vector machines (SVM) with different kernels. There are other space-time descriptors that represent each video using bag-of-features based on vectors of interest instead of points, such as Fisher vectors [18,19]. Unfortunately, under bag-of-features frameworks the computational cost of spatio-temporal features extraction is very high and not useful for real-time applications. In recent years, novel feature are being explored such as Lagrangian direction fields [20], motion weber local descriptor [21], features from motion blobs after binarizing the absolute difference between consecutive frames [22] or low-level visual features such as local histogram of oriented gradient [23] and local histogram of optical flow [24].

Other methods are focused specifically on optical flow [25,26]. The latter used a Gaussian model of optical flow to obtain regions that prepare a descriptor (orientation histogram of optical flow) that is classified by a SVM. This approach was further developed with new feature descriptors such as histogram of optical flow magnitude and orientation [27]. Non-deep learning methods that mark the state of the art employ Dense Trajectories for spatio-temporal feature extraction, Fisher vector in feature coding and SVM for classification [18,19].

2.2. Deep Learning Methods

Deep learning methods began to be used in 2014 to perform action recognition [28,29]. The specialization in violence recognition came with convolutional networks [30] that analyze visual spatio-temporal features and LSTM layers to encode temporal features [31].

The following works are described in depth given the importance to understand their architectures and determine the influence of each block in the results.

In [32] a network composed of convolutional, batch normalization and pooling layers followed by reduction operations to extract spatial features was proposed. Then a recurrent convolutional layer (ConvLSTM) is used to encode the level changes of the frames or temporal features, which characterize violent scenes. After all the frames have been applied sequentially, the classification is carried out with fully connected layers. Authors use the AlexNet network [33] pre-trained with ImageNet as the CNN model to extract features from frames.

In [34] the authors proposed a model divided into a spatial encoder, a temporal encoder and a classifier. Subtraction operation between frames is used as input for the next block, the spatial encoder, that consists of a modified version of a convolutional neural network VGG13 [35]. As a result, feature maps of spatial features are obtained for each frame. This information is passed to the temporal encoder, a bidirectional convolutional LSTM layer (BiConvLSTM). An element-wise maximum operation is used to combine the results in a rendering of the video. Finally, this representation is applied to a classifier to determine whether or not it is a video with violent action.

In [36] the Xception Bi-LSTM Attention model was proposed. This model first performs an uniform frame extraction process obtaining between 5 and 10 frames from each video. Then, a CNN network based in the Xception architecture called Fight-CNN with an expanded kernel size is used to extract and capture spatial features of violent scenes. After that, a bidirectional LSTM layer is used as it can learn the dependency between past and current information including temporal features. Finally, in order to distinguish important parts, an attention layer [37] is incorporated. This combination, the attention layer in conjunction with bidirectional LSTM layers, determines if the video contains violence or not.

The models described so far are sequential one-stream models. That is, input frames are in the RGB format or their optical flow. However, there are other models that use both

formats and more as inputs. These are convolutional multi-stream models in which each stream analyzes a type of video feature. In [38] the authors presented a multi-stream model called FightNet with three types of input modes, that is, RGB images, optical flow images, and acceleration images, each one with an associated network. The input video is divided into segments and for each segment the feature maps of each stream are obtained. Finally, all these feature maps are merged. The final output of the network is the average score of each segment of the video.

Our architecture is based on the recurrent convolutional architecture with attention mechanism. The most similar work to our proposal is that of [36], but we improve it on the following relevant points: (1) the optical flow of the video as input to the network instead of the RGB format, (2) the inclusion of the DenseNet architecture adapted to three dimensions, using 3D convolutional layers instead of 2D ones and (3) the use of multi-head self-attention layer [39] instead of attention mechanisms [37], creating a novel architecture for the detection of violence. Once the new model is obtained, a cross-dataset experiment is carried out to check whether it is capable of generalizing violent actions. The following section describes it in detail.

3. Model Architecture

To correctly classify violence in videos, the generation of a robust video encoding is fundamental to later classify it using a fully connected network. To achieve it, each video is transformed from RGB to optical flow. Then a Dense network is used to encode the optical flow as a sequence of feature maps. These feature maps are passed through a multi-head self-attention layer and then through a bidirectional ConvLSTM layer to apply the attention mechanism in both temporal directions of the video (forward and backward pass). This spatio-temporal encoder with attention mechanism extracts relevant spatial and temporal features for each video. Finally, the encoded features are fed into a four-layer classifier that classifies the video into two categories (violence and non-violence).

The architecture of the model, called ViolenceNet, is shown in Figure 1.

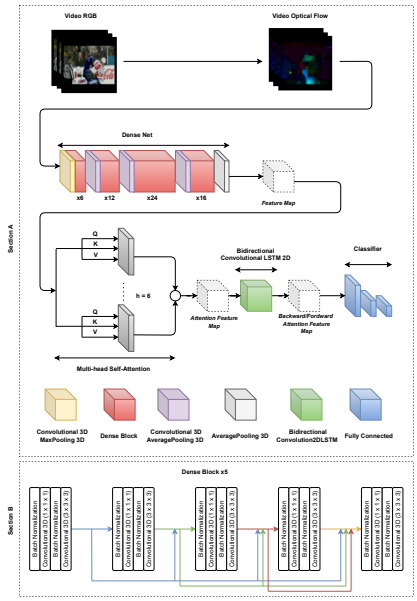

Figure 1. In Section **A** the architecture of the ViolenceNet model, that takes the optical flow as input, is shown. It is composed of four parts: a DenseNet-121 network spatio-temporal encoder, a multi-head self-attention layer [39], a bidirectional convolution 2D LSTM (BiConvLSTM2D) layer and a classifier. Below each Dense Block its number of components is indicated. The variable **h** corresponds to the number of heads used in parallel by the multi-head self-attention layer and the variables **Q**, **K**, **V** their inputs. Section **B** shows the internal architecture of a five-component Dense Block (X5).

