

A lightweight video violence detection method based on multi-scale temporal pyramid networks

Liu Pan^{1,*}, Liu Guangbin², Tadiwa Elisha Nyamasvisva¹

¹Centre for Postgraduate Studies, Kuala Lumpur University of Science and Technology, Kuala Lumpur, Malaysia

²College of Electrical and Information Engineering, Guangdong Baiyun University, Guangzhou, China

Email: 242924784@s.klust.edu.my

Abstract.

Video violence detection plays a critical role in public safety surveillance and online content moderation. This paper proposes a Multi-Scale Temporal Pyramid Network that employs three parallel temporal pyramid branches with strides of 1, 2, and 4 to explicitly model fine-grained, intermediate, and coarse temporal dynamics. The model adopts a lightweight MobileNetV3-Large backbone for spatial feature extraction, followed by spatial-to-temporal projection to generate compact token sequences. Three temporal depth-wise separable residual blocks independently process each temporal scale, while a temporal scale attention fusion module adaptively integrates multi-scale representations. Learnable temporal attention pooling then compresses variable-length sequences into a compact video-level descriptor for classification. Systematic experiments on three public benchmark datasets: RWF-2000, Hockey Fights, and Real Life Violence Situations, demonstrate that MPPN achieves competitive detection performance, validating the effectiveness of explicit multi-scale temporal modeling for practical video violence detection.

Keywords:

Violence detection;
Temporal pyramid
network; Multi
scale modeling;
Temporal attention

1. INTRODUCTION

With the widespread deployment of video surveillance systems and the explosive growth of social media video content, automatic detection of violent behavior in videos has become an urgent need in the fields of public safety and content management. Traditional manual review methods are not only costly and inefficient, but also susceptible to subjective factors^[1]. The core challenge of video violence detection lies in the effective modeling of temporal information. Unlike static image classification, the identification of violent behavior often depends on the dynamic changes between frames, the instantaneous burst of striking action, the continuous confrontation during pushing and shoving, the spatial diffusion in group conflicts, etc^{[2][3]}. How to capture these motion features at different temporal scales simultaneously within the same frame is a key challenge faced by current methods. In recent years, deep learning technology has made significant progress in the field of video understanding, providing a new technical path for the automatic detection of violent behavior. Existing mainstream methods can be broadly organized into four categories.

(1) Temporal shift methods, such as TSM, achieve efficient temporal fusion by shifting feature channels along the time axis with negligible overhead, but essentially restrict information flow to adjacent frames and lack explicit multi-scale temporal modeling^[4].

(2) 3D convolution-based methods extend 2D convolutions along the temporal dimension, jointly learning spatio-temporal features from video clips, but incur high computational overhead due to cubic parameter growth^[5].

(3) RNN-based methods sequentially aggregate frame-level CNN features through recurrent units, such as LSTM, enabling variable-length temporal modeling, yet suffer from inefficient sequential computation and difficulty in capturing multi-scale patterns^[6].

(4) Transformer-based methods model global temporal dependencies via self-attention mechanisms, achieving strong long-range modeling capacity at the cost of quadratic complexity with respect to sequence length^[7].

In summary, a persistent trade-off exists: methods with strong temporal modeling capacity impose prohibitive computational costs, while efficient alternatives sacrifice the ability to simultaneously capture motion patterns at multiple temporal scales. This motivates our Multi-Scale Temporal Pyramid Network(MPPN), which achieves multi-scale temporal understanding through parallel 1D convolutions while maintaining only approximately 3 million parameters. The main contributions of this paper include:

(1) Proposing a multi-scale temporal pyramid architecture, achieving efficient multi-scale temporal modeling through branched 1D convolutions;

(2) Designing a temporal scale attention fusion mechanism to adaptively integrate features from different time scales;

(3) Validating the effectiveness of the method on three public datasets, achieving competitive detection performance while maintaining a low parameter count.

