Neurocomputing 514 (2022) 148–161

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Semantic multimodal violence detection based on local-to-global embedding

Yujiang Pu [a], Xiaoyu Wu [a,*], Shengjin Wang [b], Yuming Huang [c], Zihao Liu [a], Chaonan Gu [a]

[a] State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing 100024, China
[b] Department of Electronic Engineering, Tsinghua University, Beijing 100080, China
[c] Department of Neurosurgery, University of Massachusetts Medical School, Worcester, MA, USA

ARTICLE INFO

Article history:
Received 31 August 2021
Revised 16 August 2022
Accepted 15 September 2022
Available online 20 September 2022
Communicated by Zidong Wang

Keywords:
Violence detection
Semantic embedding
Multimodal fusion
Deep learning

ABSTRACT

Automatic violence detection has received continuous attention due to its broad application prospects. However, most previous work prefers building a generalized pipeline while ignoring the complexity and diversity of violent scenes. In most cases, people judge violence by a variety of sub-concepts, such as blood, fighting, screams, explosions, etc., which may show certain co-occurrence trends. Therefore, we argue that parsing abstract violence into specific semantics helps to obtain the essential representation of violence. In this paper, we propose a semantic multimodal violence detection framework based on local-to-global embedding. The local semantic detection is designed to capture fine-grained violent elements in the video via a set of local semantic detectors, which is generated from a variety of external word embeddings. Also, we introduce a global semantic alignment branch to mitigate the intra-class variance of violence, in which violent video embeddings are guided to form a compact cluster while keeping a semantic gap with non-violent embeddings. Furthermore, we construct a multimodal cross-fusion network (MCN) for multimodal feature fusion, which consists of a cross-adaptive module and a cross-perceptual module. The former aims to eliminate inter-modal heterogeneity, while the latter suppresses task-irrelevant redundancies to obtain robust video representations. Extensive experiments demonstrate the effectiveness of the proposed method, which has a superior generalization capacity and achieves competitive performance on five violence datasets.

1. Introduction

Violence not only seriously affects the physical and mental health of individuals but also endangers social stability. Psychological studies [1–3] show that adverse environmental stimuli around adolescents, especially long-term violence exposure [4], can fuel their violent emotions and even cause impulsive behaviors. The violent content of media represented by news reports, film and television works, and online games enriches the resources for individuals to imitate aggressive behaviors [5,6]. Relying on manual regulation of these contents can no longer meet the practical needs of today's society. Therefore, autonomous violence detection is of great significance for preventing criminal acts and maintaining public security.

Since most violent videos are associated with specific entities and actions, some early work focuses mainly on rich visual information, with shallow features represented by handcrafted features used for characterization. [7] proposes the DOVE system, which determines violence by detecting whether a video frame contains blood within skin regions. Hassner et al. [8] use information about the optical flow between adjacent frames to express violent motion features. Later, it has been realized that audio contains violent information that may not be covered by visual elements, such as gunshots, explosions, screams, etc., which also play a non-negligible role in violence detection. In [9], Zajdel et al. utilize a mid-level descriptor of scream to establish complementarity between audio and video sensing, aiming to detect instances of human aggression in public. [10] composes the audio view under weak supervision to detect violent scenes in movies, and the video and audio classifiers are jointly trained to yield potent decision scores.

The rise of deep neural networks has made efficient violence detection possible. Convolutional neural networks, or CNNs, are widely used to extract static frame features, which are generally fed into a long short-term memory (LSTM) network [11] for temporal encoding. [12] constructs a two-stream structure to extract appearance features and motion features, respectively, followed

* Corresponding author.
E-mail address: wuxiaoyu@cuc.edu.cn (X. Wu).

https://doi.org/10.1016/j.neucom.2022.09.090
0925-2312/

by an LSTM network to capture long-range dependencies. Both [13,14] exploit a convolutional LSTM network for spatiotemporal modeling of visual features. In recent years, 3D convolutional networks are introduced to capture the spatiotemporal correlation of videos simultaneously. Song et al. [15] develop a modified 3D-CNN using keyframe localization to distinguish violence. In [16], Li et al. construct an efficient 3D convolutional network for end-to-end violence recognition, which significantly outperforms the previous ConvLSTM structure with fewer parameters. Gu et al. [17] cascade a P3D network with an LSTM network to capture the spatiotemporal dependences of the video clip, and an early fusion of three modalities is introduced to obtain robust violence representations.

However, most existing studies focus on generalized violence and ignore the complex semantics of violent scenarios. We believe that the high-level cognition of violence comes from specific entities or behaviors, such as blood, gunshots, fights, explosions, etc, as shown in Fig. 1. The combination of these entities or behaviors somehow forms the abstract human understanding of violence. We also notice that these entities and behaviors of violence exhibit a clear co-occurrence trend: physical conflicts such as abuse, rape, and fights are usually accompanied by blood, ropes, and cudgels, while riots and wars are generally associated with screams, explosions, and gunshots. The complex scenarios lead to the distribution of violent videos with significant intra-class variance, while some dramatic non-violent videos are prone to false alarms. In addition, with a wide range of data sources, including movies, TV dramas, surveillance cameras, and Internet videos, it becomes challenging work to build a unified framework to accommodate datasets across different source domains.

To address the above problems, we propose a semantic multimodal violence detection model with local-to-global embedding, which dynamically captures local violence elements while recalibrating the abstract semantics of violent videos. Specifically, we propose two branches corresponding to the above issues: (1) **Local Semantic Detection** translates abstract violence into specific sub-concepts, which we call local semantics, and these local semantics are dynamically captured by a set of local semantic detectors generated from external word embeddings. Thus, different violent videos can be detected in a fine-grained way to improve classification performance. (2) **Global Semantic Alignment** aims to re-calibrate the abstract semantics across different violent videos, thus alleviating the intra-class variance. The global semantic descriptor constructed by violence word embeddings guides the violent class to cluster in a joint semantic space while maintaining a distance from the non-violent class. This allows erroneous samples with similar abstract semantics to be further corrected. To validate the effectiveness of the proposed method, we conduct experiments on five different violence datasets. Comprehensive results indicate that our model outperforms other approaches and achieves competitive results on these datasets. To the best of our knowledge, this is the first work to introduce external semantics for violence detection.

To summarize, the main contributions of this paper are threefold.

- We propose a Local Semantic Detection (LSD) branch to dynamically capture diverse sub-concepts of violence by a set of local semantic detectors, which is generated from external word embeddings updated by a graph convolutional network.
- We introduce a Global Semantic Alignment (GSA) strategy to recalibrate the abstract semantics across violent videos, where a global semantic descriptor is constructed to guide the clustering of violent samples while enlarging the semantic gap from non-violent class.
- We build a Multimodal Cross-fusion Network (MCN) to eliminate the heterogeneity gap and suppress irrelevant redundancies across modalities, which consists of a cross-adaptive module and a cross-perceptual module.

The remainder of this paper is organized as follows. In Section 2, we explore the related work in violence detection. In Section 3, the proposed semantic multimodal violence detection framework is further specified. In Section 4, we present the setup details and

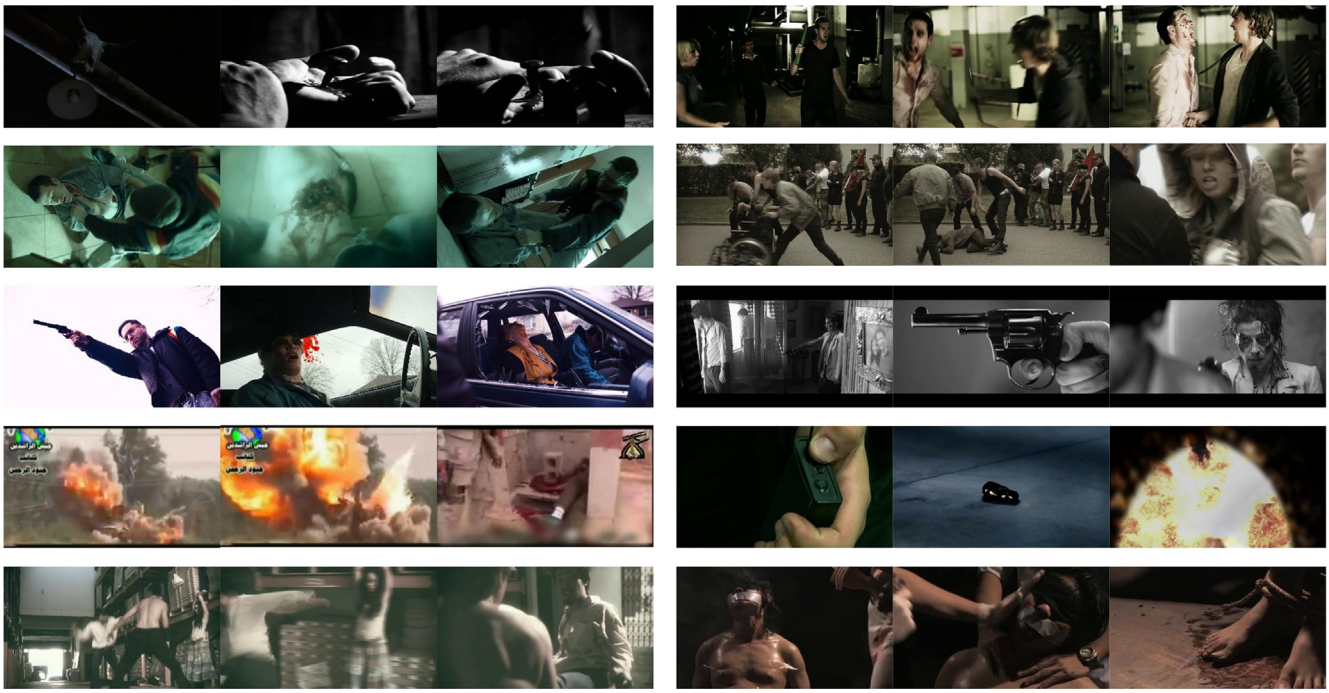

Fig. 1. Complex semantics of violence in diverse scenarios.

analyze the experimental results. Finally, in Section 5, we conclude the paper and discuss possible future work.

