Attention Inspiring Receptive-Fields Multi-Task Network via Self- supervised Learning for Violence Recognition

Suyuan Li
Northeastern University
Xin Song (sxin78916@neuq.edu.cn)
Northeastern University

Research Article

Keywords: Violence recognition, attention module, self-supervised learning CNN

Posted Date: April 6th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2778719/v1

Additional Declarations: No competing interests reported.

Attention Inspiring Receptive-Fields Multi-Task Network via Self-supervised Learning for Violence Recognition

Suyuan Li[a], Xin Song[a,b,*]

[a] School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China

[b] School of Computer and Communication Engineering, Northeastern University at Qinhuangdao, Qinhuangdao 066004, China

E-mail: 2010649@stu.neu.edu.cn (S. Li), sxin78916@neuq.edu.cn (X. Song).

Abstract: Generally, a large amount of training data is essential to train deep learning model for obtaining more accurate detection performance in computer vision domain. However, to collect and annotate datasets will lead to extensive cost. In this letter, we propose a self-supervised auxiliary task to learn general videos features without adding any human-annotated labels, aiming at improving the performance of violence recognition. Firstly, we propose a violence recognition method based on convolutional neural network with self-supervised auxiliary task, which can learn visual feature for improving down-stream task (recognizing violence). Secondly, we establish a balance-weighting scheme to solve the crucial problem of balancing the self-supervised auxiliary task and violence recognition task. Thirdly, we develop an attention receptive-field module, indicating that the proper use of the spatial attention mechanism can effectively expand the receptive fields of the module, further improving semantically meaningful representation of the network. To evaluate the proposed method, two benchmark datasets have been used, and better performance is shown by the experimental results comparing with other state-of-the-art methods.

Key words: Violence recognition, attention module, self-supervised learning CNN.

1. Introduction

Usually, violent behaviors pose a major threat to social and personal safety. In recent years, violence recognition becomes a fascinating and challenging problem in surveillance video that has attracted the interest of a wide number of researchers. Due to the rapid popularity of the internet, video surveillance technology gradually becomes an essential security monitoring approach. Up to now, some developmental systems for detecting violence based on audio and vision information have achieved great progress in practice. With the development of neural network, the achievement of deep learning has ushered in a new phase of artificial intelligence, bringing with creative solutions and technological advancements in intelligent video analysis. Generally, better detection performance requires a large amount of video data for training the deep learning model. To avoid collecting and annotating datasets, self-supervised learning can usually be considered as effective solution. Therefore, we design self-supervised auxiliary task to learn general videos features without adding any human-annotated labels, improving the performance of violence recognition.

Commonly the existing methods of violence recognition can be primarily divided into two categories: (1) Methods based on hand-crafted features. These methods mainly concentrate on the extraction of hand-crafted features used for modeling a video event (e.g., Invariant Feature Transform (MoSIFT) [2], Histogram of Oriented Gradients [3], Space Time Interest Points (STIP) [4, 5] and Histogram of Oriented Flows (HOF) [6]). (2) Methods based on deep learning features. Convolutional Neural Networks (CNNs) have recently gained from the implementation in a wide range of image

processing tasks, including image classification [7-10], object recognition [11, 12], and many other anomaly detection tasks [13, 14]. As a result, a wide variety of deep learning-based methods for recognizing violence have been proposed [15-18]. For instant, in [15], combination of MCNN, key-point and LSTM are proposed to recognize violent behaviors. However, the majority of currently used CNN-based violence recognition systems are single-task, deep learning models only accomplish one goal. In contrast, multi-task deep learning models [19, 20] allow for the simultaneous learning of numerous tasks, and the extracted representations are somewhat shared throughout these tasks. While the model is trained using these single-task deep learning methods, the underlying relevance between several tasks will be neglected. Despite the superiority of MTL in a variety of visual tasks, the core issue is the design of a multi-task network that is the most suitable for video violence recognition.