Related work

Temporal shift lightweight method

Chen et al. proposed the DualCascade TSF-MobileNetV2 lightweight violent behavior recognition model, which uses MobileNetV2 as a lightweight backbone network and relies on depthwise separable convolution and inverse residual structure to significantly reduce the model parameters and computational overhead. Based on the temporal shift module, a dual cascaded temporal shift fusion module DualCascade TSF is designed to sequentially fuse the spatiotemporal features before and after two temporal shifts, alleviating the problem of sparse temporal features caused by multi-level temporal shifts and enhancing the model's ability to model long-term video behavior dependencies. Competitive accuracy was achieved on datasets such as RWF-2000, but with a model size of 16.99 MB, there is still room for improvement in optimization^[4].

3D Convolution-based method

Liu et al. proposed a lightweight 3D-CNN enhanced with the Squeeze and Excitation(SE) attention mechanism for campus violence detection. The SE module adaptively recalibrates channel-wise feature responses by explicitly modeling inter-channel dependencies, allowing the 3D backbone to emphasize discriminative spatio-temporal features. Through structural optimization of the C3D architecture, the model reduces the parameter count by 88% compared to the original C3D while achieving improved accuracy on a self-constructed campus violence dataset and the Hockey Fights benchmark. However, despite significant compression, the resulting model size of 36 MB remains relatively heavy for resource-constrained edge deployment. Moreover, the method relies on only 16-frame input clips, which inherently limits its capacity to capture long-range temporal dependencies and multi-scale motion patterns^[5].

RNN-based method

Vosta and Yow ^[6] proposed KianNet, an attention-based CNN-LSTM architecture for video violence detection. The model employs a ResNet50 backbone to extract per-frame spatial features, followed by a ConvLSTM layer to capture temporal dependencies across frame sequences. A Multi-Head Self-Attention (MHSA) module is strategically inserted between two ConvLSTM layers, mimicking the human vision saccade mechanism to allow the model to simultaneously attend to multiple salient regions within each frame while preserving their temporal context. However, the reliance on ResNet50 as the backbone introduces a substantial parameter overhead, resulting in a heavy model unsuitable for resource-constrained edge deployment. Additionally, the sequential nature of ConvLSTM inherently limits the model's ability to capture long-range temporal dependencies and multi-scale motion patterns, as information propagation across distant frames suffers from the well-known vanishing gradient problem in recurrent architectures.

Islam et al^[8] proposed an efficient two-stream network for video violence detection that leverages Separable Convolutional LSTM and a truncated MobileNet backbone. The method achieves 89.75% accuracy on RWF-2000 with only 0.333 million parameters. However, despite its parameter efficiency, the sequential processing nature of LSTM fundamentally constrains the model's capacity to capture long-range temporal dependencies and multi-scale motion patterns, as information propagation across distant frames remains limited by the recurrent architecture.

Transformer-based methods

Rendón-Segador et al^[7] proposed CrimeNet, a violence detection framework that combines Vision Transformer (ViT) with Neural Structured Learning(NSL). Each video frame is first transformed via optical flow, then partitioned into fixed-size patches and encoded by a ViT-S backbone to capture spatial dependencies. Simultaneously, a similarity graph is constructed using Locality-Sensitive Hashing and adversarial perturbation, encoding structural relationships among training samples. NSL then jointly minimizes both the standard supervised classification loss and a neighbor-based regularization loss, forcing the ViT to learn representations that respect the underlying sample similarity structure. The model achieves near-perfect in-domain performance, with 100% ROC-AUC on NTU CCTV Fights and UBI Fights, and 99.98% on XD Violence and UCF Crime. However, the reliance on optical flow preprocessing and adversarial neighbor generation introduces substantial computational overhead, and the method is evaluated exclusively at the frame level, requiring laborious per-frame annotations that are rarely available in practical deployment scenarios. Bai et al^[9] proposed a Transformer-based campus violence recognition method specifically designed for thermal infrared videos, addressing the limitation of conventional optical sensors under low-light and nighttime conditions. The method achieves an AUC of 88.6% on their self-constructed campus violence dataset and demonstrates an 8.3% AUC improvement brought by the U-Net-based image transformation module. However, the evaluation is restricted to a single self-collected dataset with limited scale, and the performance on widely adopted public violence detection benchmarks like RWF-2000 and Hockey Fights, remains unvalidated, making it difficult to assess the model's generalization capability. Overall, Transformer-based violence detection methods benefit from the self-attention mechanism's ability to capture long-range temporal dependencies, achieving strong performance on several benchmarks. However, the quadratic computational complexity of standard self-attention with respect to sequence length imposes high memory and inference costs, making deployment on resource-constrained edge devices challenging.