2. Related work

The goal of violence detection varies according to different dataset types, and our approach focuses on determining whether a video clip is violent or not. In this section, we mainly review some existing violence detection methods, including traditional methods and deep learning methods. In general, deep learning methods achieve superior performance with powerful encoding capacities while traditional methods have better theoretical interpretability.

2.1. Traditional methods

In earlier work, researchers mainly use handcrafted features for violence detection. Manually designed descriptors like Scale-Invariant Feature Transform (SIFT) [18] and Histogram of Oriented Gradients (HOG) [19] are used to characterize visual features, which are then fed into a specific classifier (e.g., SVM, k-Nearest Neighbor, Naive Bayes) for classification.

Common motion descriptors [20–22] including Space–Time Interest Points (STIP), improved Dense Trajectories (iDT), and Histograms of Oriented Optical Flow (HOOF) are also widely used for violence detection. Gao et al. [23] propose an improved OViF descriptor to detect violence in a crowd. Zhang et al. [24] develop an extension of spatial descriptors by adding temporal components to capture local motion information. HOMO is proposed in [25], which obtains amplitude and directional gradients by comparing the optical flow of adjacent frames.

Audio can also be practical complementary information [26–28] and descriptors like pitch, spectrogram, Zero-Crossing Rate (ZCR), Mel Frequency Cepstrum Coefficient (MFCC), etc., have been introduced for violence detection. For different modalities, early fusion methods or late fusion strategies are commonly used to achieve better performance [29–31]. Dong et al. [32] combine video frames, optical flows, and accelerated flow maps as multiple inputs to detect person-to-person violence. Although these well-established traditional algorithms are relatively fast and reliable in some specific cases, they rely too much on feature engineering with poor robustness and generalization.

2.2. Deep learning methods

In recent years, some researchers have tried deep neural networks for violence detection. Ersa et al. [33] propose a coarse-to-fine violence detection framework, where coarse-grained MFCC audio features are used for timing efficiency, while fine-grained advanced motion features are used when necessary. In [34], Zhou et al. build the FightNet based on a temporal segment network for detecting visual violence in complex scenes. [35] deploys a hybrid framework in which handcrafted and deep features are effectively combined to obtain better violence classification results. Later, Xu et al. [36] propose a dual-stream structure consisting of an object detector and FlowNet for both localization and recognition of fight actions. [37] exploits two network structures, C3D and CNN-LSTM, to explore high-level concepts and the objective form of violence in videos, respectively, and merged both to identify different violent subjects. Wu et al. [38] build an HL-Net with two branches to capture the long-range dependencies and local positional relations of the video, respectively. In [39], Su et al. propose a SPIL model for fight detection, which uses a graph convolutional network to capture physical action interactions. Similarly, Liu et al. [40] construct a skeleton-based monitoring system for violent action recognition and abuser tracking.

Recently, [41] proposes a Flow Gated Network that combines the advantages of 3D-CNN and optical flow for violence detection in surveillance footage. Fernando et al. [42] use a bidirectional ConvLSTM network followed by a multi-head self-attention block to detect violence. In [43], Islam et al. introduce a Separable Convolutional LSTM network for violence recognition, and a dual-stream structure with background suppression and frame difference is constructed to capture motion information. Asad et al. [44] explore a multi-level feature fusion strategy to merge diverse motion patterns, in which a Wide-Dense Residual Block (WDRB) is constructed to capture wide-range spatial features. Iqbal et al. [45] propose a weakly supervised Orientation Aware Object Detection (OAOD) algorithm for the automatic detection of firearms. However, most of these methods detect generalized violent events without exploring the semantic properties of violent elements, which is difficult for parsing complex violence scenarios.

In contrast to the above approaches, we propose a novel semantic multimodal violence detection model with local-to-global embedding. First, features of different modalities, i.e., appearance, motion, and audio (if available), are extracted by deep neural networks. Subsequently, a multimodal cross-fusion network (MCN) enables information sharing and complementarity by alleviating the heterogeneity gap across different modalities. Most importantly, we introduce a local-to-global embedding strategy, in which the generated local semantic detectors capture various sub-concepts of violence in the video while the global semantic descriptor reduces the intra-class variance of violence and enlarges the margin with non-violent samples.

3. The proposed method

In this section, we present the proposed violence detection framework in detail, which mainly consists of three parts: multimodal feature extraction, multimodal cross-fusion network, and local-to-global embedding, as shown in Fig. 2. First, we extract multimodal features using different deep neural networks. Subsequently, a multimodal cross-fusion network is constructed for effective multimodal feature fusion, followed by a multilayer perceptron (MLP) for violence detection. Finally, local semantic detection and global semantic alignment are jointly introduced in a multi-task learning manner for optimization.

3.1. Multimodal feature extraction

For the intelligent detection of violent videos, it is important to effectively mine multimodal information. In this part, we firstly utilize different neural network backbones to extract multimodal features, i.e., appearance, motion, and audio.

3.1.1. Appearance feature extraction

RGB frames often contain rich visual information, which is widely used for violence detection. Considering a large amount of redundancy between adjacent video frames, reasonable time-domain sampling is necessary. Dense sampling can capture the motion trajectory of frame sequences but has a limited span, while uniform sampling has an expansive sampling range but covers few violent frames. To compensate for the shortcomings mentioned above, we try to obtain a better video representation by increasing the number of sampled clips.

Assuming that the total number of frames of the original video V is T and the length of the sampled segments is l, the hybrid sampling process is defined as

$$v_j^i = V^{s+(i-1)\times\tau+k}, s \in [(j-1)\times\frac{T}{n}, j\times\frac{T}{n} - l\times\tau], \quad k \in [0,\tau], \tag{1}$$

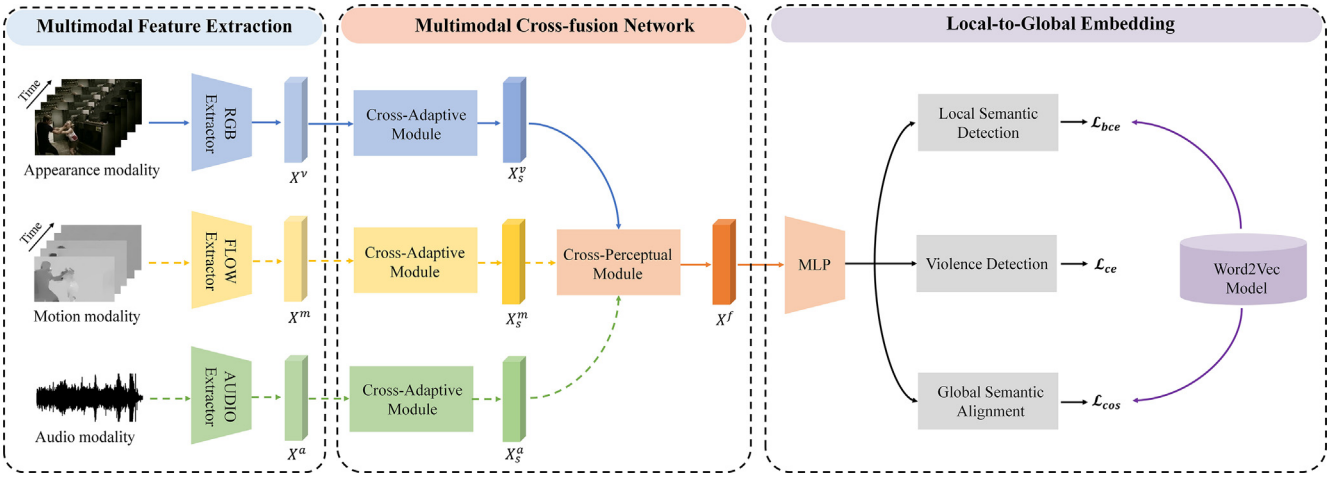

Fig. 2. An overview of the proposed violence detection framework. We take the appearance feature as the primary modality, while the motion or audio feature is treated as an auxiliary modality. The Multimodal Cross-fusion Network (MCN) aims to alleviate the inter-modal heterogeneity while generating the enhanced fusion feature X^f. The fusion feature is projected to a latent semantic space by a multilayer perceptron (MLP). Furthermore, we extract the corresponding sub-concept word embeddings from an external corpus (i.e., Word2Vec) for semantic parsing. The local semantic detection branch captures fine-grained violence elements by multi-label classification loss L_{bce}, while the global semantic alignment branch re-calibrates the high-level semantics of violent videos by cosine embedding loss L_{cos}. Finally, we apply the violence detection loss (i.e., L_{ce}) and semantic embedding losses (i.e., L_{bce} and L_{cos}) to jointly optimize the whole framework.

where v_j^i represents the i^{th} frame of the j^{th} sampling clip, τ is the sampling rate, and k is a random jitter in the time domain for frame selection. Specifically, the whole frame sequence interval $[0, T]$ is equally divided into n subintervals, which are randomly performed dense or uniform sampling operations. These n clips of length l are then fed into a pretrained X3D-L network [46] independently to obtain clip-level features $X = \{x_i\}_{i=1}^n$. Then, a scaled dot-product attention mechanism [47] is applied to aggregate the global contextual information, after which the aggregated clip features are averaged over the temporal dimension to generate video-level representation X^v.