3.1. Model Justification

The blocks that compose the architecture of the model have been successfully tested in the field of human action recognition and, specifically, in the field of violent actions, as shown in Section 2.2. In addition, 3D DenseNet variant has been used for video classification [40], bidirectional recurrent convolutional block convolutional improved the efficiency in detecting violent actions [34], as it allows analyzing features in both temporal directions. Attention mechanisms have been used in recent years with good results in the task of recognizing human actions [41,42] and the combination of convolutional networks and bidirectional convolutional recurrent blocks has already shown efficiency in learning spatial and temporal information in videos [43]. These facts prove that the model is based on blocks that are useful to recognize human actions in videos and have led us to use them to develop our proposal.

The following subsections describe the architecture of the ViolenceNet model in detail.

3.2. Optical Flow

One of the inputs of our network is the dense optical flow [44]. This algorithm generates a sequence of frames where those pixels that move the most between consecutive frames are represented with greater intensity. This information is a key element in violent scenes since the most important components are contact and speed: pixels tend to move more during that segment of the video than in the rest and tend to cluster in one area of the scene.

Once applied the algorithm, a 2-channel matrix with optical flow vectors that include magnitude and direction, is obtained. The direction corresponds to the hue value of the image, while the magnitude corresponds to the value plane. The hue value is used for visualization only.

We selected dense optical flow over sparse optical flow because the former provides the flow vectors for the full-frame, up to one flow vector per pixel, while sparse flow only provides the flow vectors some "interesting features", such as some pixels that represent the edges or corners of an object within the frame. In deep learning models, like the proposed one, the selection of features is unsupervised, therefore it is better to have a wide range of features than a limited one. Mainly for this reason dense optical flow is used as input to the model.

3.3. DenseNet Convolutional 3D

The DenseNet architecture [45] was designed to be used with images so it is composed of 2D convolutional layers. However, it can be adapted so that it can process videos. Two modifications to the original have been made: the first one is to replace the 2D convolutional layers with 3D ones; the second, to replace the 2D reduction layers with 3D reduction ones. DenseNet uses the reduction layers MaxPool2D and AveragePool2D with a pool size of *(2, 2)* and *(7, 7)*. The reduction layers MaxPool3D and AveragePool3D were used with a pool size of *(2, 2, 2)* and *(7, 7, 7)*.

DenseNet takes its name from the dense blocks that are the foundation of the network architecture. These blocks concatenate the feature maps of a layer with all of its descendants. In our proposal, four dense blocks have been used in total, each one with a different size. Each dense block is made up of a series of layers that follow the sequence: batch normalization-convolutional 3D-batch normalization-convolutional 3D as it can be seen in Figure 1 (Section B).

The DenseNet model has been chosen because of the way in which it concatenates the feature maps, which is simpler than models like Inception [46] or ResNet [47]. Its architecture is more robust and requires a very low number of filters and parameters to achieve high efficiency, unlike other models [45]. It has been used successfully for the treatment of biomedical images [48,49], showing better results than other architectures. From the point of view of detecting violence in videos, the DenseNet model extracts the

necessary features to perform the detection task more efficiently than other models in terms of numbers of trainable parameters and training time and inference time.

3.4. Multi-Head Self-Attention

The multi-head self-attention [39] is an attention mechanism that links different positions of a single sequence and thus generates a representation of it focusing on the most relevant parts of the sequence. It is based on the attention mechanism that was first introduced in 2014 [37]. Self-attention has had remarkable success being used in natural processing language tasks and text analysis [50,51], to determine which other words are relevant while the current one is being processed.

Multi-head self-attention is a layer that essentially applies multiple self-attention mechanisms in parallel. The procedure is based on projecting the input data applying different linear projections learned from the same data. Then the attention mechanisms are applied to each one of them concatenated.

We use the multi-head self-attention layer in combination with the recurrent convolutional bidirectional layer, to determine which relevant elements are common in both temporal directions, generating a weighted matrix that holds more relevant past and future information simultaneously. The parameters of the multi-head self-attention layer are the next: number of heads $h = 6$, dimension of questions $d\_q = 32$, dimension of values $d\_v = 32$ and dimension of keys $d\_k = 32$. An ablation study is carried out in Section 5 to test the improvements provided by this layer. In the task of detecting violent actions in videos, multi-head self-attention mechanisms establish new relationships between features, determining which of them are the most important in determining whether or not it is a violent action.

3.5. Bidirectional Convolutional LSTM 2D

A bidirectional recurrent cell is a recurrent cell with two states. The two state are, the past state (backward) and the future state (forward). In this way, the output layer to which the bidirectional recurrent layer is connected can obtain information on both states simultaneously. The principle of bidirectional recurrent layers is to divide the neurons of a regular recurrent layer in two directions, positive and negative time directions. This is especially useful in the context of detecting violent actions in videos, as performance improvement can be gained by having the ability to look back.

In a standard recurrent neural network (RNN), temporal features are extracted but spatial ones are lost. To avoid this problem, fully connected layers are replaced with convolutional ones. This is how the ConvLSTM layer can learn the spatio-temporal features of a video, allowing us to take full advantage of the spatio-temporal information that arises from the correlation between convolution and recurrent operations.

BiConvLSTM is an enhancement on ConvLSTM that allows to analyze sequences forward and backward in time simultaneously. A BiConvLSTM layer can access information in both directions of a video's timeline. In this way, a better overall understanding of the video is achieved.

The bidirectional convolutional LSTM 2D module is known in the field of video and image classification to extract spatio-temporal features, being used successfully in other proposals. Zhang et al. [52] used it to recognize gestures in videos and classify them by learning long-term spatio-temporal features. In [53] it is used to classify hyperspectral images, but instead of learning spatial and temporal features, the latter are replaced by spectral features.

3.6. Classifier

The classifier part is made up of four fully connected layers. The number of nodes in each layer, ordered sequentially, is 1024, 128, 16 and 2. Hidden layers use the ReLu activation function. The output of the last layer output is a binary predictor that employs the Sigmoid activation function classifying the input into Violence and Non-Violence categories.

