An Efficient Violence Detection Method Based on Temporal Attention Mechanism

WANG Binxu, ZHANG Xuguang

(*The Communication Engineering Department*, *HangZhou Dianzi University*, *Hang Zhou*, *China*, *310020*)

Abstract: Violence detection is very important for public safety. However, violence detection is not an easy task. Because recognizing violence in surveillance video requires not only spatial information but also sufficient temporal information. In order to highlight the time information, we propose an efficient deep learning architecture for violence detection based on temporal attention mechanism, which utilizes pre-trained MobileNetV3, convolutional LSTM and temporal attention block Temporal Adaptive (TA). TA block can focus on further refining temporal information from spatial information extracted from backbone. Experimental results show the proposed model is validated on three publicly datasets: Hockey Fight, Movies, and RWF-2000 datasets.

Keywords: Violence Detection, Temporal Attention, Convolutional LSTM, CNN-RNN

1 Introduction

Violent behavior usually refers to behaviors in which people solve problems or express emotions by using force, hitting others, inflicting harm, etc. It is very dangerous. Especially in some group incidents, since many different groups are involved, once a violent conflict occurs, it may cause irreparable consequences. Therefore, it is very urgent to detect violent behavior in time and intervene.

With the development of urbanization and the concept of smart cities[1]. Urban safety becomes an important issue, in order to build a smart and safe city more and more surveillance cameras are being installed in all corners of the city to detect acts of violence. However, detecting violent behavior detection is different from general computer vision tasks and faces many challenges. It is a multi-branch integrated task, which needs to combine target tracking, action recognition and so on. Also, since violence occurs not instantaneously but continuously during a certain period of time, the recognition of violence also requires the use of temporal information, which can only be solved by fusing temporal and spatial information.

Before deep learning became popular in the field of computer vision, researchers mostly used traditional methods to create manual models for extracting features from video frames. Hassner et al.[2] used the Violent Flows (ViF) descriptor which can take advantage of the magnitude series of optical flow over time for violence detection. Gao et al.[3] proposed a novel feature extraction method named Oriented VIolent Flows (OViF) which make the orientation of the violent flow features into the ViF descriptor. However, handcrafted framework have a huge disadvantage in generalization ability, and it cannot reliably and efficiently process data that has never been seen before. With the widespread application of deep learning in computer vision, CNN network architecture has shown significant advantages in image recognition. Sudhakaran et al.[4] combined convolutional neural networks and recurrent neural networks to extract spatial information using convolutional neural networks and temporal information using recurrent neural networks, and achieved success in violence detection. Despite its effectiveness in modeling sequential data, Recurrent Neural Networks (RNNs) suffer from a significant drawback[5] in which the information from the first input is diluted or overwritten by subsequent inputs. This issue becomes increasingly severe as the sequence length grows, thus limiting the capability of RNNs in capturing long-term dependencies.To remedy this drawback, we propose an efficient violence detection method based on the temporal attention mechanism, which uses the temporal attention mechanism to strengthen temporal features and thus avoid the problem of dilution of temporal feature information after input to the recurrent neural network.

The proposed method makes several key contributions, including:

- A combination of MobileNetV3 and Convolutional LSTM is used for anomaly detection from surveillance cameras.
- In order to solve the problem that the time information produced by LSTM is lost when the long sequence list picture is input, we use the TA module to strengthen the time information and then input the picture frame into the LSTM.

In the remainder of this paper, we present our analysis of related work in Section 2. We then provide a detailed description of our proposed model in Section 3, followed by an explanation of our training methods and experimental results on different datasets in Section 4. Finally, we conclude with a summary of our work.

2 Related Work

The deep learning model needs to extract a large amount of feature information for training to ensure the robustness of the model. For building a model for violent behavior detection, it is necessary to extract a large amount of Spatio-temporal information. The Spatio-temporal feature means that the model must obtain the spatial information of the object in one frame of video and obtain the temporal information of the object movement between different frames.To achieve the above goals, Nievas et al.[6] used the well-known Bag-of-Words framework used for violence detection and they also constructed two commonly used standard datasets which is Hockey Fight, Movies.