3.1.2. Motion feature extraction

Optical flow is the instantaneous velocity of the pixel motion of a moving object on the viewing plane. It suppresses redundant spatial noise and focuses information on the moving target, making it more suitable for movement description. First, we use the TV-L1 algorithm [48] to extract dense optical flow, where the horizontal d_x and the corresponding vertical d_y of adjacent frames are stacked to form an optical flow graph $f = (d_x, d_y)$. Subsequently, it is fed into a pretrained ResNet50 [49] with a temporal shift module (TSM) [50] to extract the high-level motion feature X^m.

Different from the appearance feature extraction process, the optical flow sequence in the time domain is generated by

$$f^i = F^{s+k}, s \in [(i-1) \times \frac{T}{n}, i \times \frac{T}{n}], \quad k \in [0, \tau], \tag{2}$$

where f^i denotes the i^{th} optical flow graph in the sampled clip, and F^i is an optical flow sequence of length T. Since optical flow graphs are inherently low-level motion features, further dense sampling and multi-clip input may lead to overfitting. In addition, motion features are more dependent on the complete video sequence, and such a sampling strategy can cover more time-domain information.

3.1.3. Audio feature extraction

As for audio feature extraction, a PANNs model [51] pre-trained on AudioSet [52] is exploited. We first separate the audio from its corresponding video and convert it to a logarithmic mel-spectrogram. Then, the raw audio waveform and mel-spectrogram are fed into the PANNs network simultaneously to capture both time-domain and frequency-domain information. Concretely, we choose the output vector of CNN14 at the top of the PANNs as the audio features X^a. More preprocessing details can be found in the experimental setup in Section 4.

3.2. Multimodal cross-fusion network

Since the above features are extracted from encoders pre-trained on different datasets, they may contain some task-irrelevant redundancies that interfere with the fusion process. Meanwhile, these features exist in their respective representation spaces as different media forms. Subject to the heterogeneity gap, direct concatenation or addition may affect the initial distribution of features and even degrade the quality of the unimodal representation. Therefore, we develop a novel multimodal cross-fusion network (MCN) to obtain the enhanced fusion feature, as shown in Fig. 3.

The network mainly consists of two parts: a cross-adaptive module (top) and a cross-perceptual module (bottom left). In the cross-adaptive module, we first perform cross-modal adaptation through a gate mechanism, in which the generated attention weights are used to re-calibrate the initial feature distribution. Then, a group convolution layer followed by a ReLU function is applied to project the modality into a joint space to eliminate the heterogeneity gap.

Here we take the appearance feature X^v as the primary modality and the audio feature X^a as the auxiliary modality, which is fed into our proposed MCN network. As some videos lack audio information, the motion feature is introduced as the auxiliary. The process is formulated as

$$\begin{aligned} X_m^v &= ReLU(\phi(X^v + \sigma(X^a)X^v)), \\ X_m^a &= ReLU(\psi(X^a + \sigma(X^v)X^a)), \end{aligned} \tag{3}$$

where $\sigma(\cdot)$ is a sigmoid function, $\phi(\cdot)$ and $\psi(\cdot)$ are two group convolution layers. Subsequently, we conduct the multimodal interaction in the common subspace, in which a shared group convolution layer $\kappa(\cdot)$ is used to enhance the joint feature representation. On the one hand, the kernel of group convolution can effectively capture the local context over the channel dimension. On the other hand, channel grouping greatly reduces the number of parameters and pre-

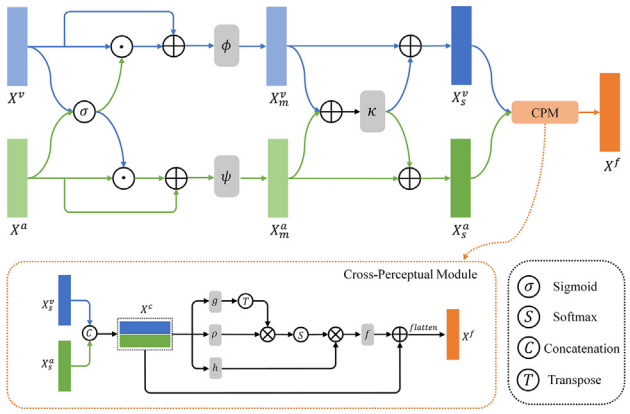

Fig. 3. The structure of the multimodal cross-fusion network.

vents overfitting. A residual connection is also performed to maintain the independence of each modality. The whole process can be expressed as

$$X_s^v = X_m^v + ReLU(\kappa(X_m^v + X_m^a)),$$
$$X_s^a = X_m^a + ReLU(\kappa(X_m^v + X_m^a)). \tag{4}$$

The outputs after the cross-adaptive module are concatenated as $X^c = (X_s^v, X_s^a)$, which are fed into the cross-perceptual module to filter task-irrelevant redundancies and activate the inter-modal consensus. Accordingly, the cross-perceptual module can be formulated as

$$X^h = softmax\left(\frac{\rho(X^c)g(X^c)^T}{\sqrt{d}}\right)h(X^c), \tag{5}$$

$$X^f = flatten(LN(X^c + f(X^h))), \tag{6}$$

where $\rho(\cdot), g(\cdot)$ and $h(\cdot)$ are three different linear layers, d is the hidden dimension and $LN(\cdot)$ is the layer normalization operation. In the cross-perceptual module, we use the auxiliary modality to capture local information across modalities and suppress task-irrelevant redundant noise in the primary modality.

3.3. Violence detection

After obtaining the fused feature, we use a multilayer perceptron (MLP) to generate video embeddings with high-level semantics, which are finally fed into a linear layer for classification. The process is expressed as

$$X^e = MLP(X^f), \tag{7}$$

$$y^s = softmax(WX^e + b), \tag{8}$$

where y^s denotes the prediction score, W is a linear projection function and b is a bias term. A binary cross-entropy loss function is used for supervised learning of violence detection, which is calculated as

$$L_{ce} = \frac{1}{N}\sum_{i=1}^{N} -y_i log(y_i^s), \tag{9}$$

where N is the total number of videos, $\{y_i\}_{i=1}^N$ and $\{y_i^s\}_{i=1}^N$ are the ground truth and prediction score of the videos, respectively.

3.4. Local-to-global embedding

Previous efforts mainly focus on the generalized detection of violent behavior, ignoring the explicit semantics of violent events and the co-occurrence trend of violent elements. In this section, we propose a novel semantic embedding strategy to guide violence detection, including local semantic detection and global semantic alignment. The former uses word embeddings of violence sub-concepts to generate a set of local semantic detectors for capturing specific concepts in video embeddings. The latter aims to align violence embeddings with a global semantic descriptor, eliminating the intra-class variance of violent classes while enlarging the semantic gap from nonviolence embeddings.

3.4.1. Local semantic detection

The concept of violence itself is vague and diverse, and human determinations of violence are more often derived from specific entities in some scenarios, such as blood, physical conflict, guns, explosions, etc. Inspired by this observation, we consider that parsing the abstract concept of violence into localized violent elements will facilitate the model to learn the essential representation of violence. Therefore, we first propose a local semantic detection method to generate local semantic detectors, as shown in Fig. 4. The combination of these detectors reveals the co-occurrence trend of violent elements and helps semantic relation reasoning in complex scenes.

Specifically, we extract word embeddings corresponding to violence categories from a large text corpus, i.e., Word2Vec [53], and subsequently construct the semantic relation graph based on conditional probabilities of label co-occurrence (case 1). In particular, when multi-label annotations are not available, we initialize the semantic relation graph via a co-similarity matrix (case 2). This process can be expressed as

$$G_{ij} = \begin{cases} A(e_i, e_j) & \text{if case 1,} \\ S(e_i, e_j) & \text{if case 2.} \end{cases} \tag{10}$$

$$A(e_i, e_j) = \Gamma\left(\frac{1}{2}(P(e_j|e_i) + P(e_i|e_j))\right), \tag{11}$$

$$S(e_i, e_j) = \frac{e_i e_j^T}{\|e_i\|\|e_j\|}, \tag{12}$$

where $\{e_i\}_{i=1}^C$ is the original label embeddings, $P(e_j|e_i)$ denotes the probability of occurrence of e_j under the condition that e_i occurs, and vice versa. $\Gamma(\cdot)$ is a threshold function to smooth the long-tailed distribution. By constructing the semantic relation graph G, each independent category is associated with a specific prior. Generally, a larger G_{ij} indicates that the i^{th} category is more closely related to the j^{th} category, and this relationship is reflected in the co-occurrence trend or feature similarity of violent elements.