4. Data

In the experiments, the four datasets that appear the most in studies on the detection of violent actions were selected. They are widely accepted and used to compare approaches to detecting violent behaviour. These datasets are as follows:

- Hockey Fights (HFs) [17] a collection of hockey games from the USA's National Hockey League (NHL) that includes fights between players.
- Movies Fights (MFs) [17] a 200-clip collection of scenes from action movies that includes fight and non fight events.
- Violent Flows (VFs) [25] a collection of videos that include violence in crowds. It differs from the previous ones in that it is focused on crowds and not in person-to-person violence but it is interesting to check the versatility of the model.
- Real Life Violence Situations (RLVSs) [54] a collection of 1000 violence and 1000 non-violence videos collected from youtube, violence videos contain many real street fights situations in several environments and conditions. Additionally, non-violence videos are collected from many different human actions like sports, eating, walking, etc.

All four datasets had the same labels, were balanced and were split in a 80–20% ratio for training and testing respectively. Table 1 shows the information of each dataset.

The datasets cover indoor and outdoor scenarios as well as different weather conditions. The Hockey Fights dataset only shows indoor scenarios, specifically an ice hockey arena. The Movies Fights dataset scenes vary between indoor and outdoor scenes, but none of them show adverse weather conditions. The Violent Flows dataset focuses on mass violence that always occurs outdoors. Some scenes contain adverse weather conditions such as rain, fog and snow. The Real Life Violence Situations dataset shows a great variability of indoor and outdoor scenarios ranging from the street to different venues for sporting events, different rooms in a house, stages for music shows, etc. It also shows different adverse weather conditions, although the most frequent is rain.

Table 1. Hockey Fights, Movie Fights, Violent Flows and Real Life Violence Situations dataset features.

Dataset	Number of Clips	Average Frames
Hockey Fights [17]	1000	50
Movies Fights [17]	200	50
Violent Flows [25]	246	100
Real Life Violence Situations [54]	2000	100

5. Experiments

This section summarizes the training methodology and proposes an ablation study to test the importance of the self-attention mechanism. In addition, a cross-dataset experimentation is proposed to evaluate the level of generalization of violent acts. All the experiments are available through Github (https://github.com/FernandoJRS/violence-detection-deeplearning) (accessed on 2 July 2021).

5.1. Training Methodology

For the model, the weights of all neurons were randomly initialized. The pixel values of each frame were normalized to be in the range of 0 to 1. The number of frames of the input video was the average of the frames of all the videos in the dataset. If an input video had more frames than the average, the excess frames were eliminated, if there were fewer frames than the average, the last frame was repeated until the average was reached [55,56]. Frames were resized to $224 \times 224 \times 3$, the standard size for Keras pre-trained models.

A base learning rate of 10^{-4}, a batch size of 12 videos and 100 epochs were selected. Weight decay was initiated to 0.1. Furthermore, the default configuration of the Adam optimizer was used. The Binary Crossentropy function was chosen as the loss function and the Sigmoid function as the activation function for the last layer of the classifier.

To perform the experiments, the CUDA toolbox was used to extract deep features on Nvidia RTX 2070 Super GPU. The operating system was Windows 10 using Intel Core i7. The performances with the datasets were carried out using a random permutation cross-validator method. A five-fold cross validation was chosen.

The experiments were carried out with two kinds of inputs, the first batch with optical flow and a second batch with adjacent frames subtraction, which we called pseudo-optical flow. Both entries implicitly represented the temporal dimension, but in different ways. This was to find out which kind of input obtained the best results for our model.

The pseudo-optical flow was obtained by subtracting two adjacent frames [57]. Given a sequence of frames $(f_0 \dots f_k)$ a matrix subtraction was applied to each pair of adjacent frames, $\forall n < k \in [0, k] : s_n = f_n - f_{n+1}$. With this method, any difference between the pixels of two adjacent frames was represented. Figure 2 shows the transformation of three violent scenes into their respective optical flows and pseudo-optical flows.

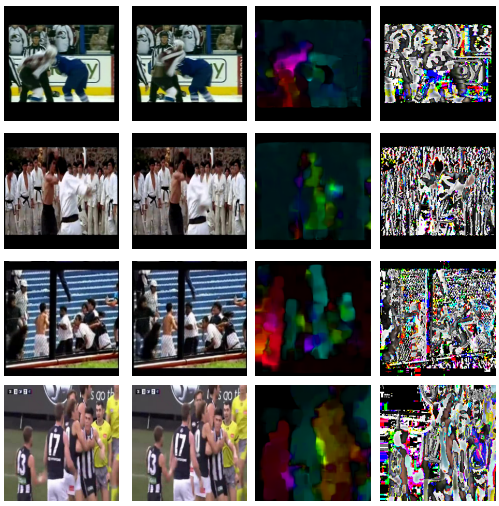

Figure 2. Transformation of sequences of two consecutive frames, of four violent scenes for each dataset (columns 1 and 2), in their dense optical flow (column 3) and adjacent frame subtraction (column 4). The color in (column 3), is for better visualization.

The main difference between both methods is that the pseudo-optical flow also represents those pixels that did not move between two consecutive frames. If the same pixel in both frames did not change in value, by subtracting both frames that pixel turned black, regardless of whether it moved. For the optical flow method, the pixels that turn black are those that have not moved between two of consecutive frames.

5.2. Metrics

To measure the efficiency of our model, the following set of metrics was used:

- Train accuracy: The amount of correct classifications of the model on examples it was constructed on divided by the total amount of classifications.
- Test accuracy: The amount of correct classifications of the model on examples it has not seen divided by the total amount of classifications.
- Test inference time: The average latency time when making predictions atomically on the test dataset.

5.3. Ablation Study

For the ablation study, a double experiment was proposed to test the importance of the self-attention mechanism and to determine how the use of optical flow and pseudo-optical flow affected the results. Avoiding the self-attention mechanism involved the direct connection of the DenseNet to the bidirectional convolutional LSTM.