As discussed above, we therefore propose to perform a self-supervised auxiliary task by reconstructing frames for improving the violence recognition in this letter. Specifically, we carefully investigate the multi-task issue of video violence recognition, where the main task is to recognize violent behaviors and the auxiliary task based on self-supervised learning is to reconstruct frames. The auxiliary task can reconstruct the original frame without adding any human-annotated labels, which is helpful to learn more abundant visual feathers. For that, the performance of down-stream main task (recognizing violence) will be improved. In addition, a modular encoder-decoder-based convolutional neural network architectural unit is built to motivate multi-level receptive fields. The key contributions of this paper can be summarized as follows:

i) We proposed a multi-task AMCNN network to recognize violent behaviors via self-supervised learning, where violence recognition and frame reconstruction are respectively considered as the discrimination task and the generation task, and so that more discriminative and informative visual features can be obtained. ii) We design an attention module to enlarge the receptive fields for enhancing more semantic representation of the network. To our knowledge, we are the first to propose a multi-task learning approach that integrates novel self-supervised task for violence recognition in video.

2. Proposed method

This section will introduce the details of AMCNN for video violence recognition. The proposed algorithm can mainly be divided into the following blocks: overall architecture, attention module, multi-task module and loss function.

2.1 Overall Architecture

In this letter, a novel method for detecting video violent behaviors is proposed, which combines a convolutional neural network with an attention receptive-field module and a multi-task mechanism. **Fig. 1** shows the framework of AMCNN. Firstly, taking account for the motion of consecutive video frames, optical flow images are utilized as input to the network. Secondly, the CNN that mainly consists of one attention module and four convolutional blocks is adopted to extract more semantic latent visual features on optical flows. In order to categorize latent visual features, frame reconstruction is carried out as a self-supervised auxiliary task to reconstruct the original frames, improving the representativeness of the latent visual features, correspondingly, recognizing violent behaviors is regarded as the main task Accordingly, after optimizing a balance-weighting multi-task loss function, the best AMCNN model can be obtained, further recognizing violent behaviors in a testing procedure.

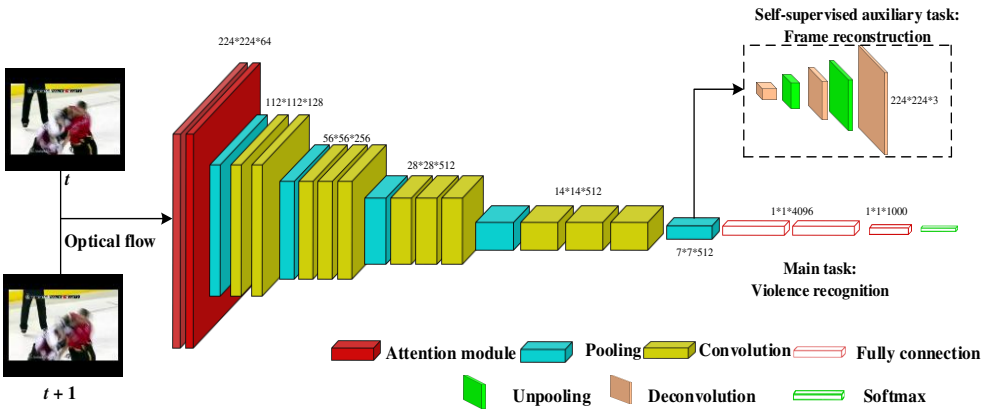

Fig. 1. Overall structure of the proposed AMCNN method.

2.2 Attention Module

To enlarge the receptive fields for enhancing the semantic representation of latent features, the spatial attention mechanism is utilized as the mask branch, adding it to the master branch of CNN. As depicted as in **Fig. 2**, the attention module mainly includes two branches: the master branch of CNN and the attention mask branch. The master branch is two conventional convolutional layers in the VGG16, which can be represented as:

$$F_{1,2}\left(\boldsymbol{x},\{W_1,W_2\}\right) \tag{1}$$

in which $F_{1,2}$ is the master branch, $\boldsymbol{x}$ is the input vector and $\{W_1,W_2\}$ represent the weights of the master branch. The attention mask branch is a micro encoder-decoder structure, in which the encoder adopts convolutional layers and a sampling operation, while the decoder adopts an upsampling operation. Accordingly, the output $\boldsymbol{y}$ of attention module can be represented as:

$$\boldsymbol{y} = F_{1,2}\left(\boldsymbol{x},\{W_1,W_2\}\right) * M\left(\boldsymbol{x},\{W_m\}\right) \tag{2}$$