Proposed Method

Overview

To address these intertwined challenges of multi-scale modeling and computational efficiency, we propose the Multi-Scale Temporal Pyramid Network (MPPN). Figure 1 illustrates the overall architecture of the proposed Multi-Scale Temporal Pyramid Network. The network comprises five main stages:

- (1) Spatial feature extraction with a frozen MobileNetV3-Large backbone.
- (2) Spatial-to-temporal projection via 1×1 convolution and global average pooling.
- (3) Multi-scale temporal pyramid modeling through three parallel 1D convolutional branches operating at different temporal strides.
- (4) Scale-adaptive fusion that dynamically integrates branch-wise representations.
- (5) A lightweight classification head with temporal attention pooling. Unlike prior approaches that inject temporal shift modules directly into the 2D backbone, MPPN explicitly decouples spatial and temporal modeling, enabling each component to specialize and facilitating interpretability.

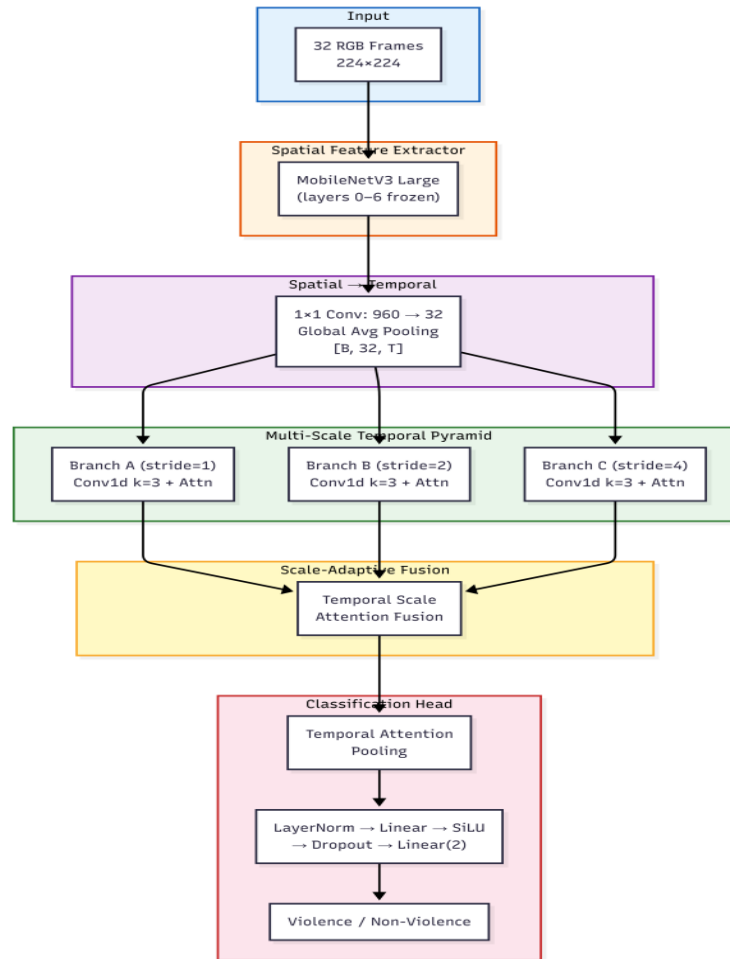


Figure 1: The overall architecture of the proposed method

Model input and output

Given an input video, we uniformly sample $T = 32$ RGB frames and resize each to $H = W = 224$ pixels, forming an input tensor $\mathbf{X} \in \mathbb{R}^{B \times T \times 3 \times H \times W}$, where B denotes the batch size. The network outputs a binary prediction $\hat{\mathbf{y}} \in \mathbb{R}^{B \times 2}$, corresponding to violence and non-violence classes. All intermediate representations and their dimensionalities are detailed in the following subsections.