5.4. Cross-Dataset Experimentation

Cross-dataset experimentation is intended to determine whether a model trained on one dataset can correctly evaluate instances of another dataset. The main objective is to determine if the concept of violence learned by the model is general enough to be able to correctly evaluate other datasets.

Two kinds of cross-dataset setups were tested. In the first one the model was trained with one of the datasets and evaluated with the rest. In the second one the model was trained with a combination of three datasets and evaluated with the remaining one.

6. Results

In this section results obtained from the ablation study, the comparison with state of the art and the cross-dataset experimentation are shown.

6.1. Ablation Study Results

Although a more powerful backbone network was used than in previous work, we considered it interesting to check how the performance improved by changing the network input (optical flow and pseudo-optical flow) and using the attention mechanism. When comparing the two versions of the proposed model (optical flow vs pseudo-optical flow input) with the ones without the self-attention module two main advantages were obtained: better accuracy and shorter inference time. As can be seen in Table 2, both accuracy and inference time were consistently better when the attention module was used. Inference time with attention mechanisms was shorter than without them because bidirectional convolutional recurrent layer operations took longer to apply to featured maps resulting from a convolutional network than to concatenated sequences of attention layers. Luong et al. [58] showed how attention mechanisms could reduce inference time.

It can be seen that when the datasets were composed of clips of 50 frames on average (HF and MF), the differences in accuracy were small, but when the clips were 100 frames on average (VF and RLVS), the accuracy of the model with self-attention outperformed the others by 2 points. Inference time was consistently shorter on each dataset and went from 4% (VF) to 16% (HF) less than without attention.

Finally, comparing the results obtained in each dataset in the pseudo-optical flow without attention and the optical-flow with attention versions, relevant gains were observed in all cases, except in the Movies Fights dataset (very simple). Specifically, the gains were 2 points in HF, 4.4 points in VF and 3.4 points in RLVS.

Table 2. Ablation study of architecture Bi-Dense attention and Bi-Dense without attention.

Dataset	Input	Test Accuracy (with Attention)	Test Accuracy (without Att.)	Test Inference Time (with Attention)	Test Inference Time (without Att.)
HF	Optical Flow	**99.20 ± 0.6%**	99.00 ± 1.0%	**0.1397 ± 0.0024 s**	0.1626 ± 0.0034 s
HF	Pseudo-Optical Flow	97.50 ± 1.0%	97.20 ± 1.0%		
MF	Optical Flow	**100.00 ± 0.0%**	**100.00 ± 0.0%**	**0.1916 ± 0.0093 s**	0.2019 ± 0.0045 s
MF	Pseudo-Optical Flow	**100.00 ± 0.0%**	**100.00 ± 0.0%**		
VF	Optical Flow	**96.90 ± 0.5%**	94.00 ± 1.0%	**0.2991 ± 0.0030 s**	0.3114 ± 0.0073 s
VF	Pseudo-Optical Flow	94.80 ± 0.5%	92.50 ± 0.5%		
RLVS	Optical Flow	**95.60 ± 0.6%**	93.40 ± 1.0%	**0.2767 ± 0.020 s**	0.3019 ± 0.0059 s
RLVS	Pseudo-Optical Flow	94.10 ± 0.8%	92.20 ± 0.8%		

6.2. State of the Art Comparison

After the experiments were carried out, better results were observed for the input of optical flow than with that of pseudo-optical flow. The results of training and testing procedure of one iteration for each dataset and for each type of input to the model are shown in Table 3.

Table 3. Performance comparison for one iteration of our model for Hockey Fights, Movies Fights, Violent Flows and Real Life Violence Situations datasets.

Dataset	Input	Training Accuracy	Training Loss	Test Accuracy Violence	Test Accuracy Non-Violence	Test Accuracy
HF	Optical Flow	100%	1.20×10^{-5}	**99.00%**	**100.00%**	**99.50%**
HF	Pseudo-Optical Flow	99%	1.35×10^{-5}	97.00%	98.00%	97.50%
MF	Optical Flow	100%	1.18×10^{-5}	**100%**	**100%**	**100%**
MF	Pseudo-Optical Flow	100%	1.19×10^{-5}	**100%**	**100%**	**100%**
VF	Optical flow	98%	1.50×10^{-4}	**97.00%**	**96.00%**	**96.50%**
VF	Pseudo-Optical Flow	97%	2.94×10^{-4}	95.00%	94.00%	94.50%
RLVS	Optical Flow	97%	3.10×10^{-4}	**96.00%**	**95.00%**	**95.50%**
RLVS	Pseudo-Optical Flow	95%	7.31×10^{-4}	94.00%	93.00%	93.50%

As can be seen, the optical flow allowed the spatio-temporal dimension of the videos to be highlighted better than the pseudo-optical flow, and favored the training to achieve a greater decrease of the loss function.

When comparing our proposal with the state of the art (Table 4), it is observed that it outperformed previous studies, even those that did not use cross-validation, maintaining a low number of parameters. Best results were obtained for HF and MF where person-to-person violence was present. In particular, the MF dataset was the most homogeneous and the least challenging. The model also worked very well with violence in crowds.

Table 4. State of the art for HF, MF, VF and RLVS datasets. OF stands for optical flow.