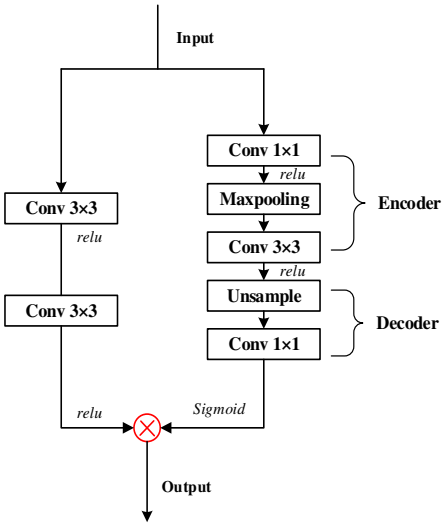

Fig. 2. Attention receptive-field module.

in which M represents the multiple learnable convolutional weights of the attention mask branch and $\{W_m\}$ represents mask branch with a micro encoder-decoder structure. The component of encoder in the attention module expands the receptive fields while swiftly gathering context data from the input. The later decoder recovers the dimension of the input feature by combining the global data with the original feature maps. To the specific, the primary semantic information of the entire attention mask branch is represented by the output of the 3×3 convolutional layer, and two 1×1 convolutional layers are utilized to decrease or increase dimensions. Additionally, batch normalization [21] and ReLU [22] activation function are adopted after every convolutional layer in the attention module. Specially, the sigmoid operation is used in the mask attention branch to normalize the output into the range between 0 and 1. And the location of violent behaviors will be more likely to have a higher output value, which can enhance the semantic representation of the master branch.

2.3 Multi-task Module and Loss Function

In the multi-task framework, the VGG16 with the attention module is considered as shared task layer, recognizing violent behaviors and reconstructing frames are considered as task-specific layers. Two tasks are closely related in shared layer. Self-supervised auxiliary task can learn kernels to capture both low-level and high-level characteristics that are useful for recognizing violent behaviors. Consequently, the completion of the self-supervised auxiliary task will be beneficial for making the latent visual features better represent the original inputs, which can further improve the accuracy of violence recognition. For the main task and self-supervised side task, our network can be trained to minimize the loss as follows:

$$\begin{aligned} &\min \sum_{i=1}^{N} Loss = \alpha \cdot L_1 + \beta \cdot L_2 \\ &s.t. \quad \alpha + \beta = 1, \alpha > \beta \end{aligned} \tag{3}$$

where α is the weight of the main task and β is the weight of self-supervised auxiliary task, limiting the impact of different tasks in the multi-task learning loss function. Finally, the optimal values of α and β are respectively set as 0.7 and 0.3, which can be proved in the next experiments. Specifically, for recognizing violent behaviors, the cross-entropy loss function for main task can be adopted:

$$L_1 = \frac{1}{N}\sum_{i=1}^{N}\left[t_i \cdot \log(p_i) + (1-t_i) \cdot \log(1-p_i)\right] \tag{4}$$

where t_i is the ground label of the i_{th} sample, and p_i is the predicted result of recognizing violence, and N denotes the total number of samples. Specifically, for reconstructing frames, the mean square error loss function of the side task can be adopted:

$$L_2 = \frac{1}{N}\sum_{1}^{N}\|\boldsymbol{x}_i - \boldsymbol{y}_i\|^2 \tag{5}$$

in which $\boldsymbol{x}_i$ is the i_{th} original frame, $\boldsymbol{y}_i$ is the i_{th} reconstructed frame and N is the total number of samples.