Following the success of convolutional neural networks, Seranno et al.[7] proposed a hybrid framework that combines handcrafted and learned features to detect violent behavior. This approach utilizes both Hough Forests and 2D CNNs, resulting in an effective method for identifying instances of violence. However, due to the singularity of features and poor generalization ability of handcrafted models, it is difficult to produce excellent performance in complex environments. Since its inception, deep learning has revolutionized the field of computer vision, and many tasks related to violence detection have been developed using end-to-end trainable neural network architectures that require no preprocessing and are highly effective. Inspired by the success of Two-Stream networks[8] in the field of action recognition, Dai et al.[9] employed a dual-stream CNN framework to extract features from both static frames and motion optical streams, allowing for a more comprehensive analysis of the data. They then used SVM for the final classification. To solve the drawback that CNN structural model can only extract spatial feature information but not temporal feature information, Dong et al.[10] utilized a CNN-LSTM framework to extract spatial-temporal features, but this approach may not fully preserve the spatial features extracted by the CNN. Nonetheless, it allowed for effective analysis of the data's temporal relationships. sudhakaran et al.[4] proposed to use ConvLSTM[11], and ConvLSTM can extract not only spatial information but also temporal information.Islam et al.[12] created the Two-Stream network with Separable Convolutional LSTM and achieved excellent results in violence detection. Ding et al.[13] suggested the use of 3D CNN to tackle violence detection, which allows for direct processing of an entire video frame. This approach is capable of effectively capturing both spatial and temporal features in the data. In our work, we utilized MobileNetV3[14], a lightweight network that can deliver impressive results using fewer parameters.

3 Proposed Methodology

This paper presents an end-to-end trainable deep neural network that efficiently captures Spatio-temporal features in violent videos, culminating in accurate judgments.

To achieve the above effect, we propose to combine temporal attention block TA with ConvLSTM to enhance temporal features for perfect results.

3.1 Pre-processing

Typically, conventional models use optical flow as the input to capture both temporal and spatial information simultaneously. However, due to the computational complexity of optical flow data, we use differences between video frames to represent subject movement information instead. This method captures spatial information also captures some temporal information, and It is simpler in calculation than optical flow. Previous works[4][12][15] have demonstrated the effectiveness of this approach.

The calculation method of the frame difference method is shown in equation 1.

$$input_i = |frame_{i+1} - frame_i| \qquad (1)$$

In equation 1, $input_i$ represents the i-th of frame difference. $frame_i$ is i-th frame. If a video has t frames will produce $t - 1$ frames.

3.2 Model

Violence detection requires not only temporal information, but also spatial information. Using spatial information, it is possible to determine the relative position relationship between characters at a certain moment, and whether a certain behavior has occurred. In order to balance the performance and speed of the model, we used a lightweight model named MobilenetV3 to extract spatial information to help the model make more accurate judgments about violent behavior. The MobileNetV3 is pretrained on ImageNet[16] dataset. The MobileNetV3 bneck is composed of MobileNetV2[17] layer (Inverted Residual and Linear Bottleneck) and Squeeze-and-Excite[18]. In contrast with[18] MobileNetV3 applies the squeeze and excite in the residual layer. At the same time, in order to extract more temporal information, we proposed to both use temporal attention module TA and ConvLSTM. As shown in Fig.1, input the preprocessed frame difference image into MobilenetV3, after the model extracts the spatial features from MobilenetV3, then input the data into TA, and then TA extracts the time information, and then enters ConvLSTM to further extract the spatiotemporal information to enrich the spatiotemporal information proposed by the model, and finally make a correct judgment.

3.3 Temporal Adaptive block

In the video-based violence detection task, it is necessary to determine whether a person has committed