Spatial Feature Extraction

We employ a MobileNetV3-Large backbone pre-trained on ImageNet for frame-level spatial feature extraction, with the first six blocks frozen during training. Each frame is processed independently, producing a feature map of size 7×7 with 960 channels per frame for each of the T time steps.

Spatial-to-Temporal Projection

To bridge the spatial and temporal modeling stages, the 960-dimensional spatial features are first compressed to $D=32$ channels through a 1×1 convolution followed by batch normalization and SiLU activation. Global average pooling then collapses the spatial dimensions, converting the frame-level feature maps into a compact temporal token sequence of shape $B \times D \times T$:

$$\mathbf{Z} = \text{GAP}(\text{SiLU}(\text{BN}(\text{Conv2D}_{1 \times 1}(\mathbf{F})))) \in \mathbb{R}^{B \times D \times T} \quad (1)$$

where each $\mathbf{z}_t \in \mathbb{R}^D$ represents the semantic embedding of frame t .

Multi-Scale Temporal Pyramid

The temporal token sequence is processed by three parallel branches with strides $s_i \in \{1, 2, 4\}$, which sample the sequence at the original, half, and quarter frame rates, respectively. This multi-rate sampling strategy enables the network to simultaneously observe motion patterns at fine-grained, intermediate, and coarse temporal resolutions.

Each branch contains a temporal depthwise separable residual block. Specifically, the sampled sequence first passes through a depthwise 1D convolution for temporal feature extraction, followed by a pointwise convolution with batch normalization and SiLU activation. A channel-wise gating mechanism is then applied to selectively recalibrate feature responses:

$$\mathbf{Y}^{(i)} \& = \mathbf{Y}^{(i)} \odot \sigma(\text{Conv1D}_{K=3}(\text{mean}(\mathbf{Y}^{(i)}))) \quad (2)$$

where σ denotes the sigmoid function and $\text{mean}()$ computes the average across the temporal dimension. This gating operation suppresses irrelevant channels while amplifying informative ones, functioning similarly to a lightweight temporal attention mechanism. Finally, a residual connection with a learnable scalar α is added to stabilize training. Branch outputs are interpolated back to the original temporal length T to enable subsequent fusion. Stochastic branch dropout with probability $p=0.1$ is applied during training to prevent co-adaptation among branches.

Scale-Adaptive Fusion

Once the three branches produce aligned feature sequences, a scale-adaptive fusion module dynamically integrates their representations. The outputs are stacked along a new dimension, and a set of learned per-timestep, per-scale importance weights is computed via a softmax-normalized 1D convolution over the temporally-averaged branch features:

$$\mathbf{W} = \text{softmax}\left(\frac{\text{Conv1D}_{K=3}(\text{mean}(\mathbf{S}, \text{dim}=2))}{\tau}\right) \quad (3)$$

where S denotes the stacked features, and the temperature parameter τ controls the sharpness of the attention distribution. The fused representation is obtained by element-wise multiplying each branch's output with its corresponding weight and summing across branches. A final 1×1 convolution followed by batch normalization and SiLU activation projects the fused features to the output dimension. This design allows the fusion module to adaptively emphasize, for example, the fine-scale branch during fast striking actions and the coarse-scale branch during prolonged confrontations, enabling context-dependent multi-scale integration.