After obtaining the semantic relation graph, we use a graph convolutional network (GCN) to generate the local semantic detectors. Specifically, the initial word embeddings are projected to a latent semantic space by a linear function, and the semantic vectors are subsequently updated by the semantic relation graph. The update process is formulated as

$$\tilde{G}_{ij} = \frac{exp(G_{ij})}{\sum_{k=1}^{C} exp(G_{ik})}, \tag{13}$$

$$E_{l+1} = F^{\Theta}(\tilde{G} E_l W_l), \tag{14}$$

where $\tilde{G} \in \mathbb{R}^{C\times C}$ is a normalized semantic graph, $W_l \in \mathbb{R}^{D_e \times D_l}$ is a learned transformation matrix, $E_l \in \mathbb{R}^{C\times D_e}$ is the hidden representa-

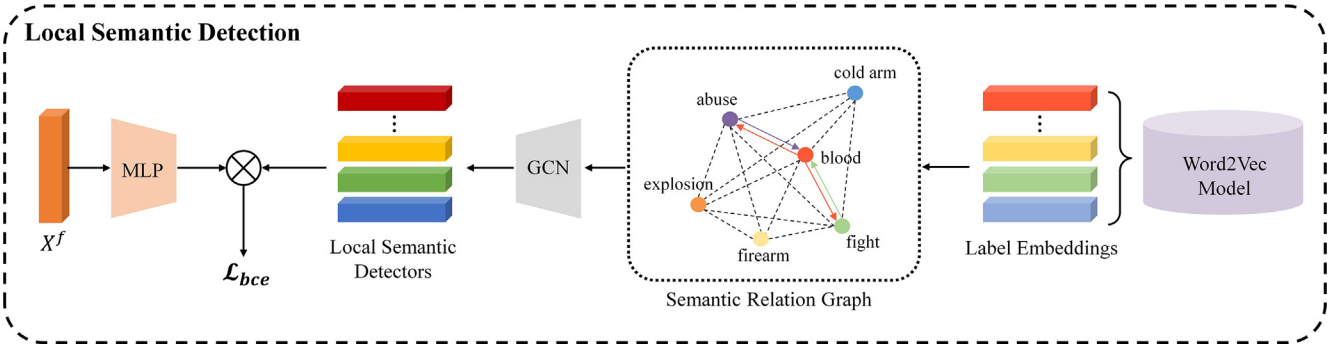

Fig. 4. The overall structure of local semantic detection. We extract word embeddings for a series of violent sub-concepts from a pretrained Word2Vec model. A semantic relation graph is then constructed based on the label co-occurrence probabilities or co-similarity matrix. The graph convolutional network (GCN) is used to update the word embeddings to generate local semantic detectors, which are applied to the video embedding to obtain local semantic responses.

tion, and $F^{\Theta}(\cdot)$represents a LeakyReLU unit with a dropout layer. This operation ensures dynamic knowledge propagation between different categories, resulting in enhanced word representations, which we call the local semantic detectors.

Guided by local semantics, the classifier will learn to project video embeddings from the visual space to the semantic space, responding to a specific semantics of violence. By applying the detectors to video representations, we can obtain local violence prediction scores by

$$y^p = \sigma(EX^e + b'), \tag{15}$$

where $E \in \mathbb{R}^{C\times D}$ is the generated detector and $X^e \in \mathbb{R}^{1\times D}$ is the high-level video embedding.

In the semantic space, the detector can guide the response of video embeddings with specific local semantics, enabling the figurative parsing of violence concepts. Accordingly, a binary cross entropy loss function is defined as

$$L_{bce} = \frac{1}{N}\sum_{n=1}^{N} y^c log(y^p) + (1-y^c) log(1-log(y^p)), \tag{16}$$

where $y^c \in \mathbb{R}^{1\times C}$ is the multi-label annotation. This loss function describes the probability distribution of multiple non-mutually exclusive objects appearing in the same video, which is suitable for local semantic detection. For datasets lacking multi-label annotations, the local semantic detection loss correspondingly degrades to a regular binary cross-entropy.

3.4.2. Global semantic alignment

Local semantic detection ensures that different types of violent elements are distinguished, however, it does not precisely describe the high-level semantics of violence, i.e., the abstract notion of the combination of multiple elements. Therefore, we further propose a global semantic alignment strategy to unify the abstract concept of element combinations. The constructed global semantic descriptor conduce to the clustering of violent categories in the semantic space while maintaining a certain distance from non-violent samples, as shown in Fig. 5.

To obtain a uniform semantic representation, we construct the global semantic descriptor using local violence detectors generated in the joint space. Specifically, we take the mean value of the local violence detectors as the semantic center of violence, which is formulated as

$$P = \frac{1}{C}\sum_{i=1}^{C} E_i, \tag{17}$$

where $E_i \in \mathbb{R}^{1\times D}$ is a specific local semantic detector, and C is the number of violence categories. Since these detectors have similar

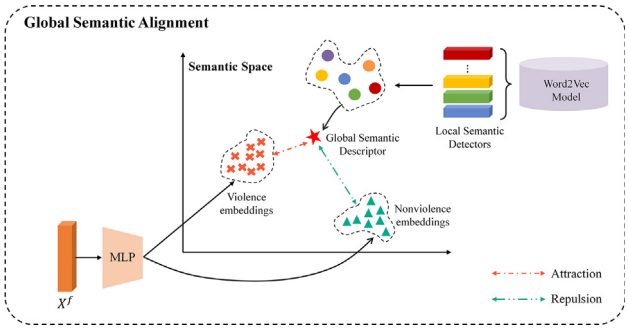

Fig. 5. The overview of global semantic alignment. The global semantic descriptor is constructed to attract violent embeddings while repulsing non-violent embeddings.

distribution properties and are dynamically updated, their average value in the semantic space can be used as a semantic prototype of violence, which is adaptive to datasets from different sources.

Subsequently, we expect different violent video embeddings to have a certain clustered tendency in the joint semantic space, maintaining semantic consistency with the global semantic descriptor. Therefore, we introduce a cosine embedding loss to narrow the gap between the violent samples and the descriptor, which is defined as

$$L_{cos} = \begin{cases} 1-\cos(X^e, P) & \text{if } \mu = 1, \\ max(0, \cos(X^e, P)) & \text{if } \mu = -1. \end{cases} \tag{18}$$

where X^e is the abstract video embedding in the semantic space. $\mu = 1$ or -1 denotes whether the video embedding has the same semantics as the descriptor P, i.e, when X_e is a violent video, $\mu = 1$, otherwise $\mu = -1$.

Global semantic alignment promotes a compact distribution of violent samples in the semantic space, alleviating the large intra-class variation caused by the diversity of violent elements. Meanwhile, the descriptor maintains a certain margin with the non-violent class, clarifying the decision boundary between the two within the semantic space. By synergizing with local semantic detection, harmony in diversity of violence categories is further achieved.

3.5. Objective function

Finally, we merge the local-to-global embedding losses (i.e., L_{bce} and L_{cos}) with the violence detection loss (i.e., L_{ce}) to form the final objective function for the optimization of the whole framework, which is expressed as

$$L = L_{ce} + \lambda_1 L_{bce} + \lambda_2 L_{cos} \tag{19}$$

where λ_1 and λ_2 are two weight hyperparameters. By optimizing this final objective function, our framework can yield discriminative video representations to obtain more accurate prediction results. The detailed description of our model optimization procedure is displayed in Algorithm 1.1.

Algorithm 1: Semantic Multimodal Violence Detection

input : multimodal features $\{X_i^v, X_i^a\}_{i=1}^N$ or $\{X_i^v, X_i^m\}_{i=1}^N$, multi-label annotations $\{y_i, y_i^c, \mu_i\}_{i=1}^N$, violence word embeddings $\{e_j\}_{j=1}^C$.
output: violence prediction scores $\{y_i^s\}_{i=1}^N$

initial the network parameters of MCN, MLP and GCN;
for $i \leftarrow 1$ **to** *mini-batch M* **do**
  Generate fusion features X_i^f from MCN;
  Generate high-level semantics X_i^e from MLP;
  Calculate the violence detection loss L_{ce} according to Eq. (9);
  Constract the semantic relation graph G according to Eq. (10);
  Generate the local semantic detectors $\{E_j\}_{j=1}^C$ from GCN;
  Generate the global semantic descriptor P according to Eq. (17);
  Calculate the local detection loss L_{bce} according to Eq. (16);
  Calculate the global alignment loss L_{cos} according to Eq. (18);
  Update the network parameters to minimize L in Eq. (19);
end
return *The optimal network parameters*

4. Experiments and validation

To verify the effectiveness of the proposed method, we have conducted experiments on five different types of violence datasets: Crowd Violence (Violent Flow) [8], Hockey Fight [54], RLVS [55], RWF-2000 [41] and VSD2015 [56]. A brief summary of the datasets is presented in Table 1.

In this work, we mainly focus on the VSD2015 dataset, as it is not only the largest violence dataset with audio information but also suffers from severe a category imbalance problem, which poses a significant challenge for violence detection. Here, we first introduce the datasets and their corresponding evaluation metrics. Next, we provide the implementation details and hyperparameter settings. A series of quantitative and qualitative experiments are finally presented to demonstrate the effectiveness of the framework in this paper.