3. Experiments

In this section, the description of two benchmark datasets and details of implementation are firstly provided, and then, the influence of attention module and multi-task loss can be verified. Finally, we evaluate and compare the effectiveness of proposed method with other methods.

3.1 Datasets and Experiment Setting

Datasets: The Hockey Fight [23] and Movies Dataset [24] are adopted for training and evaluating the proposed AMCNN. The Hockey Fight includes 1000 video clips, which consist of an equal number of violent and non-violent behaviors. Some sample frames the Hockey Fight dataset is shown in **Fig. 3**.

Furthermore, Movies Dataset contains 100 violent clips and 100 non-violent clips. Some sample frames the Movies Dataset is shown in **Fig. 4**. Specifically, the training set, validation set, and testing set will be divided in 4:1:1.

Fig. 3. Sample frames from the Hockey Fight dataset

Fig. 4. Sample frames from the Movies Dataset

Details for Implementation: we implement the AMCNN using PyTorch on a server with two RTX 2080Ti GPUs. In the training processing, the initial learning rate is set to 0.001, an ADAM optimizer with a mini-batch size of 128 is adopted. Moreover, the revolution of the input frames is resized to 224×224, further applying to the network.

3.2 Ablation Experiments

This section conducts an ablation analysis using the Hockey Fight dataset. The attention module, various downsampling operation types, and various multi-task weights can all be used to assess performance. After that, by comparing different metrics, the best AMCNN parameters can be obtained.

Attention module: To demonstrate and analyze the effect of the attention module used in the AMCNN, we compare different number of the embedded attention modules. Generally, the outputs of these convolutional blocks in VGG16 are respectively denoted as {C1, C2, C3, C4, C5} for conv1, conv2, conv3, conv4, and conv5. Therefore, the outputs of these attention modules are correspondingly denoted

as {A1, A2, A3, A4, A5}. **Table 1** shows metrics of different number of embedded attention modules. Apparently, as shown in the Table 1, the model size of AMCNN grows as the number of attention modules increases. Therefore, we must control the number of embedded attention modules to reduce computation and parameters. Moreover, the accuracy of only embedded one attention module in the first convolutional block is obviously higher, which can fully prove the effectiveness, and it is worth noting that the accuracy decreases as the number of attention modules increases. It's because the size of feature maps decreases with the increase of the network, which will cause the mask to be too rough, further influencing performance of the network. Therefore, we only utilize one attention module in our proposed method.

Table 1. Metrics of different number of attention modules

Stage	Model Size/MB	Accuracy
No Attention	545.7	94.77%
A_1	**546.1**	**95.71%**
A_1, A_2,	547.6	93.90%
A_1, A_2, A_3,	552.6	93.75%
A_1, A_2, A_3, A_4,	573.7	92.93%
A_1, A_2, A_3, A_4, A_5	595.7	92.29%

Downsampling Operation Type: To verify the performance of different downsampling operation types in the attention module, average pooling, max pooling, or convolution with stride=2 and kernel=3 are performed. As shown in **Table 2**, the accuracy of AMCNN will be impacted by these several operations. The attention module with the convolution consumes more parameters and achieves lower accuracy, and the attention module with max pooling can achieve the best accuracy than others. Consequently, in the attention module, we employ max pooling as a downsampling method.

Table 2. Metrics of different downsampling operation

Type	Parameters	Time/Min	Accuracy
Maxpool	**45,544,645**	**6.26**	**95.71%**
Avgpool	45,544,645	6.23	94.97%
Conv	45,546,181	5.41	92.36%