Temporal Attention Pooling

To aggregate the fused temporal sequence into a compact video-level representation, we employ a learnable temporal attention pooling mechanism:

$$\begin{aligned} \mathbf{P} &= \text{softmax}(\text{Conv1D}_{K=3}(\mathbf{F}_{\text{out}})) \in \mathbb{R}^{B \times 1 \times T}, \\ \mathbf{v} &= \sum_{t=1}^T \mathbf{P}_{:,1,t} \cdot \mathbf{F}_{\text{out},:,t} \in \mathbb{R}^{B \times D} \end{aligned} \quad (4)$$

A 1D convolution maps the fused features to a single-channel attention score for each time step. After softmax normalization, the resulting weights quantify the relative importance of each frame to the final decision. The weighted sum across all T time steps produces a fixed-dimensional video descriptor $\mathbf{v} \in \mathbb{R}^D$. Unlike simple average pooling, this mechanism enables the model to focus on the most discriminative frames—such as the moment of physical impact—while suppressing irrelevant or redundant temporal segments.

Classification

The video representation $\mathbf{v}$ is classified through a two-layer MLP with a hidden dimension of 32. Layer normalization and SiLU activation are applied before the hidden layer, and dropout rates of 0.5 and 0.3 are used before and after the hidden layer, respectively, to mitigate overfitting. The final linear layer produces a two-dimensional output for binary violence/non-violence classification. The network is trained with cross-entropy loss using label smoothing with coefficient $\epsilon=0.2$:

$$\mathcal{L} = -\frac{1}{B} \sum_{b=1}^B \sum_{c=0}^1 \tilde{y}_{b,c} \log(\text{softmax}(\hat{\mathbf{y}}_b)_c) \quad (5)$$

where $\tilde{y}_{b,c}$ is the smoothed target label. Label smoothing prevents the model from becoming overconfident on training samples and improves generalization to unseen data.

Experiment analysis

Datasets

The experiments were conducted on three publicly available benchmark datasets.

(1) Hockey Fights^[10]: Contains 1000 hockey game video clips, 500 of which include fight scenes and 500 are normal game footage. The videos feature diverse perspectives, including challenging scenarios such as fast movement, occlusion, and crowd interaction.

(2) RWF-2000^[10]: Contains 2000 violent and non-violent video clips from 2000 surveillance perspectives, with 1000 samples from each. Each video is approximately 5 seconds long and runs at 30fps. The datasets are divided into training, validation, and test sets in a 7:2:1 ratio.

(3) Real-Life Violence Situations(RLVS)^[12]: Contains 2000 real-life scene video clips, covering various forms of violence such as street fights and school bullying, as well as non-violent scenarios such as daily activities and sports. The videos come from a wide range of sources and vary

significantly in perspective, lighting, and resolution, placing high demands on the model's generalization ability.

Experimental setup

The backbone network used MobileNetV3-Large, which was pre-trained on ImageNet, with parameters frozen for first 6 blocks. Input resolution was 224x224, batch size was 8, using the AdamW optimizer with an initial learning rate of 2×10^{-5} , weight decay of 1.5×10^{-2} , and cosine annealing to 1×10^{-6} . Label smoothing coefficient was 0.2. The maximum of epochs is 100, with the early stop for 10. All experiments were conducted on a single NVIDIA RTX 5060 Laptop GPU.

Evaluation metrics

Evaluation metrics include accuracy, F1 score, and AUC. Accuracy measures the overall proportion of correctly classified samples among all predictions, providing a holistic view of model performance. Precision quantifies the fraction of video clips predicted as violent that are genuinely violent, reflecting the model's ability to avoid false alarms, which is critical in surveillance scenarios where excessive false positives can overwhelm human operators. Recall measures the proportion of actual violent instances that are successfully detected, indicating the model's sensitivity to identifying all violent events and its capacity to minimize missed detections. Since precision and recall often trade off against each other, the F1 score serves as their harmonic mean, offering a balanced single-number summary of detection quality. AUC evaluates the model's discriminative ability across all possible classification thresholds, assessing its overall ranking capacity irrespective of the specific decision boundary. In addition to detection performance, we also report the number of parameters as an efficiency metric, which directly reflects the model's memory footprint and inference cost, serving as a practical indicator of deployability on resource-constrained edge devices commonly used in real-world surveillance systems. Together, these complementary metrics provide a thorough and multi-faceted assessment of both detection reliability and practical deployability in real-world violence detection systems.