4.1. Datasets

1) Crowd Violence contains 246 videos downloaded from YouTube, with a 50/50 split between violent and non-violent videos. The clip length ranges from 1.04 to 6.52 s. The dataset focuses on violence in crowded scenes with low image resolution.

2) Hockey Fight has a total of 1,000 video clips that are filmed in hockey games of the National Hockey League. The average length of these videos is less than 2 s, most of which are physical conflicts that occurred during the game. The video was shot from a single scene and in a simple setting.

3) Real-Life Violence Situations (RLVS) contains 2,000 video clips with an average length of 5 s, with 1,000 violent and 1,000 non-violent videos each. The dataset is collected from streets, schools, and prisons to ensure diversity, and the rest are from YouTube. Some of the videos contain audio files.

4) RWF-2000 is collected from surveillance videos in real scenes without audio information. It has a total of 2000 clips, of which 1600 videos are the training set and 400 videos are used for testing. The number of violent and non-violent video clips is well balanced and the average duration is about 5 s. Violent contents include fights, robberies, explosions, assaults,.etc.

5) Violent Scene Detection (VSD2015) is a large audio-visual dataset released by the Mediaeval 2015 VSD competition, which contains 10,900 clips from Hollywood movies and YouTube videos ranging from 8 s to 12 s. In particular, we divide this dataset into six subclasses, including abuse, blood, cold arms, explosion, fighting, and firearms. Considering that multiple violence categories may exist in the same video, we provide multi-label annotations for the training set. In addition, there is a severe data imbalance with only 502 videos being violent and the rest being non-violent. Therefore, we use random mirroring, rotation, and cropping for data augmentation in the training phase.

Table 1
Comparison of different violent video datasets.

Dataset	Size	Duration	Scenario	Audio
Crowd Violence [8]	246 Clips	1~6s	Street Crowd	✓
Hockey Fight [54]	1,000 Clips	1~2s	Ice Hockey	×
RLVS [55]	2,000 Clips	3~7s	Movies and Sports	✓*
RWF-2000 [41]	2,000 Clips	5s	Surveillance	×
VSD2015 [56]	10,900 Clips	8~12s	Movies	✓

* Partially available.

4.2. Evaluation metrics

Since the number of violent and non-violent videos in the VSD2015 dataset is imbalanced, we use the officially specified average precision (AP) for evaluation, which is more sensitive to less numerous categories (e.g., violence). The metric is calculated as

$$AP = \frac{1}{P}\sum_{i=1}^{N} L_i \cdot \frac{P_i}{i}, \tag{20}$$

where N is the total number of test sets, P is the number of testing violent videos, and $L_i = 1$ or 0 indicates whether the i^{th} video is violent or not. The prediction scores of the testing videos are arranged in descending order, and P_i videos are predicted correctly in the first i predicted samples.

For the remaining four datasets, with an equal number of violent and non-violent classes, we use the testing accuracy to evaluate the model performance, which is formulated as

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%, \tag{21}$$

where TP and TN are the numbers of correct hits, while FP and FN are the numbers of false alarms. Specifically, Crowd, Hockey, and RLVS are not split to train/test partitions, so we use fivefold cross-validation for comparison.

4.3. Implementation details

For appearance feature extraction, the number of video subintervals n is adopted as 10, the length l of the clip is set to 16 and the sampling rate τ is set to 5. After size transformation and central cropping, we get an input sequence with a spatial–temporal resolution of $16 \times 3 \times 312 \times 312$. These sequences are sent into a pretrained X3D-L model to obtain 2048-dimensional appearance features.

The dense optical flow is uniformly sampled in the time domain to obtain a stacked sequence of $l \times 2 \times H \times W$, where 2 denotes the horizontal and vertical directions, respectively. The scaled transformed sequence is randomly cropped to a spatial size of 224×224. Finally, the sequence is fed into a ResNet50 network with a TSM block to obtain the 2048-dimensional motion feature.

For audio feature extraction, we first separate the audio file from the video using the moviepy tool. Subsequently, the audio signal is filtered by a Hamming window with a window size of 1024 and a jump distance of 320. After performing a short-time Fourier transform, a logarithmic mel-spectrogram with a size of 100×64 is obtained. Finally, the original audio and spectrogram are fed into the PANNs network in parallel to generate 2048-dimensional audio features.

In the MCN network, the output dimension of the group convolution layers in the cross-adaptive module is 2048 with 512 groups, while the linear layers of the cross-perceptual module contain 2048 nodes with an output dimension of 128. For the MLP, a group convolution layer with 1024 units and two linear layers with 300 and 2 units, respectively, are included. The two graph convolution layers in GCN contain 64 and 300 nodes, with a dropout rate of 0.2.

As for hyperparameters, we empirically set $\lambda_1 = 2$ and $\lambda_2 = 3$ in the final objective function. The SGD optimizer with momentum is adopted for optimization, of which the weight decay rate is 1×10^{-5} and the momentum term is 0.9. The learning rate is set to 1×10^{-3} with a cosine decay strategy. We train our model for 50 epochs in total with a mini-batch size of 128. All experiments are conducted on a GTX 1080Ti GPU based on PyTorch.

4.4. Comparison with state-of-the-art methods

We first compare our proposed violence detection framework with the current state-of-the-art methods. Due to the different evaluation metrics, we report the AP performance on VSD2015 in Table 2, while Table 3 shows the test accuracies for the remaining four datasets. As shown in Table 2, the AP value of our model on the VSD2015 dataset is significantly better than those of existing methods, exceeding the latest SOTA results by 6.08%. Compared to those early fusion methods (e.g., Gu et al.[17] and Peixoto et al.[57]) and late fusion methods (e.g., Zheng et al.[58]), our method obtains considerable improvements. This result suggests that a rational fusion approach can effectively reduce the information redundancy across modalities and enhance multimodal synergy, while the local-to-global embedding strategy contributes to discriminative representations for more reliable results.

The results reported in Table 3 further verify the generalizability of the proposed method. We can find that our approach achieves competitive results overall, with 1.9%, 1.25%, and 0.51% improvements on RLVS, RWF-2000 (abbreviated as RWF), and Crowd, respectively. However, we notice that the improvement on these datasets is not as remarkable as on the VSD2015 dataset. On the one hand, most of these datasets are shot by fixed-view devices without scene switching, resulting in a single form of violence. On the other hand, the image resolution is relatively low due to hardware constraints. In contrast, the VSD2015 dataset originates from film footage, with artistic expressions such as depth of field changes and camera movements, thus producing complex and diverse scenes of violence. This also confirms the advantages of the proposed method for semantic parsing in complex scenarios.

Table 2
Comparisons with state-of-the art methods on the VSD2015 dataset.

Method	AP
MIC-TJU [59]	0.2848
FSP [33]	0.2947
Fudan–Huawei [12]	0.2959
Constantin et al. [60]	0.2968
Peixoto et al. [57]	0.301
Li et al. [61]	0.303
Zheng et al. [58]	0.3242
Gu et al. [17]	0.4131
Ours	**0.4739**

Table 3
Comparisons with other state-of-the art methods on different datasets. The best results are highlighted in bold and the second best results are underlined.

Method	Accuracy(%)			
	Crowd	Hockey	RLVS	RWF
VGG16 + LSTM [55]	90.01	95.1	-	-
ConvLSTM [13]	94.57	97.1	-	-
Efficient Conv3D [16]	97.17	98.3	-	-
VGG16 + WDRB + LSTM [44]	97.1	98.8	-	-
P3D + LSTM [17]	<u>97.69</u>	94	-	-
Inception-Resnet-V2 [62]	93.33	-	86.79	-
Conv2D + LSTM [63]	-	94.5	92	-
Flow Gated Network [41]	88.87	98	-	87.25
SPIL [39]	94.5	96.8	-	89.3
SepConvLSTM [43]	-	**99.5**	-	<u>89.75</u>
DeVTr [64]	-	-	<u>96.25</u>	-
ViolenceNet [42]	96.9	<u>99.2</u>	95.6	-
Ours	**98.22**	<u>99.2</u>	**98.15**	**91**

In addition, the annotations of the other datasets are coarser than those of the VSD2015 dataset, and the following ablation studies demonstrate the prominent contribution of fine-grained annotation for semantic embedding.

4.5. Ablation study

Here, we evaluate the effectiveness of the proposed method in both quantitative and qualitative terms. We first perform ablation studies to investigate the contribution of each modality. Subsequently, we discuss different multimodal fusion approaches to verify the capacity of MCN. Finally, we analyze the effect of each component in the local-to-global embedding.

4.5.1. The effect of unimodality

We first report the effect of each modality on different datasets in Table 4. We can find that the appearance modality has a clear advantage over the motion modality in most cases, thanks to the rich information on visual violence. For the Hockey dataset, presenting a single physical conflict in sports, the effect of the motion modality is slightly better than the appearance. This also indicates that rapidly changing dynamic violence is more easily perceived compared to static violence. In addition, we note that the audio performance of the Crowd dataset is far from that of the VSD2015 dataset. Because the former is captured from real scenes of crowd violence with noisy background sounds, while the latter is derived from movie footage with clear human and ambient sounds. Surprisingly, the effect of audio modality in the VSD2015 dataset even surpasses the appearance modality. We deem that the corresponding auditory violence, such as screams, explosions, gunshots, etc., is more intuitive than the complex and variable visual information, making it easier to distinguish audio in violent and non-violent scenes.