Model	HF test Accuracy	MF Test Accuracy	VF Test Accuracy	RLVS Test Accuracy	Train—Test	Validation	Params
VGG13-BiConvLSTM [34]	$96.54 \pm 1.01\%$	**100 ± 0%**	$92.18 \pm 3.29\%$	–	80 – 20%	5 fold cross	–
Spatial Encoder VGG13 [34]	$96.96 \pm 1.08\%$	**100 ± 0%**	$90.63 \pm 2.82\%$	–	80 – 20%	**5 fold cross**	–
FightNet [38]	$97.00 \pm 0\%$	**100 ± 0%**	–	–	80 – 20%	hold-out	–
Three streams + LSTM [31]	$93.90 \pm 0\%$	–	–	–	–	–	–
AlexNet - ConvLSTM [32]	$97.10 \pm 0.55\%$	**100 ± 0%**	$94.57 \pm 2.34\%$	–	80 – 20%	**5 fold cross**	9.6 M
Hough Forest + CNN [22]	$94.60 \pm 0.6\%$	$99.00 \pm 0.5\%$	–	–	80 – 20%	**5 fold cross**	-
FlowGatedNetwork [59]	$48.10 \pm 0\%$	$59.00 \pm 0\%$	$50.00 \pm 0\%$	–	80 – 20%	hold-out	**5.07 K**
Fine-Tuning Mobile-Net [60]	$87.00 \pm 0\%$	$99.50 \pm 0\%$	–	–	80 – 20%	hold-out	–
Xception BiLSTM Attention 10 [36]	$97.50 \pm 0\%$	**100 ± 0%**	–	–	80 – 20%	hold-out	9 M
Xception BiLSTM Attention 5 [36]	$98.00 \pm 0\%$	**100 ± 0%**	–	–	80 – 20%	hold-out	9 M
SELayer-C3D [61]	$99.00 \pm 0\%$	–	–	–	80 – 20%	hold-out	–
Conv2D LSTM [62]	$94.50 \pm 0\%$	–	–	$92.00 \pm 0\%$	80 – 20%	hold-out	–
ViolenceNet Pseudo-OF	$97.50 \pm 1.0\%$	**100 ± 0%**	$94.80 \pm 0.5\%$	$94.10 \pm 0.8\%$	80 – 20%	**5 fold cross**	4.5 M
ViolenceNet OF	**99.20 ± 0.6%**	**100 ± 0%**	**96.90 ± 0.5%**	**95.60 ± 0.6%**	80 – 20%	**5 fold cross**	4.5 M

In the HF dataset, our model generalized very well the amount of movement of the elements given that hockey is a sport where the players are in constant movement and sometimes there is physical contact. The temporal features that our model learned were based on energetic movement when it occurred.

The difficulty of generalizing the concept of violence in the VF dataset was different from that in the HF dataset. The videos from the VF dataset showed violent acts at mass events such as demonstrations, concerts, etc. In massive events, many actions occurred simultaneously. The viewpoints were far from the scene and thus captured many people appearing in low resolution. Many actions in a single video from a far viewpoint made

them seem small and made it harder to distinguish if a touch action was violent or not, even by a person. Furthermore, the context of the mass event included specific situations such as catching a golf ball by a crowd (that could seem the beginning of a fight). It was also difficult to generalize the concept of violence with the RLVS dataset as it is very heterogeneous. Unlike the other three datasets, the RLVS scenes were not topic-specific. The heterogeneity of the dataset was more visible in the non-violence category, where the actions of each scene were very different from each other.

The test accuracy obtained for the different datasets reached the state of the art. Our model was in a good position compared to others. Before our proposal, other models were applied for the HF, MF, VF and RLVS datasets.

For the MF dataset there were several models that reached a test accuracy of $100 \pm 0\%$ which made them unbeatable.

For the VF dataset our proposal outperformed the closest one, [32], by more than 2 points.

The RLVS dataset was only tested with a model prior to ours, [54]. This model had a test accuracy for the RLVS dataset of 92.00% using hold-out validation. Our model achieved a value of 95.60%, again improving on the state of the art.

The test accuracy values were slightly higher with the optical flow input than with the pseudo-optical flow input. This occurred with all datasets except the MF dataset, which had the same test accuracy value for both.

Even comparing our model with those who used a hold-out validation methodology, it can be seen that our method improved the state of the art.

Another remarkable advantage of our proposal was the number of trainable parameters used, less than for the rest of the models for which these data were available. This is due to the Dense architecture and its feature map concatenation method. The only model with a number of trainable parameters lower than the proposed one was $FlowGatedNetwork - 3DCNN - Flow - RGB$ [59], however, this model was not relevant because its test accuracy did not exceed 60% for any of the datasets.

6.3. Cross-Dataset Experimentation Results

After performing the cross-dataset experiments, two logical facts were observed: on the one hand, there was a slight correlation between more heterogeneous datasets and a better generalization of the concept of violent actions. On the other hand, the experiments performed by training the model with unions of different datasets showed a better generalization of the concept of violence than in those experiments in which the model was trained with a single dataset.

It can also be observed that for pairs of datasets very different in the type of context of violence, the results did not show effective generalization. An example of this was the pair MF and VF, where the contexts were clearly different (MF violence between two people or a very small group and VF focuses on mass violence), obtaining a very low accuracy (52.32%) in the MF -> VF direction and somewhat higher in the opposite direction (60.02%), given that the length and variability of VF was higher.

The best result obtained during cross-dataset experimentation was the one where the model was trained with the combination of the HF, RLVS and VF datasets, achieving a test accuracy of 81.51%, but very low compared to cross-validation using the same dataset (100%). Finally, experiments that included the RLVS dataset for model training obtained better results than experiments in which the model was not trained with it. The RLVS dataset was the largest and most heterogeneous of the four.

The results in testing for five iterations for each type of input are shown in Table 5.

Table 5. Cross-dataset experiment results.