Multi-task loss: To validate the benefit of self-supervised auxiliary task for violence recognition, we analyze the accuracy of various weights in the loss function. In our method, $\alpha > \beta$ denotes that violence recognition is considered as the main task, and conversely, $\alpha < \beta$ denotes that self-supervised auxiliary task is considered as the main task. **Fig. 5** shows the accuracy of different weights in the loss function. As we can see that the accuracy on $\alpha > \beta$ is much greater than that on $\alpha < \beta$. This supports our claim that the completion of the auxiliary task can make the latent visual features better represent the original inputs, and further improving the accuracy of violence recognition in the main task. It is worth noting that $\alpha = 1$ denotes single-task violence recognition task, the accuracy on $\alpha = 1$ is obviously lower than those on $0.5 \le \alpha < 1$. This also can verify that the ability to recognize violent behaviors can be boosted by the multi-task mechanism with appropriative self-supervised auxiliary task. Additionally, the best accuracy is based on $\alpha = 0.7$ and $\beta = 0.3$, which means that self-supervised auxiliary task is

most helpful to recognize violent behaviors at this time. Therefore, the optimal values of α and β are respectively set as 0.7 and 0.3 in our method.

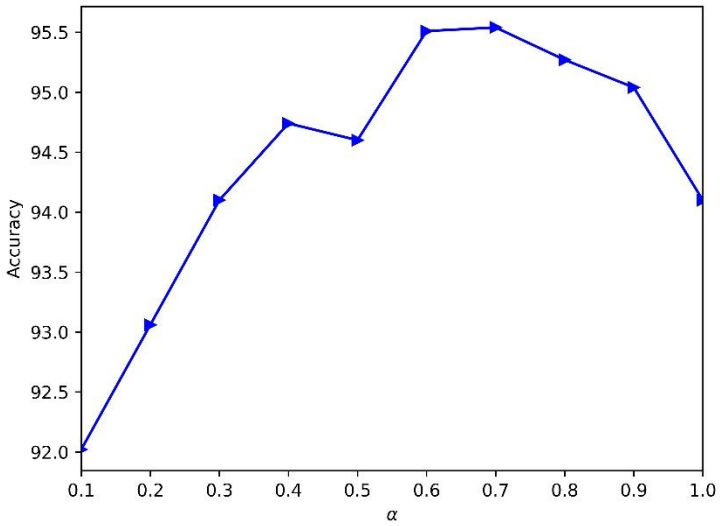

Fig. 5. The accuracy of different weights in the loss function.

3.3 Comparison Results with Other Models

In this section, we compare our method against various recent existing methods on the Hockey Fights dataset and Movies dataset in **Table 3** and **Table 4** to illustrate the effectiveness of AMCNN. **Table 3** illustrates the accuracy of conventional methods, which mainly rely on different kinds of hand-crafted features. For instance, Lagrangian direction fields based on a spatio-temporal model and appearance are combined to recognize violence [25]. According to **Table 3**, the proposed AMCNN achieves the best accuracy than other standard traditional methods on two benchmark datasets. This is because these hand-crafted features of traditional methods will reduce the effect in real complex scenes.

Table 3. The accuracy of conventional methods

Methods	Hockey	Movies
HOG3D [24]	95.09%	99.90%
BoW [25]	94.42%	94.95%
MoSIFT [26]	94.30%	89.50%
STIFV [27]	93.70%	99.50%
VIF [28]	82.60%	91.30%
Proposed method	**95.71%**	**100%**

Table 4. The accuracy of deep-learning-based methods

Methods	Hockey	Movies
CNN [29]	94.60%	99%
P3D + LSTM [30]	94.40%	97.97%
3D CNN [31]	91.00%	-
Multi-CNN [32]	89.10%	100%
ResNet50 [33]	83.19%	88.74%
Proposed method	**95.71%**	**100%**

Table 4 illustrates accuracy of these deep-learning-based methods, which mainly depend on automatically extracted deep features. For instance, the binary robust invariant scalable key points (BRISK) and Hough forests are combined to obtain the representative image, and 2D CNN is utilized to recognize violent behaviors. Comparing other existing deep-learning-based methods in **Table 4**, our method can also achieve excellent recognition performance. This is mainly because multi-task self-supervised mechanism and attention module can further enhance the semantic representation. Obviously, it is worth noting that AMCNN can achieve the best accuracy on two benchmark datasets comparing to traditional methods and deep-learning-based methods, further demonstrating the effectiveness of the proposed AMCNN.