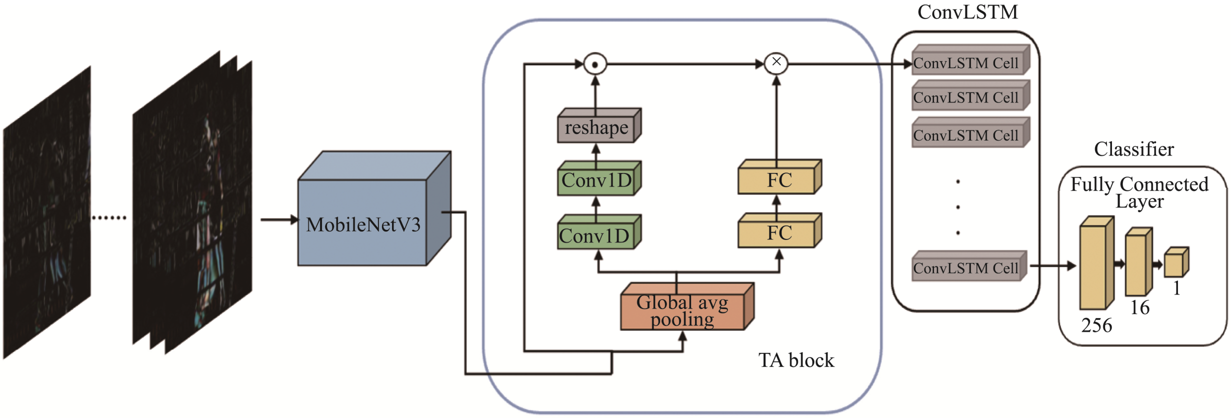


Fig.1 The Detail of Temporal Adaptive (TA) Block Shows How It Works. ⊙ Is Element-wise Multiplication, ⊗ Is Convolution Operation.

violent behavior in a series of successive actions, so the model needs to be able to obtain sufficient temporal information. Moreover, video sequences are generally long sequences, and LSTM will cause the problem of time information loss when there is a long sequence, so we use TA block which is proposed by liu et al[19]. The convolution kernel dynamically generated by the TA module can adaptively generate convolution kernels of different sizes according to different scenes of the video frame, enhance the ability to capture time information, and thus improve the accuracy and precision of identifying violent behavior. The overall architecture of the TA block is shown in Fig.2.

Generally, $X \in \mathbb{R}^{C\times T\times H\times W}$ represents the feature map of a video frame. TA block focuses only on the temporal features of video frames, so the global spatial average pooling is adopted to squeeze the feature maps. At this time $X \in \mathbb{R}^{C\times T}$ After then, performing a 1-dimensional convolution operation on X can extract the spatiotemporal features in the video frame. At the same time, the fully connected layer is combined with the global background information, and finally the softmax activation function is used to calculate the final dynamic convolution kernel.

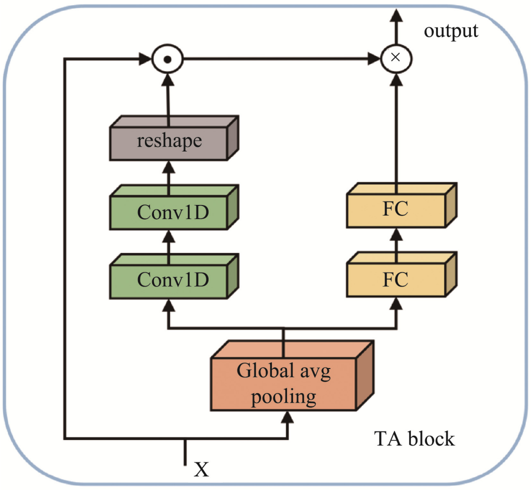


Fig.2 The Detail of Temporal Adaptive (TA) Block Shows How It Works. ⊙ Is Element-wise Multiplication, ⊗ Is Convolution Operation

3.4 Convolutional LSTM

In order to further enhance the ability of the model to extract temporal and spatial information, we adopted Convolutional LSTM which uses a convolutional network structure. ConvLSTM has the advantages of CNN and RNN at the same time. The spatial local features are extracted by CNN, and the time series features are extracted by LSTM, so that the spatiotemporal feature information in the time series can be extracted at the same time. Fig.3 demonstrates the structure of ConvLSTM, and Equations 2 represent the operations inside a ConvLSTM cell.

$$\begin{gathered}
i_t = \sigma(W_{xi} * x_t + W_{hi} * H_{t-1} + W_{ci} \cdot C_{t-1} + b_i) \\
f_t = \sigma(W_{xf} * x_t + W_{hf} * H_{t-1} + W_{cf} \cdot C_{t-1} + b_f) \\
\tilde{C}_t = W_{xc} * X_t + W_{hc} * H_{t-1} + b_c \\
C_t = f_t \cdot C_{t-1} + i_t \cdot \tanh(\tilde{C}_t) \\
O_t = \sigma(W_{xo} * x_t + W_{ho} * H_{t-1} + W_{co} \cdot C_t + b_o) \\
H_t = O_t \cdot \tanh(C_t) \qquad (2)
\end{gathered}$$

Where, σ is the activation function; ∗ represents a convolution operation; · means Hadamard product; X_t is the input of the network layer at time t; H_{t-1} is the hidden state at time t-1; H_t is result set, and the gate activationsi_t, f_t, O_t are all 3-dimensional tensors.