Dataset Training	Dataset Testing	Test Accuracy Optical Flow	Test Accuracy Pseudo-Optical Flow
HF	MF	65.18 ± 0.34	64.86 ± 0.41
HF	VF	62.56 ± 0.33	61.22 ± 0.22
HF	RLVS	58.22 ± 0.24	57.36 ± 0.22
MF	HF	54.92 ± 0.33	53.50 ± 0.12
MF	VF	52.32 ± 0.34	51.77 ± 0.30
MF	RLVS	56.72 ± 0.19	55.80 ± 0.20
VF	HF	65.16 ± 0.59	64.76 ± 0.49
VF	MF	60.02 ± 0.24	59.48 ± 0.16
VF	RLVS	58.76 ± 0.49	58.32 ± 0.27
RLVS	HF	69.24 ± 0.27	68.86 ± 0.14
RLVS	MF	75.82 ± 0.17	74.64 ± 0.22
RVLS	VF	67.84 ± 0.32	66.68 ± 0.22
HF + MF + VF	RLVS	**70.08 ± 0.19**	69.84 ± 0.14
HF + MF + RLVS	VF	**76.00 ± 0.20**	75.68 ± 0.14
HF + RLVS + VF	MF	**81.51 ± 0.09**	80.49 ± 0.05
RLVS + MF + VF	HF	**79.87 ± 0.33**	78.63 ± 0.01

6.4. Detection Process in CCTV

The ViolenceNet model allowed classifying videos between the categories of violence and non-violence. In a CCTV system, our model worked with short video fragments of equal length. Each of these fragments was preprocessed, applying the dense optical flow algorithm that generated the input to the ViolenceNet model that was in charge of classifying these fragments into violence or non-violence, as shown in Figure 3.

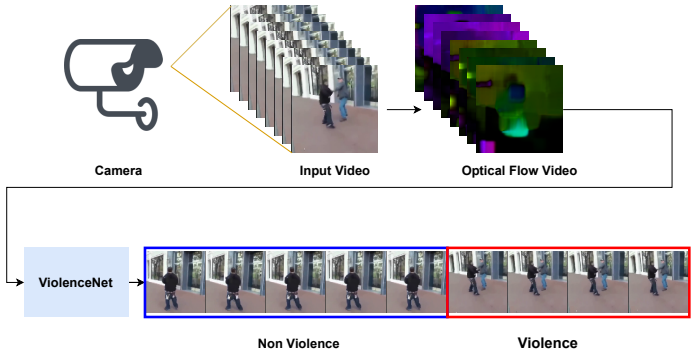

Figure 3. CCTV scheme for detection of violence scenes. In a CCTV system, our model receives the optical flow of the fragments provided by a camera and classifies each of them. The red box represents violent segments and the blue box represents non-violent segments.

7. Discussion

The developed model for the detection of violent actions has been more efficient in improving the state of the art than in generalizing the concept of violent actions. This is demonstrated in the results presented in the previous section. However, some results obtained in cross-dataset experiments show that it is possible to improve the generalization of violence if large and heterogeneous datasets with a large number of instances that contemplate different contexts and scenarios are used. Datasets with few instances like MF or HF with a single type of context are not effective in generalizing the concept of violence regardless of whether the input is optical flow or pseudo-optical flow.

8. Conclusions and Future Work

We have proposed ViolenceNet, a space-time encoder architecture for the detection of violent actions that improves the state of the art. While several studies have used

RCNN and bi-directional RCNN for video problems, our main contributions have been an architecture that combines a modified DenseNet with a multi-head self-attention module and a 2D LSTM bi-directional convolutional module; a comparative analysis of two types of input (optical flow and pseudo-optical flow) including an ablation study of the self-attention mechanism; an experimentation with four datasets that are benchmarks in the field of violence detection that show that our proposal overpass the state of the art; and finally a cross-dataset experimentation to carry out an analysis of the generalization of violent actions.

This last analysis on short video datasets shows how the accuracy drops from accuracy values between 95% and 100% using the same dataset cross-validation to values between 70.08% and 81.51% in the cross-dataset experiments, which leads us to think that future work should focus on anomaly detection on long video datasets. Among other datasets, UCF-Crime [63], XD-Violence [64], UBI-Fights [65] or CCTV-Fights [66] are worth mentioning. In these cases, it would be necessary to analyze chunks of the video to obtain not only whether there is violence or not, but at what moment it occurs. For this reason, an architecture like the one presented, capable of capturing temporal features in both directions, is an efficient way of dealing with more heterogeneous datasets. Another interesting line of research to follow is the use of new deep learning techniques based on transformers [67].

Finally, our model does not include any human features, and it works correctly given the input dataset, but to achieve a generalization of violence involving people it would be necessary to include pose estimation or at least face detection in our future work. We believe that further research on this can lead to fruitful results.

Author Contributions: Conceptualization, J.A.Á.-G.; Funding acquisition, J.A.A.-G., O.D.; Investigation, F.J.R-S.; Methodology, F.J.R-S., F.E.; Software, F.J.R-S.; Writing—original draft, F.J.R-S. All authors have read and agreed to the published version of the manuscript.

Funding: This research is partially supported by The Spanish Ministry of Economy and Competitiveness MINECO/FEDER R&D, UE through the project VICTORY (Grant No.: TIN2017-82113-C2-1-R and TIN2017-82113-C2-2-R).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data available in a publicly accessible repository. The data presented in this study is openly available in https://github.com/FernandoJRS/violence-detection-deeplearning

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chaquet, J.M.; Carmona, E.J.; Fernández-Caballero, A. A survey of video datasets for human action and activity recognition. *Comput. Vis. Image Underst.* **2013**, *117*, 633–659. [CrossRef]
2. Guo, G.; Lai, A. A survey on still image based human action recognition. *Pattern Recognit.* **2014**, *47*, 3343–3361. [CrossRef]
3. Carranza-García, M.; Torres-Mateo, J.; Lara-Benítez, P.; García-Gutiérrez, J. On the Performance of One-Stage and Two-Stage Object Detectors in Autonomous Vehicles Using Camera Data. *Remote Sens.* **2021**,*13*, 89. [CrossRef]
4. Velastin, S.A.; Boghossian, B.A.; Vicencio-Silva, M.A. A motion-based image processing system for detecting potentially dangerous situations in underground railway stations. *Transp. Res. Part Emerg. Technol.* **2006**, *14*, 96–113. [CrossRef]
5. Ainsworth, T. Buyer beware. *Secur. Oz* **2002**, *19*, 18–26.
6. Cheng, G.; Wan, Y.; Saudagar, A.N.; Namuduri, K.; Buckles, B.P. Advances in human action recognition: A survey. *arXiv* **2015**, arXiv:1501.05964.
7. Kooij, J.; Liem, M.; Krijnders, J.; Andringa, T.; Gavrila, D. Multi-modal human aggression detection. *Comput. Vis. Image Underst.* **2016**, *144*, 106–120. [CrossRef]
8. Nazare, A.C., Jr.; Schwartz, W.R. A scalable and flexible framework for smart video surveillance. *Comput. Vis. Image Underst.* **2016**, *144*, 258–275. [CrossRef]
9. Salazar-González, J.L.; Zaccaro, C.; Álvarez-García, J.A.; Soria-Morillo, L.M.; Caparrini, F.S. Real-time gun detection in CCTV: An open problem. *Neural Netw.* **2020**, *132*, 297–308. [CrossRef]
10. Vallez, N.; Velasco-Mata, A.; Deniz, O. Deep autoencoder for false positive reduction in handgun detection. *Neural Comput. Appl.* **2020**, 1–11. [CrossRef]