Finally, we comprehensively conclude experiments on the Hockey Fight dataset and Movies dataset with respect to the value of area under curve (AUC), sensitivity and specificity. **Table 5** shows metrics of our proposed method, the value of AUC, sensitivity and specificity on Hockey are respectively 97.8%, 96% and 96%, and the value of AUC, sensitivity and specificity on Movies are all 100%. It's obvious that the performance on the Movies is higher, which can prove that our proposed method performs better in severe crowded scenes.

Table 5. Metrics of the proposed method

Methods	AUC	Sensitivity	Specificity
Hockey	97.8%	96%	96%
Movies	100%	100%	100%

4. Conclusion

In this paper, we propose a visual violence recognition algorithm based on multi-task self-supervised learning and attention mechanism, in which recognizing violent behaviors is regard as the main task, and reconstructing frames is regard as the self-supervised auxiliary task. The accuracy of violence recognition in the main task can be further enhanced by the balance-weight multi-task self-supervised mechanism. Additionally, we adopt the spatial attention mechanism as the mask branch and add it into the master branch of CNN in order to increase the receptive fields for improving the semantic representation of latent features. Experimental results demonstrate that the AMCNN model can actually accomplish accurate recognition when compared to other existing methods. And researchers should concentrate on dealing with complicated environments in the future to better preserve safety.

5. Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant 61473066 and Grant 61601109, in part by the 2023 Hebei Provincial doctoral candidate Innovation Ability training funding project under Grant CXZZBS2023170 and in part by the Natural Science Foundation of Hebei Province under Grant F2021501020.

References

[1] T. Zhang, W. Jia, X. He and J. Yang, "Discriminative Dictionary Learning with Motion Weber Local Descriptor for Violence Detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 696-709, Mar. 2017.

[2] M. Chen and A. Hauptmann, "Mosift: Recognizing human actions in surveillance videos," *Annals of*

pharmacotherapy, vol. 1, no. 1, pp.1-16, Sep. 2009.
[3] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *in Proc. ICLR*, San Diego, CA, USA, 2005, pp. 886-893.
[4] D. D. Das and S. H. Shaikh, "A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector," *The Visual Computer*, vol. 32, no. 3, pp. 289-306, Mar. 2016
[5] I. Laptev, "On space-time interest points," *International journal of computer vision*, vol. 64, no. 2, pp. 107-123, Sep. 2005.
[6] N. Dalal, B. Triggs and C. Schmid, "Human detection using oriented histograms of flow and appearance," *in European conference on computer vision*, Berlin, Germany, 2006, pp. 428-441.
[7] A. Krizhevsky, I. Sutskever and G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Communications of the ACM*, vol. 60, no. 6, pp.84-90, Jun. 2017.
[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *in Proc. NIPS, Lake Tahoe*, CA, USA, 2012, pp. 1097–1105.
[9] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *in Proc. ICLR, San Diego*, CA. UAE, 2015, pp. 1–14.
[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *in Proc. CVPR*, Las Vegas, NV, USA, 2016, pp. 770–778.
[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *in Proc. CVPR*, Columbus, Ohio, USA, 2014, pp. 580–587.
[12] W .Y. Zou, X. Wang, M Sun and Y. Q. Lin, "Generic object detection with dense neural patterns and regionlets," *arXiv preprint*, vol. 1, no. 1, pp. 1-9, Apr. 2014.
[13] W. X. Luo, W. Liu, D. Z. Lian, J. H. Tang, L. X. Duan, X. Peng and S. H. Gao, "Video Anomaly Detection with Sparse Coding Inspired Deep Neural Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 3, pp. 1070-1084, Mar. 2021.
[14] P. Wu, J. Liu and F. Shen, "A deep one-class neural network for anomalous event detection in complex scenes," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 7, pp. 2609-2622, Jul. 2020.
[15] A. J. Naik and M. T. Gopalakrishna, "Deep-violence: individual person violent activity detection in video," *Multimedia Tools and Applications*, vol. 80, no. 12, pp. 18365-18380, Feb. 2021.
[16] P Wang, P Wang and E Fan, "Violence detection and face recognition based on deep learning," *Pattern Recognition Letters*, vol. 142, no. 1, pp. 20-24, Feb. 2021.
[17] A. Srivastava, T. Badal and P. Saxena P, A. Vidyarthi and R. Singh, "UAV surveillance for violence detection and individual identification," *Automated Software Engineering*, vol. 29, no. 1, pp. 1-28, Mar. 2022.
[18] R. Halder and R. Chatterjee, "CNN-BiLSTM model for violence detection in smart surveillance," *SN Computer science*, vol. 1, no. 1, pp. 1-9, Jun. 2020.
[19] R. Caruana, "Multitask Learning," *Machine Learning*, vol. 28 no.1, pp. 41–75, Jul. 1997.
[20] S. Vandenhende, S. Georgoulis, W. V. Gansbeke, M. Proesmans, D. Dai and L. V. Gool, "Multi-Task Learning for Dense Prediction Tasks: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3614-3633, July. 2022.
[21] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *in Proc. ICML*, Lile, France, 2015, pp. 448–456.
[22] V. Nair and G. Hinton, "Rectified linear units improve restricted Boltzmann machines," *in Proc. ICML*, Haifa, Israel, 2010, pp. 807–814.