Table 4
The Effect of unimodality on different datasets (accuracy in percentage except for VSD).

Modality	Dataset				
	Crowd	Hockey	RLVS	RWF	VSD2015
Appearance	96.49	97.2	95.3	88.25	0.3041
Motion	93.04	97.4	94.8	82.5	0.2205
Audio	77.87	-	-	-	0.356

Considering the model size and inference efficiency, we use the two more effective modalities for the subsequent multimodal feature fusion. In particular, most of the audio in the Crowd dataset is background noise without significant semantic information, we use motion modality to capture violence such as assault and fighting instead. As for the evaluated dataset lacking audio information (i.e., Hockey, RLVS, and RWF), the motion feature is adopted as an auxiliary modality.

4.5.2. Comparison with different fusion methods

To illustrate the advantages of the proposed multimodal cross-fusion network, we compare the effect of different fusion strategies on five datasets, including both early fusion and late fusion methods. As Table 5 shows, the MCN structure achieves an impressive improvement on all datasets. Compared with direct concatenation, the performance of the MCN-based framework on the RWF, RLVS, and VSD dataset is improved by 1%, 0.95%, and 0.92%, respectively. This benefits from a rational two-stage fusion strategy, in which the cross-adaptive module eliminates the inter-modal heterogeneity, while the cross-perceptual module further filters task-irrelevant redundancies to obtain the enhanced fusion feature.

In addition, compared with a single modality, multimodal fusion can effectively boost the model performance, which is in line with the expectation of complementarity across different modalities. In some cases, however, feature fusion methods perform worse than the fusion of decision scores, which may be attributed to the simpler nature of these datasets themselves and the tendency to cause overfitting using deep fusion models.

4.5.3. The effect of local-to-global embedding

Finally, we conduct ablation studies to explore the effect of local-to-global embedding, as shown in Table 6. After introducing semantic embedding objective functions (i.e., L_{bce} and L_{cos}), the test AP of our model on the VSD2015 dataset is improved by 1.07% and 0.75%, respectively. On the one hand, the local semantic detectors generated by the semantic correlation graph can dynamically parse complex audiovisual violence scenes. At the same time, the global semantic descriptor drives the violent samples to form a compact clustering while keeping a distance from the non-violent embeddings. The inconsistency within the violent classes is further mitigated.

Remarkably, without using additional multi-label annotations, the local semantic detectors and the global descriptor both generated by co-similarity reasoning still perform well, improving 0.44% and 0.68% over the baseline result, respectively. When we jointly introduce local and global branches, the best AP value on the VSD2015 test dataset is obtained. This result shows that the two have a certain synergistic effect, which further verifies the effectiveness of our contribution.

In addition, we report the effect of local-to-global embedding on the other four datasets in Table 7. Notably, although not as significant as the boost on the VSD2015 dataset, the embeddings updated by the co-similarity graph also achieve good results on these datasets. Moreover, the contribution of global semantic alignment is generally lower than that of local semantic detection due to the smaller intra-class variance and the greater discriminability of these datasets. In contrast, the local semantic detectors cover a wide range of violence and are more sensitive to some violence variations. The above results also illustrate the robustness and generalizability of our method.

Table 5
Comparison with different fusion methods (accuracy in percentage except for VSD).

Fusion Level	Method	Dataset				
		Crowd	Hockey	RLVS	RWF	VSD2015
Late Fusion	WA[1]	96.93	97.8	96.1	88.5	0.4304
	LR[2]	97.03	97.6	95.8	88.75	0.4311
Early Fusion	Concat	96.93	97.8	96.45	89.5	0.4435
	DMRN [65]	97.16	97.9	96.35	90	0.4461
	MCN (Ours)	**97.38**	**98.3**	**97.4**	**90.5**	**0.4527**

[1] Weighted Average.
[2] Linear Regression.

Table 6
Ablation Studies of local-to-global embedding on the VSD2015 dataset.

L_{ce}	L_{bce}	L_{cos}	AP	
			Labeled	Unlabeled
✓			0.4527	0.4527
✓	✓		0.4634	0.4571
✓		✓	0.4602	0.4595
✓	✓	✓	**0.4739**	**0.4629**

Table 7
Ablation Studies of local-to-global embedding on different datasets.

L_{ce}	L_{bce}	L_{cos}	Accuracy (%)			
			Crowd	Hockey	RLVS	RWF
✓			97.38	98.3	97.4	90.5
✓	✓		97.64	**99.2**	98.05	91
✓		✓	97.57	99	97.5	90.75
✓	✓	✓	**98.22**	99	**98.15**	**91**

4.6. Evaluation of efficiency

In this subsection, we report the time and space complexity of our method, as shown in Table 8. Here, we do not take the feature extraction stage into account since it is not the main contribution of this paper. Our model consists of a Multimodal Cross-fusion Network (MCN), a Multilayer Perceptron (MLP), and a Graph Convolutional Network (GCN), corresponding to multimodal fusion, violence detection, and local-to-global embedding, respectively. Therefore, we mainly analyze the time–space complexity in these modules, which include fully connected (FC) layers, group convolution and graph convolution layers.

For each FC layer, the time complexity can be expressed as

$$Time \sim O(L \times K \times D_{in} \times D_{out}), \tag{22}$$

where L is the length of the input feature sequence and K is the kernel size. For the FC layer, $K = 1$. D_{in} and D_{out} are the input and output feature dimensions, respectively. In terms of the space complexity, we consider the learned parameters at each layer, denoted as

$$Space \sim O(D_{in} \times D_{out}). \tag{23}$$

The time complexity of the group convolution layer is formulated as

$$Time \sim O(L \times K \times D_{in} \times \frac{D_{out}}{N_{group}}), \tag{24}$$

where N_{group} indicates the number of grouped channels. For $N_{group} = 1$, this operation is equivalent to the FC layer. Similarly, its space complexity can be expressed as

$$Space \sim O(D_{in} \times \frac{D_{out}}{N_{group}}). \tag{25}$$

With respect to GCN layers, we take the time complexity of matrix multiplication into account. Given two matrices with sizes $M \times N$ and $N \times Q$, respectively, the time complexity is denoted as

$$Time \sim O(M \times N \times Q). \tag{26}$$

Table 8
Analysis of time–space complexity of our method.

Module	Layer	Layer num	Time	Space
MCN	Group Conv	3	$L\times1\times2048\times\frac{2048}{512}$	$2048\times\frac{2048}{512}$
	FC	3	$2L\times1\times2048\times128$	2048×128
	FC	1	$2L\times1\times128\times2048$	128×2048
MLP	Group Conv	1	$L\times1\times4096\times\frac{1024}{8}$	$4096\times\frac{1024}{8}$
	FC	1	$L\times1\times1024\times300$	1024×300
	FC	1	$L\times1\times300\times2$	300×2
GCN			$L_c\times L_c\times64$	$L_c\times L_c$
	GCN	2	$L_c\times300\times64$	300×64
			$L_c\times L_c\times300$	$L_c\times L_c$

L_c indicates the number of graph nodes.

Table 9
Comparison of efficiency with existing methods.

Model	#Params	FLOPs
ConvLSTM [13]	9.6 M	14.40G
Efficient Conv3D [16]	7.4 M	10.43G
VGG16 + LSTM [55]	2.06 M	24.13 M
Flow Gated Network [41]	0.27 M	0.54 M
SepConvLSTM [43]	0.33 M	1.93 M
ViolenceNet [42]	4.5 M	-
Ours	1.98 M	6.35 M

Regarding space complexity, we mainly consider the size of the adjacency matrix and the learned weights, which can be formulated as

$$Space \sim O(M \times N). \tag{27}$$

In addition, We compare the efficiency of our framework with several SOTA methods, in terms of model size and algorithm complexity. As shown in Table 9, our model achieves a good balance between the number of parameters and floating-point operations (FLOPs). Compared to the algorithms proposed in [13]16, our model has a lower number of trainable parameters and requires significantly fewer FLOPs, ensuring faster computation and inference efficiency. In particular, our framework outperforms ViolenceNet [42] by 1.32% and 2.55% on the Crowd and RLVS datasets with fewer parameters, respectively, which is attributed to efficient group convolution layers in MCN. Moreover, our model surpasses Flow Gated Network [41] and SepConvLSTM [43] by 3.75% and 1.25% on the RWF-2000 dataset, achieving significant performance gains with a few additional parameters.