Analysis of Experimental Results

This experiment focuses on the performance on three open datasets, RWF-2000, Hockey Fights and RLVS, which was shown on Table 1.

Table 1: Performance of MPPN on three benchmark datasets

Datasets	Accuracy(%)	F1-Score(%)	AUC
RWF-2000	82.75	83.37	0.9014
Hockey Fights	97.00	97.03	0.9816
RLVS	97.50	97.46	0.9991

As shown in Table 1, MPPN achieves 82.75% accuracy on RWF-2000, substantially higher performance on Hockey Fights at 97.00%, and RLVS get the accuracy of 97.50%. A striking feature of the results is the near-identical accuracy and F1-score on every dataset, with the gap never exceeding 0.5 percentage points, which indicates that the model maintains a well-balanced precision–recall trade-off with no systematic class bias. MPPN achieves strong discriminative performance across all three datasets, as reflected by the AUC values. On RLVS, the near-perfect AUC of 0.9991 indicates that the model almost perfectly separates violent and non-violent clips. On Hockey Fights, the high AUC of 0.9816 reflects the benefit of fixed rink-side viewpoints, which provide consistent visual context for capturing fight-specific motion patterns. On RWF-2000, the AUC of 0.9014, while lower than the other two, still represents strong discriminative ability considering this dataset's inherent difficulty: it comprises uncurated surveillance footage captured from diverse and often distant camera angles under uncontrolled lighting, where fine-grained motion cues are significantly harder to resolve.



Figure 2: The Grad-CAM of violence(left) and non-violence(right) on RLVS

Figure 2 is the Grad-CAM visualizations of MPPN on RLVS. In violent scenes, high-response regions are sharply concentrated on the interacting body parts, notably the arms, and upper torsos of individuals engaged in physical altercation, mirroring the high-intensity, localized motion cues characteristic of aggressive behavior. In non-violent scenes, activation is broadly distributed across the entire human figure without pronounced concentration on any specific body region, reflecting the absence of anomalous localized motion. Across both classes, the model consistently attends to human body regions while largely ignoring background elements such as walls and floors, confirming that MPPN relies on semantically meaningful, task-relevant features rather than spurious scene-level correlations.

Table 2: The performance on Hockey Fights Datasets based on different methods

Methods	Accuracy(%)	F1-Score(%)	Parameters(M)
Ref [5]	93.66	-	-
Ref [8]	89.75	-	0.333M
Ref [13]	97.2	95.1	4.47M
Ref [14]	95.5	-	0.05M
Ref [15]	96.5	-	-
Ours	97	97.03	3.01 M

Table 2 presents the performance comparison between MPPN and several representative lightweight deep learning methods on the Hockey Fights dataset. Our method achieves competitive performance on both accuracy and F1-score, each reaching approximately 97%, with the two metrics nearly identical, indicating a well-balanced precision–recall trade-off free of systematic class bias.

In terms of parameter efficiency, MPPN maintains a significantly more compact architecture than the top-performing competitor, requiring substantially fewer parameters while delivering comparable detection accuracy. Compared to the most lightweight alternative, which achieves only moderate accuracy at an extremely small parameter budget, our method demonstrates that aggressive parameter reduction inevitably compromises detection performance, underscoring the importance of a carefully designed temporal modeling mechanism rather than simply minimizing model size. Notably, several compared methods either lack F1-score reporting or omit parameter count entirely, making a complete and fair assessment difficult.

Taken together, MPPN achieves an optimal balance between detection performance and computational efficiency: it delivers top-tier accuracy rivaling the best-performing lightweight

model while preserving a compact parameter footprint, demonstrating that explicit multi-scale temporal modeling can close the performance gap with heavier architectures without incurring their computational cost, a critical advantage for practical deployment in resource-constrained real-world surveillance scenarios.