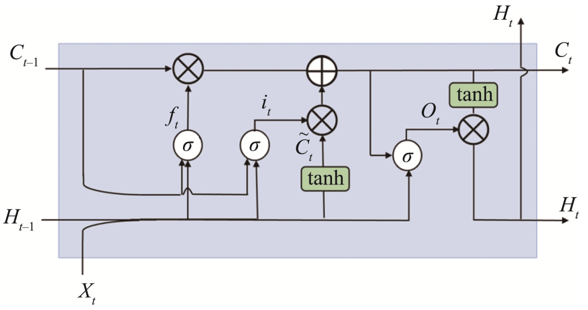


Fig.3 The Basic ConvLSTM Cell.

4 Implementation Details and Results

In this section, we will introduce some details of our network from training methods, datasets, ablation experiments, and result analysis.

4.1 Training Methodology

Two adjacent frames in a video often contain a lot of repeated information, so for each video, we use the average sampling method to extract 31 frames and use the frame difference method to process a total of 30 frames. Finally, the shape of input is 30×3×320×320.

We implemented the proposed method using the PyTorch Python library, training the model for 100 epochs or until it began overfitting. Specifically, we utilized the MobileNetV3 Large version and eliminated the final convolutional and linear layers, while leveraging weights pre-trained on the ImageNet dataset. Common image augmentation techniques are also used in training such as: random rotation, Adjusting brightness and hue. For model optimization, we used Adam[20] optimizer. The learning rate is 0.0001 and Cosine Annealing[21] is used to reduce learning rate for improve its anytime performance when training our model. The loss function is binary cross-entropy (BCE).

4.2 Datasets

To gauge the performance of our proposed model, we assessed it on three distinct violence detection datasets: the Hockey Fights[6] dataset, the Movies[6] dataset, and the RWF-2000[22] dataset.

4.2.1 Hockey

The Hockey Fights dataset consists of clips from National Hockey League (NHL) ice-hockey matches, with a total of 500 fighting and 500 non-fighting video clips. All videos share similar backgrounds and feature violent actions.

4.2.2 Movies

The Movies dataset consists of video clips showcasing action sequences from various films, whereas the non-fight scenes are collected from action recognition datasets. With a total of 200 video clips, this dataset is relatively small in size.

4.2.3 RWF-2000

The RWF-2000 dataset is a newly curated collection of real-world combat videos obtained from YouTube, consisting of 2000 video clips captured by surveillance cameras in real-life settings. Half of the videos depict violent behavior, while the other half depict normal behavior. Each video clip has a duration of 5 seconds and is captured at 30 frames per second. This dataset is particularly valuable for practical applications as all videos are captured by surveillance cameras in real-life scenarios, providing a close representation of actual violent incidents. Additionally, with a larger number of videos compared to previous violence detection datasets, the RWF-2000 dataset is well-suited for deep learning training.

4.3 Experiment on Standard Benchmark Datasets

In the experiment, 80% of the dataset is used to train our model and the rest is used to validate the performance of our model. From Table 1, we can see that our proposed method basically achieves the best performance compared with other methods.

In Table 2, we demonstrate a comparison of the parameters of our proposed model and other methods. Although our model parameter quantity is about 1M higher than Vijeikis[23] et al , our method achieves much

Table 1 A Comparison of Classification Outcomes on Prevalent Benchmark Datasets

Method	RWF-2000	Hockey	Movies
ViF [2]	–	82.90%	–
ViF + OViF [3]	–	87.50%	–
Hough Forests + CNN [7]	–	94.60%	99%
Three Streams + LSTM [10]	–	93.90%	–
ResNet50+NN [24]		96%	100%
FightNet [25]	–	97%	100%
Multi-frame Fusion [26]	–	98.80%	99.10%
3DCNN [27]	–	98.3%	100%
ConvLSTM [4]	77%	97.1%	100%
U-Net + LSTM [23]	82%	96.10%	99.50%
C3D [28]	82.75	96.50%	100%
Flow Gated Net [22]	87.25%	98.0%	100%
Two Stream SepConvLSTM [12]	89.75%	99.5%	100%
Ours	91.75%	99%	100%

Table 2 Comparison of the Parameters of Our Proposed Model with Other Models

Model	Parameters
ConvLSTM [4]	9.6M
C3D [28]	78M
U-Net + LSTM [23]	4.074M
Ours	5.17M

higher accuracy on all three benchmark datasets. Therefore, our model takes into account both the amount of parameters and the accuracy.