11. Ruiz-Santaquiteria, J.; Velasco-Mata, A.; Vallez, N.; Bueno, G.; Álvarez García, J.A.; Deniz, O. Handgun detection using combined human pose and weapon appearance. *arXiv* **2021**, arXiv:2010.13753.
12. United Nations Office on Drugs and Crime (UNODC) Global Study on Homicide 2019. Available online: https://www.unodc.org/documents/data-and-analysis/gsh/Booklet1.pdf (accesed on 2 July 2021)
13. Clarin, C.; Dionisio, J.; Echavez, M.; Naval, P. DOVE: Detection of movie violence using motion intensity analysis on skin and blood. *PCSC* **2005**, *6*, 150–156.
14. Chen, D.; Wactlar, H.; Chen, M.y.; Gao, C.; Bharucha, A.; Hauptmann, A. Recognition of aggressive human behavior using binary local motion descriptors. In Proceedings of the 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Vancouver, BC, Canada, 20–25 August 2008; pp. 5238–5241.
15. Xu, L.; Gong, C.; Yang, J.; Wu, Q.; Yao, L. Violent video detection based on MoSIFT feature and sparse coding. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 3538–3542.
16. Ribeiro, P.C.; Audigier, R.; Pham, Q.C. RIMOC, a feature to discriminate unstructured motions: Application to violence detection for video-surveillance. *Comput. Vis. Image Underst.* **2016**, *144*, 121–143. [CrossRef]
17. Bermejo, E.; Deniz, O.; Bueno, G.; Sukthankar, R. Violence detection in video using computer vision techniques. In *International Conference on Computer Analysis of Images and Patterns*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 332–339.
18. Bilinski, P.; Bremond, F. Human violence recognition and detection in surveillance videos. In Proceedings of the 2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Colorado Springs, CO, USA, 23–26 August 2016; pp. 30–36.
19. Cai, H.; Jiang, H.; Huang, X.; Yang, J.; He, X. Violence detection based on spatio-temporal feature and fisher vector. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 180–190.
20. Senst, T.; Eiselein, V.; Kuhn, A.; Sikora, T. Crowd violence detection using global motion-compensated Lagrangian features and scale-sensitive video-level representation. *IEEE Trans. Inf. Forensics Secur.* **2017**, *12*, 2945–2956. [CrossRef]
21. Zhang, T.; Jia, W.; Yang, B.; Yang, J.; He, X.; Zheng, Z. MoWLD: A robust motion image descriptor for violence detection. *Multimed. Tools Appl.* **2017**, *76*, 1419–1438. [CrossRef]
22. Serrano, I.; Deniz, O.; Espinosa-Aranda, J.L.; Bueno, G. Fight recognition in video using Hough Forests and 2D convolutional neural network. *IEEE Trans. Image Process.* **2018**, *27*, 4787–4797. [CrossRef] [PubMed]
23. Das, S.; Sarker, A.; Mahmud, T. Violence Detection from Videos using HOG Features. In Proceedings of the 2019 4th International Conference on Electrical Information and Communication Technology (EICT), Khulna, Bangladesh, 20–22 December 2019; pp. 1–5. [CrossRef]
24. Zhou, P.; Ding, Q.; Luo, H.; Hou, X. Violence detection in surveillance video using low-level features. *PLoS ONE* **2018**, *13*, e0203668. [CrossRef]
25. Hassner, T.; Itcher, Y.; Kliper-Gross, O. Violent Flows: Real-time detection of violent crowd behavior. In Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, 16–21 June 2012; pp. 1–6.
26. Zhang, T.; Yang, Z.; Jia, W.; Yang, B.; Yang, J.; He, X. A new method for violence detection in surveillance scenes. *Multimed. Tools Appl.* **2016**, *75*, 7327–7349. [CrossRef]
27. Mahmoodi, J.; Salajeghe, A. A classification method based on optical flow for violence detection. *Expert Syst. Appl.* **2019**, *127*, 121–127. [CrossRef]
28. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. In Proceedings of the Advances in Neural Information Processing Systems 27, Montréal, QC, Canada, 8–13 December 2014; pp. 568–576.
29. Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Fei-Fei, L. Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1725–1732.
30. Meng, Z.; Yuan, J.; Li, Z. Trajectory-pooled deep convolutional networks for violence detection in videos. In *International Conference on Computer Vision Systems*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 437–447.
31. Dong, Z.; Qin, J.; Wang, Y. Multi-stream deep networks for person to person violence detection in videos. In *Chinese Conference on Pattern Recognition*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 517–531.
32. Sudhakaran, S.; Lanz, O. Learning to detect violent videos using Convolutional long short-term memory. In Proceedings of the 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Lecce, Italy, 29 August–1 September 2017; pp. 1–6.
33. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*; The MIT Press: Cambridge, MA, USA, 2012; pp. 1097–1105.
34. Hanson, A.; PNVR, K.; Krishnagopal, S.; Davis, L. Bidirectional Convolutional LSTM for the Detection of Violence in Videos. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, September 2018; pp. 280–295.
35. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.