Conclusion

This paper proposes a lightweight multi-scale temporal pyramid network for video violence detection. The network independently models temporal relationships at original, half, and quarter frame rates through three temporal convolution branches with varying stride lengths, enabling efficient multi-scale motion pattern capture. The temporal scale attention fusion module adaptively integrates multi-scale features and learns temporal attention pooling to focus on key frames. Experimental results on three public benchmark datasets: RWF-2000, Hockey Fights, and Real-Life Violence Situations, demonstrate that MPPN not only achieves highly competitive detection performance, surpassing several heavier counterpart models with substantially larger parameter budgets, but also maintains a compact architecture of only approximately 3 million parameters, striking a favorable balance between accuracy and efficiency that makes it well-suited for real-world surveillance deployment. Future work will explore self-supervised pre-training strategies to reduce reliance on labeled data; integrate object detection for person localization to achieve interpretable fine-grained violence recognition and deploy on edge devices to validate the model's inference efficiency in real-world surveillance scenarios.

References

- Huszar, V. D., Adhikarla, V. K., Negyesi, I., & Krasznay, C. (2023). Toward fast and accurate violence detection for automated video surveillance applications. *IEEE Access*, 11, 18772-18793.
- Binxu, W. A. N. G., & ZHANG, X. (2023). An Efficient Violence Detection Method Based on Temporal Attention Mechanism. *Instrumentation*, 10(2).
- Ullah, F. U. M., Obaidat, M. S., Ullah, A., Muhammad, K., Hijji, M., & Baik, S. W. (2023). A comprehensive review on vision-based violence detection in surveillance videos. *ACM Computing Surveys*, 55(10), 1-44.
- Chen, Y., Li, Y., Li, S., Lv, S., & Lin, F. (2025). DualCascadeTSF-MobileNetV2: a lightweight violence behavior recognition model. *Applied Sciences*, 15(7), 3862.
- Liu, P., Yan, S., Li, C., & Nyamasvisva, T. E. (2025, September). Lightweight 3D Convolutional Neural Network with SE Attention for Campus Violence Recognition. In *2025 7th International Conference on Internet of Things, Automation and Artificial Intelligence (IoTAAI)* (pp. 240-243). IEEE.
- Vosta, S., & Yow, K. C. (2023). KianNet: A violence detection model using an attention-based CNN-LSTM structure. *IEEE Access*, 12, 2198-2209.
- Rendón-Segador, F. J., Álvarez-García, J. A., Salazar-González, J. L., & Tommasi, T. (2023). Crimenet: Neural structured learning using vision transformer for violence detection. *Neural networks*, 161, 318-329.
- 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 Networ*
- Bai, Y., Yang, S., Zhang, Z., Jiang, P., Tang, L., & Yu, X. (2024, December). Recognition of Campus Violence Based on Transformer Model Using Thermal Infrared Videos. In *2024 5th International Conference on Information Science and Education (ICISE-IE)* (pp. 238-242). IEEE.

- 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
- Cheng, M., Cai, K., & Li, M. (2019). RWF-2000: An Open Large Scale Video Database for Violence Detection. <http://arxiv.org/abs/1911.05913>
- Soliman, M. M., Kamal, M. H., Nashed, M. A. E. M., Mostafa, Y. M., Chawky, B. S., & Khattab, D. (2019, December). Violence recognition from videos using deep learning techniques. In *2019 ninth international conference on intelligent computing and information systems (ICICIS)* (pp. 80-85). IEEE.
- Khan, H., Yuan, X., Qingge, L., & Roy, K. (2025). Violence Detection From Industrial Surveillance Videos Using Deep Learning. *IEEE Access*, 13, 15363–15375. <https://doi.org/10.1109/ACCESS.2025.3531213>
- Dündar, N., Keçeli, A. S., Kaya, A., & Sever, H. (2024). A shallow 3D convolutional neural network for violence detection in videos. *Egyptian Informatics Journal*, 26, 100455.
- Xia, X., Wu, H., & Yang, C. (2021). Violence detection with two-stream neural network based on C3D. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 15(4), 1-17.