4.4 Ablation Studies

Table 3 examines the impact of TA blocks on our model's performance. Upon integrating the TA module, our model achieved a 91.75% accuracy rate when tested on the RWF-2000 dataset, which greatly improved the accuracy based on the previous best results, but all this only increased the parameters of the entire model by 0.04 million. This improves the accuracy by 7.75% compared to the results withour TA, which also shows that TA block can indeed enhance the temporal feature capture of long-sequence videos to make up for the shortcomings of RNN for long-sequence information loss.

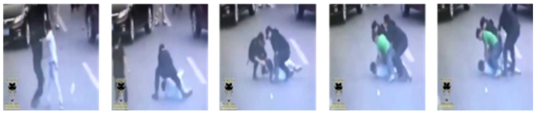
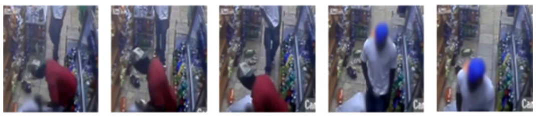
4.5 Results Analysis

Fig.4 showcases five example frames per row, highlighting both the true labels and the predicted labels that correspond to them. The initial two rows showcase instances of video clips where our model accurately predicted the outcomes. The initial example portrays characters in close contact, showcasing multiple individuals quickly punching and beating each other - clear indicators that assist the model in making accurate judgments. Similarly, in the second example, the empty scene and slow motion movements of characters offer good indications of the absence of violence, enabling our model to correctly predict non-violence.

Table 3 Analyze the Contribution of TA Blocks to Our Model on Three Datasets.

Method	RWF-2000	Hockey	Movies	Parameters
MobilenetV3+ConvLSTM	84%	97.5%	100%	5.13M
MobilenetV3+TA+ConvLSTM	91.75%	99%	100%	5.17M

Table 4 Qualitative Results of the Proposed Model for Violence Detection on the RWF-2000 Dataset.

Video Frames	Ground Truth	Predicted Label
	Violence	Violence
	NonFight	NonFight
	NonFight	Violence
	Violence	NonFight
	Violence	NonFight
	Violence	NonFight

The last four rows represent instances where our proposed model misjudges. The third and fourth rows exhibit ambiguous body movements that may lead to incorrect classifications. In the fifth example, most of the person's body engaged in the fight is occluded, leading the model to classify it as non-violent erroneously. The last row suffers from poor video quality and resolution, with a dark background that makes it difficult to discern violence accurately. These factors contribute to potential misclassifications by our model.

5 Conclusion

In this paper, we proposed an efficient violence detection method based on a temporal attention mechanism. We used the MobileNetV3 Large model with weights pre-trained on ImageNet and ConvLSTM. Our proposed network can learn spatiotemporal features simultaneously and avoid information loss caused by long time series using the Temporal Adaptive (TA) block. Through experiments on three standard benchmark datasets, we have demonstrated that our proposed method outperforms state-of-the-art techniques. Moreover, our ablation studies reveal that the incorporation of TA blocks plays a crucial role in enhancing the performance of our model. We also analyzed the quality of our model's results and provided examples of correct and incorrect predictions.Overall, our proposed method shows promising results in real-world situations and can be useful in practical applications such as surveillance systems.

References

[1] Lacinák, M., & Ristvej, J. (2017). Smart city, safety and security. Procedia engineering, 192, 522-527.

[2] Hassner, T., Itcher, Y., & Kliper-Gross, O. (2012, June). Violent flows: Real-time detection of violent crowd behavior. In 2012 IEEE computer society conference on computer vision and pattern recognition workshops (pp. 1-6). IEEE.

[3] Gao, Y., Liu, H., Sun, X., Wang, C., & Liu, Y. (2016). Violence detection using oriented violent flows. Image and vision computing, 48, 37-41.

[4] Sudhakaran, S., & Lanz, O. (2017, August). 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) (pp. 1-6). IEEE.

[5] Zhao, J., Huang, F., Lv, J., Duan, Y., Qin, Z., Li, G., & Tian, G. (2020, November). Do RNN and LSTM have long memory?. In International Conference on Machine Learning (pp. 11365-11375). PMLR.

[6] Bermejo Nievas, E., Deniz Suarez, O., Bueno García, G., & Sukthankar, R. (2011). Violence detection in video using computer vision techniques. In Computer Analysis of Images and Patterns: 14th International Conference, CAIP 2011, Seville, Spain, August 29-31, 2011, Proceedings, Part II 14 (pp. 332-339). Springer Berlin Heidelberg.