4.7. Qualitative analysis

To highlight the contribution of this paper, we visualize the results after local-to-global embedding. Fig. 6 shows several violent videos in the VSD2015 dataset after local semantic detection. We can see that these video images have a variety of styles with

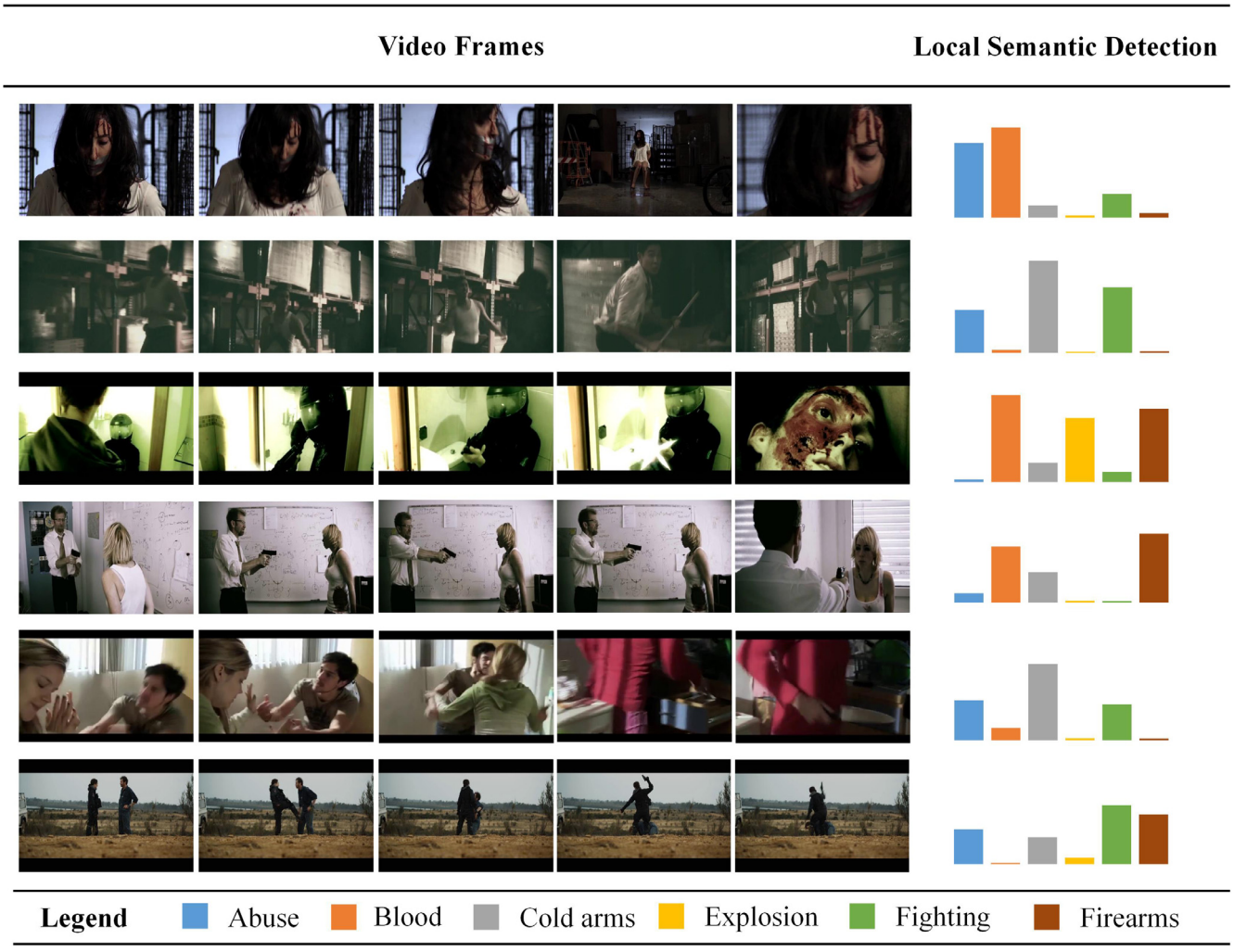

Fig. 6. Local Semantic Detection results of the proposed model on the VSD2015 dataset. The histogram represents the degree of response of these videos to different violent sub-concepts.

significant intra-class variances, and thus dynamically respond to multiple local semantics. We also notice that visual violence such as abuse, blood, and fighting show certain co-occurrence trends, while auditory violence such as explosions and gunshots have a strong semantic correlation. The local semantic detectors parse complex scenes into fine-grained entities by detecting sub-concepts of violence simultaneously, providing an interpretable basis for the judgment of violence.

In addition, we compare the video embedding distributions before and after global semantic alignment using t-SNE in Fig. 7. We can find that violent and non-violent classes are chaotically distributed in the common semantic space before alignment, reflecting a large intra-class variance of the VSD2015 dataset. After global semantic alignment, both forms compact clusters in the semantic space, respectively. This is because the global semantic descriptor guides violent samples closer to it while enlarging the interclass gap with nonviolent samples. The visualization results further demonstrate the validity and necessity of the global semantic alignment.

Finally, we present the qualitative results of semantic embedding on the RWF-2000 dataset in Fig. 8. We notice that most cases of incorrect prediction are caused by the low resolution of the surveillance videos. Also, some fixed panoramas with a larger field of view make it challenging to discern action details, where normal

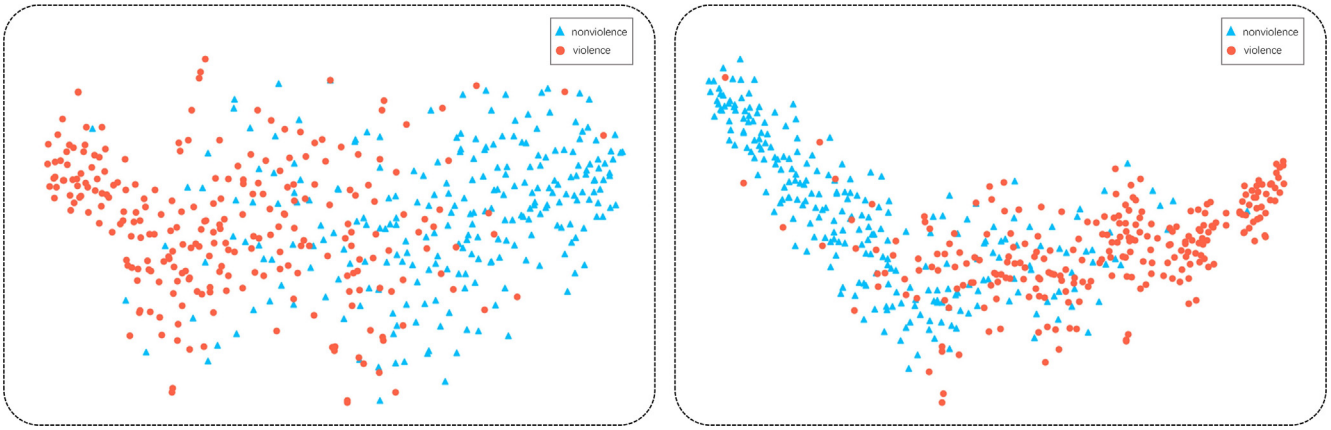

Fig. 7. Distributions of discriminative features using t-SNE on the VSD2015 dataset. The left/right panels show the results before and after global semantic alignment, respectively.

Video Frames	Ground Truth	Original Prediction	Semantic Embedding
	Non-violent	Violent	Non-violent
	Non-violent	Violent	Non-violent
	Violent	Non-violent	Violent
	Violent	Non-violent	Violent
	Violent	Non-violent	Violent
	Violent	Non-violent	Violent

Fig. 8. Several videos in the RWF-2000 dataset that are corrected after semantic embedding. Red font indicates the category of violence, while black indicates nonviolence.

body contact, such as handshakes, hugs, etc., are treated as conflicts. And in the nighttime environment, dim images can also be a cause of the error. With local-to-global embedding, these misjudged samples located at the decision boundary are further corrected, which also demonstrates that our method is feasible for violence detection in multiple views and scenarios.

5. Conclusion

In this paper, we propose a novel semantic multimodal violence detection framework with local-to-global embedding. For multimodal features, we construct a multimodal cross-fusion network (MCN) to eliminate inter-modal heterogeneity and achieve cross-modal enhancement. In particular, we parse generalized violence into a series of sub-concepts to capture the essence of violence. The corresponding word embeddings of violence sub-concepts are introduced to generate local semantic detectors and the global semantic descriptor. The former dynamically captures the specific violence semantics in video embeddings, while the latter prompts the violent classes to form a compact cluster thus reducing the intra-class variance. The two embeddings work together in a multi-task learning fashion to guide optimization. We demonstrate the effectiveness of the proposed method through a series of qualitative and quantitative experiments. In the future, we will introduce different external knowledge to adapt to more complex scenarios. Also, we will extend our work to address violence localization tasks.

CRediT authorship contribution statement

Yujiang Pu: Methodology. **Xiaoyu Wu:** Writing – review & editing, Funding acquisition. **Shengjin Wang:** Supervision. **Yuming Huang:** Writing – original draft. **Zihao Liu:** Data collection and preprocessing. **Chaonan Gu:** Data curation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the state key development program in 14th Five-Year under Grant No. 2021YFF0900701, 2021YFF0602103, 2021YFF0602102, 2021QY1702, and in part by Natural Science Foundation of China (No.61801441). We also thank the research funds under Grant No. 2019GQG0001 from the Institute for Guo Qiang, Tsinghua University, and the High-quality and Cutting-edge Disciplines Construction Project for Universities in Beijing (Internet Information, Communication University of China).