[23] N. E. Bermejo, S. O. Deniz, G. G. Bueno and Rahul Sukthankar, "Violence detection in video using computer vision techniques," *in International conference on Computer analysis of images and patterns*, Berlin, Germany, 2011, pp. 332-339.
[24] K. Deepak, L. K. P. Vignesh, G. Srivathsan, S. Roshan and S. Chandrakala, "Statistical Features-Based Violence Detection in Surveillance Videos", *in Cognitive Informatics and Soft Computing, Singapore*, 2020, pp. 197–203.
[25] T. Senst, V. Eiselein, A. Kuhn and T. Sikora, "Crowd Violence Detection Using Global Motion-Compensated Lagrangian Features and Scale-Sensitive Video-Level Representation," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 12, pp. 2945-2956, Dec. 2017.
[26] L. Xu, C. Gong, J. Yang, Q. Wu and L. Yao, "Violent video detection based on MoSIFT feature and sparse coding," *in Proc. IEEE ICASSP*, Florence, Italy, Jul, 2014, pp. 3538-3542.
[27] P. Bilinski and F. Bremond, "Human violence recognition and detection in surveillance videos," *in Proc. IEEE AVSS*, Colorado, CO, USA, 2016, pp.30-36.
[28] T. Hassner, Y. Itcher and O. Kliper, "Violent flows: real-time detection of violent crowd scene," *in Proc. IEEE CVPRW*, Providence, USA, 2012, pp. 1-6.
[29] I. Serrano, O. Deniz, J. L. Espinosa and G. Bueno, "Fight Recognition in Video Using Hough Forests and 2D Convolutional Neural Network," *IEEE Transactions on Image Processing*, vol. 27, no. 10, pp. 4787-4797, Oct. 2018.
[30] A. Mumtaz, A. B. Sargano and Z. Habib, "Violence detection in surveillance videos with deep network using transfer learning," *in Proc. IEEE EECS*, Bern, Switzerland, 2018, pp. 558-563.
[31] C. Ding, S. Fan, M. Zhu, W. Feng and B. Jia, "Violence detection in video by using 3D convolutional neural networks," *in Proc. Springer ISVC*, Nevada, USA, 2014, pp. 551-558.
[32] S. A. Carneiro, G. P. Silva, S. J. F. Guimarães and H. Pedrini, "Fight detection in video sequences based on multi-stream convolutional neural networks," *in Proc. SIBGRAPI*, Brazil, 2019, pp. 8–15.
[33] M. Sharma and R. Baghel, "Video surveillance for violence detection using deep learning," *in Proc. Advances in Data Science and Management*, Singapore, 2020, pp. 411-420.
2015, pp. 2758–2766.