[7] Serrano, I., Deniz, O., Espinosa-Aranda, J. L., & Bueno, G. (2018). Fight recognition in video using hough forests and 2D convolutional neural network. IEEE Transactions on Image Processing, 27(10), 4787-4797.

[8] Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. Advances in neural information processing systems, 27.

[9] Dai, Q., Zhao, R. W., Wu, Z., Wang, X., Gu, Z., Wu, W., & Jiang, Y. G. (2015, September). Fudan-Huawei at MediaEval 2015: Detecting Violent Scenes and Affective Impact in Movies with Deep Learning. In MediaEval (Vol. 1436).

[10] Dong, Z., Qin, J., & Wang, Y. (2016). Multi-stream deep networks for person to person violence detection in videos. In Pattern Recognition: 7th Chinese Conference, CCPR 2016, Chengdu, China, November 5-7, 2016, Proceedings, Part I 7 (pp. 517-531). Springer Singapore.

[11] Shi, X., Chen, Z., Wang, H., Yeung, D. Y., Wong, W. K., & Woo, W. C. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. Advances in neural information processing systems, 28.

[12] Islam, Z., Rukonuzzaman, M., Ahmed, R., Kabir, M. H., & Farazi, M. (2021, July). Efficient two-stream network for violence detection using separable convolutional lstm. In 2021 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.

[13] Ding, C., Fan, S., Zhu, M., Feng, W., & Jia, B. (2014). Violence detection in video by using 3D convolutional

neural networks. In Advances in Visual Computing: 10th International Symposium, ISVC 2014, Las Vegas, NV, USA, December 8-10, 2014, Proceedings, Part II 10 (pp. 551-558). Springer international publishing.

[14] Howard, A., Sandler, M., Chu, G., Chen, L. C., Chen, B., Tan, M., ... & Adam, H. (2019). Searching for mobilenetv3. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 1314-1324).

[15] Hanson, A., Pnvr, K., Krishnagopal, S., & Davis, L. (2018). Bidirectional convolutional lstm for the detection of violence in videos. In Proceedings of the European conference on computer vision (ECCV) workshops (pp. 0-0).

[16] Liu, Z., Wang, L., Wu, W., Qian, C., & Lu, T. (2021). Tam: Temporal adaptive module for video recognition. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 13708-13718).

[17] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). Ieee.

[18] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4510-4520).

[19] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7132-7141).

[20] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[21] Loshchilov, I., & Hutter, F. (2016). Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983.

[22] Cheng, M., Cai, K., & Li, M. (2021, January). RWF-2000: an open large scale video database for violence detection. In 2020 25th International Conference on Pattern Recognition (ICPR) (pp. 4183-4190). IEEE.

[23] Honarjoo, N., Abdari, A., & Mansouri, A. (2021, April). Violence detection using pre-trained models. In 2021 5th International Conference on Pattern Recognition and Image Analysis (IPRIA) (pp. 1-4). IEEE.

[24] Zhou, P., Ding, Q., Luo, H., & Hou, X. (2017, June). Violent interaction detection in video based on deep learning. In Journal of physics: conference series (Vol. 844, No. 1, p. 012044). IOP Publishing.

[25] Asad, M., Yang, J., He, J., Shamsolmoali, P., & He, X. (2021). Multi-frame feature-fusion-based model for violence detection. The Visual Computer, 37, 1415-1431.

[26] Li, J., Jiang, X., Sun, T., & Xu, K. (2019, September). Efficient violence detection using 3d convolutional neural networks. In 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) (pp. 1-8). IEEE.

[27] Vijeikis, R., Raudonis, V., & Dervinis, G. (2022). Efficient violence detection in surveillance. Sensors, 22(6), 2216.

[28] Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE international conference on computer vision (pp. 4489-4497).

Author Biographies

WANG Binxu received the M.Sc. degree from Shenyang Aerospace University in 2021. Now he is a M.Sc. candidate at Hangzhou Dianzi University. His main research interest includes video processing.

E-mail: 211080043@hdu.edu.cn

ZHANG Xuguang received the B.Sc. degree in Electrical Technology from Northeast Normal University, and Ph.D. in Machinery and electronics engineering from Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, China, in 2000 and 2008 respectively. He was a professor at the school of electrical engineering at the Yanshan University, China. He is now a professor at the School of Communication Engineering, Hangzhou Dianzi University, China. His main research interests include video and image processing, crowd behavior analysis, and human behavior understanding.

E-mail: zhangxg@hdu.edu.cn