36. Aktı, Ş.; Tataroğlu, G.A.; Ekenel, H.K. Vision-based Fight Detection from Surveillance Cameras. In Proceedings of the 2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA), Istanbul, Turkey, 6–9 November 2019; pp. 1–6.
37. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
38. Zhou, P.; Ding, Q.; Luo, H.; Hou, X. Violent interaction detection in video based on deep learning. *J. Phys. Conf. Ser. IOP Publ.* **2017**, *844*, 012044. [CrossRef]
39. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.
40. Diba, A.; Fayyaz, M.; Sharma, V.; Karami, A.H.; Arzani, M.M.; Yousefzadeh, R.; Van Gool, L. Temporal 3d convnets: New architecture and transfer learning for video classification. *arXiv* **2017**, arXiv:1711.08200.
41. Baradel, F.; Wolf, C.; Mille, J. Pose-conditioned spatio-temporal attention for human action recognition. *arXiv* **2017**, arXiv:1703.10106.
42. Cho, S.; Maqbool, M.; Liu, F.; Foroosh, H. Self-attention network for skeleton-based human action recognition. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass, CO, USA, 1–5 March 2020; pp. 635–644.
43. Courtney, L.; Sreenivas, R. Using Deep Convolutional LSTM Networks for Learning Spatiotemporal Features. In *Asian Conference on Pattern Recognition*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 307–320.
44. Farnebäck, G. Two-frame motion estimation based on polynomial expansion. In *Scandinavian Conference on Image Analysis*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 363–370.
45. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
46. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
47. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
48. Liang, S.; Zhang, R.; Liang, D.; Song, T.; Ai, T.; Xia, C.; Xia, L.; Wang, Y. Multimodal 3D DenseNet for IDH genotype prediction in gliomas. *Genes* **2018**, *9*, 382. [CrossRef]
49. Wang, H.; Shen, Y.; Wang, S.; Xiao, T.; Deng, L.; Wang, X.; Zhao, X. Ensemble of 3D densely connected convolutional network for diagnosis of mild cognitive impairment and Alzheimer's disease. *Neurocomputing* **2019**, *333*, 145–156. [CrossRef]
50. Lin, Z.; Feng, M.; Santos, C.N.d.; Yu, M.; Xiang, B.; Zhou, B.; Bengio, Y. A structured self-attentive sentence embedding. *arXiv* **2017**, arXiv:1703.03130.
51. Paulus, R.; Xiong, C.; Socher, R. A deep reinforced model for abstractive summarization. *arXiv* **2017**, arXiv:1705.04304.
52. Zhang, L.; Zhu, G.; Shen, P.; Song, J.; Afaq Shah, S.; Bennamoun, M. Learning spatiotemporal features using 3dcnn and convolutional lstm for gesture recognition. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 3120–3128.
53. Liu, Q.; Zhou, F.; Hang, R.; Yuan, X. Bidirectional-convolutional LSTM based spectral-spatial feature learning for hyperspectral image classification. *Remote Sens.* **2017**, *9*, 1330.
54. Soliman, M.M.; Kamal, M.H.; Nashed, M.A.E.M.; Mostafa, Y.M.; Chawky, B.S.; Khattab, D. Violence Recognition from Videos using Deep Learning Techniques. In Proceedings of the 2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS), Cairo, Egypt, 8–10 December 2019; pp. 80–85.
55. Varol, G.; Laptev, I.; Schmid, C. Long-term temporal convolutions for action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 1510–1517. [CrossRef] [PubMed]
56. Sanchez-Caballero, A.; de López-Diz, S.; Fuentes-Jimenez, D.; Losada-Gutiérrez, C.; Marrón-Romera, M.; Casillas-Perez, D.; Sarker, M.I. 3DFCNN: Real-Time Action Recognition using 3D Deep Neural Networks with Raw Depth Information. *arXiv* **2020**, arXiv:2006.07743.
57. Sharma, M.; Baghel, R. Video Surveillance for Violence Detection Using Deep Learning. In *Advances in Data Science and Management*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 411–420.
58. Luong, M.T.; Pham, H.; Manning, C.D. Effective approaches to attention-based neural machine translation. *arXiv* **2015**, arXiv:1508.04025.
59. Cheng, M.; Cai, K.; Li, M. RWF-2000: An Open Large Scale Video Database for Violence Detection. *arXiv* **2019**, arXiv:1911.05913.
60. Khan, S.U.; Haq, I.U.; Rho, S.; Baik, S.W.; Lee, M.Y. Cover the violence: A novel Deep-Learning-Based approach towards violence-detection in movies. *Appl. Sci.* **2019**, *9*, 4963. [CrossRef]
61. Jiang, B.; Xu, F.; Tu, W.; Yang, C. Channel-wise attention in 3d convolutional networks for violence detection. In Proceedings of the 2019 International Conference on Intelligent Computing and its Emerging Applications (ICEA), Tainan, Taiwan, 30 August–1 September 2019; pp. 59–64.
62. Moaaz, M.M.; Mohamed, E.H. Violence Detection In Surveillance Videos Using Deep Learning. *Inform. Bull. Helwan Univ.* **2020**, *2*, 1–6. [CrossRef]

63. Sultani, W.; Chen, C.; Shah, M. Real-world anomaly detection in surveillance videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6479–6488.
64. Wu, P.; Liu, J.; Shi, Y.; Sun, Y.; Shao, F.; Wu, Z.; Yang, Z. Not only Look, but also Listen: Learning Multimodal Violence Detection under Weak Supervision. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 322–339.
65. Degardin, B.; Proença, H. Iterative weak/self-supervised classification framework for abnormal events detection. *Pattern Recognit. Lett.* **2021**, *145*, 50–57. [CrossRef]
66. Perez, M.; Kot, A.C.; Rocha, A. Detection of real-world fights in surveillance videos. In Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 2662–2666.
67. Girdhar, R.; Carreira, J.; Doersch, C.; Zisserman, A. Video action transformer network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 244–253.