References

[1] D. Schwartz, L.J. Proctor, Community violence exposure and children's social adjustment in the school peer group: the mediating roles of emotion regulation and social cognition, J. Consulting Clin. Psychol. 68 (4) (2000) 670.
[2] D. Finkelhor, H. Turner, R. Ormrod, S.L. Hamby, Violence, abuse, and crime exposure in a national sample of children and youth, Pediatrics 124 (5) (2009) 1411–1423.
[3] S. Zhen, H. Xie, W. Zhang, S. Wang, D. Li, Exposure to violent computer games and chinese adolescents' physical aggression: The role of beliefs about aggression, hostile expectations, and empathy, Comput. Hum. Behav. 27 (5) (2011) 1675–1687.
[4] F. Butcher, J.D. Galanek, J.M. Kretschmar, D.J. Flannery, The impact of neighborhood disorganization on neighborhood exposure to violence, trauma symptoms, and social relationships among at-risk youth, Soc. Sci. Med. 146 (2015) 300–306.
[5] C.A. Anderson, A. Shibuya, N. Ihori, E.L. Swing, B.J. Bushman, A. Sakamoto, H.R. Rothstein, M. Saleem, Violent video game effects on aggression, empathy, and prosocial behavior in eastern and western countries: a meta-analytic review, Psychol. Bull. 136 (2) (2010) 151.
[6] J.B. Funk, H.B. Baldacci, T. Pasold, J. Baumgardner, Violence exposure in real-life, video games, television, movies, and the internet: is there desensitization?, J Adolescence 27 (1) (2004) 23–39.
[7] C. Clarin, J. Dionisio, M. Echavez, P. Naval, Dove: Detection of movie violence using motion intensity analysis on skin and blood, PCSC 6 (2005) 150–156.
[8] T. Hassner, Y. Itcher, O. Kliper-Gross, Violent flows: Real-time detection of violent crowd behavior, in: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2012, pp. 1–6.
[9] W. Zajdel, J.D. Krijnders, T. Andringa, D.M. Gavrila, Cassandra: audio-video sensor fusion for aggression detection, 2007 IEEE conference on advanced video and signal based surveillance, IEEE (2007) 200–205.
[10] J. Lin, W. Wang, Weakly-supervised violence detection in movies with audio and video based co-training, Pacific-Rim Conference on Multimedia, Springer (2009) 930–935.
[11] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[12] Q. Dai, R.-W. Zhao, Z. Wu, X. Wang, Z. Gu, W. Wu, Y.-G. Jiang, Fudan-huawei at mediaeval 2015: Detecting violent scenes and affective impact in movies with deep learning, MediaEval (2015).
[13] S. Sudhakaran, O. Lanz, Learning to detect violent videos using convolutional long short-term memory, in: 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), IEEE, 2017, pp. 1–6.
[14] A. Hanson, K. Pnvr, S. Krishnagopal, L. Davis, Bidirectional convolutional lstm for the detection of violence in videos, in: Proceedings of the European Conference on Computer Vision (ECCV), Workshops, 2018.
[15] W. Song, D. Zhang, X. Zhao, J. Yu, R. Zheng, A. Wang, A novel violent video detection scheme based on modified 3d convolutional neural networks, IEEE Access 7 (2019) 39172–39179.
[16] J. Li, X. Jiang, T. Sun, K. Xu, Efficient violence detection using 3d convolutional neural networks, in: 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), IEEE, 2019, pp. 1–8.
[17] C. Gu, X. Wu, S. Wang, Violent video detection based on semantic correspondence, IEEE Access 8 (2020) 85958–85967.
[18] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[19] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), vol. 1, Ieee, 2005, pp. 886–893.
[20] I. Laptev, On space-time interest points, Int. J. Comput. Vis. 64 (2) (2005) 107–123.
[21] H. Wang, C. Schmid, Action recognition with improved trajectories, in: Proceedings of the IEEE international conference on computer vision, 2013, pp. 3551–3558.
[22] R. Chaudhry, A. Ravichandran, G. Hager, R. Vidal, Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 1932–1939.
[23] Y. Gao, H. Liu, X. Sun, C. Wang, Y. Liu, Violence detection using oriented violent flows, Image Vis. Comput. 48 (2016) 37–41.
[24] T. Zhang, W. Jia, X. He, J. Yang, Discriminative dictionary learning with motion weber local descriptor for violence detection, IEEE Trans. Circuits Syst. Video Technol. 27 (3) (2016) 696–709.
[25] J. Mahmoodi, A. Salajeghe, A classification method based on optical flow for violence detection, Expert Syst. Appl. 127 (2019) 121–127.
[26] T. Giannakopoulos, A. Pikrakis, S. Theodoridis, A multimodal approach to violence detection in video sharing sites, in: 2010 20th International Conference on Pattern Recognition, IEEE, 2010, pp. 3244–3247.
[27] T. Giannakopoulos, A. Makris, D. Kosmopoulos, S. Perantonis, S. Theodoridis, Audio-visual fusion for detecting violent scenes in videos, in: Hellenic conference on artificial intelligence, Springer, 2010, pp. 91–100.
[28] C. Penet, C.-H. Demarty, G. Gravier, P. Gros, Multimodal information fusion and temporal integration for violence detection in movies, in: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2012, pp. 2393–2396.
[29] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, Advances in neural information processing systems 27.
[30] Z. Wu, Y.-G. Jiang, X. Wang, H. Ye, X. Xue, Multi-stream multi-class fusion of deep networks for video classification, in: Proceedings of the 24th ACM international conference on Multimedia, 2016, pp. 791–800.
[31] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, L. Van Gool, Temporal segment networks for action recognition in videos, IEEE Trans. Pattern Anal. Mach. Intell. 41 (11) (2018) 2740–2755.
[32] Z. Dong, J. Qin, Y. Wang, Multi-stream deep networks for person to person violence detection in videos, Chinese Conference on Pattern Recognition, Springer (2016) 517–531.

[33] E. Acar, F. Hopfgartner, S. Albayrak, Breaking down violence detection: Combining divide-et-impera and coarse-to-fine strategies, Neurocomputing 208 (2016) 225–237.
[34] P. Zhou, Q. Ding, H. Luo, X. Hou, Violent interaction detection in video based on deep learning, in: Journal of physics: conference series, vol. 844, IOP Publishing, 2017, p. 012044.
[35] I. Serrano, O. Deniz, J.L. Espinosa-Aranda, G. Bueno, Fight recognition in video using hough forests and 2d convolutional neural network, IEEE Trans. Image Process. 27 (10) (2018) 4787–4797.
[36] Q. Xu, J. See, W. Lin, Localization guided fight action detection in surveillance videos, in: 2019 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2019, pp. 568–573.
[37] B. Peixoto, B. Lavi, J.P.P. Martin, S. Avila, Z. Dias, A. Rocha, Toward subjective violence detection in videos, in: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 8276–8280.
[38] P. Wu, J. Liu, Y. Shi, Y. Sun, F. Shao, Z. Wu, Z. Yang, Not only look, but also listen: Learning multimodal violence detection under weak supervision, in: European Conference on Computer Vision, Springer, 2020, pp. 322–339.
[39] Y. Su, G. Lin, J. Zhu, Q. Wu, Human interaction learning on 3d skeleton point clouds for video violence recognition, European Conference on Computer Vision, Springer (2020) 74–90.
[40] H. Liu, M. Yao, L. Wang, Svrat: A skeleton-based intelligent monitoring system for violence recognition and abuser tracking, in: 2021 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2021, pp. 1–6.
[41] M. Cheng, K. Cai, M. Li, Rwf-2000: An open large scale video database for violence detection, in: 2020 25th International Conference on Pattern Recognition (ICPR), 2021, pp. 4183–4190.
[42] F.J. Rendón-Segador, J.A. Álvarez-García, F. Enríquez, O. Deniz, Violencenet: Dense multi-head self-attention with bidirectional convolutional lstm for detecting violence, Electronics 10 (13) (2021) 1601.
[43] Z. Islam, M. Rukonuzzaman, R. Ahmed, M. Kabir, M. Farazi, et al., Efficient two-stream network for violence detection using separable convolutional lstm, arXiv preprint arXiv:2102.10590.
[44] M. Asad, J. Yang, J. He, P. Shamsolmoali, X. He, Multi-frame feature-fusion-based model for violence detection, The Visual Computer 37 (6) (2021) 1415–1431.
[45] J. Iqbal, M.A. Munir, A. Mahmood, A.R. Ali, M. Ali, Leveraging orientation for weakly supervised object detection with application to firearm localization, Neurocomputing 440 (2021) 310–320.
[46] C. Feichtenhofer, X3d: Expanding architectures for efficient video recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 203–213.
[47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in neural information processing systems, 2017, pp. 5998–6008.
[48] J.S. Pérez, E. Meinhardt-Llopis, G. Facciolo, Tv-l1 optical flow estimation, Image Processing On Line 2013 (2013) 137–150.
[49] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[50] J. Lin, C. Gan, S. Han, Tsm: Temporal shift module for efficient video understanding, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7083–7093.
[51] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, M.D. Plumbley, Panns: Large-scale pretrained audio neural networks for audio pattern recognition, IEEE/ACM Trans. Audio Speech Language Process. 28 (2020) 2880–2894.
[52] J.F. Gemmeke, D.P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R.C. Moore, M. Plakal, M. Ritter, Audio set: An ontology and human-labeled dataset for audio events, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2017, pp. 776–780.
[53] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781.
[54] E. Nievas, O. Suarez, G. Garcia, R. Sukthankar, Hockey fight detection dataset, in: Computer Analysis of Images and Patterns, Springer, 2011, pp. 332–339.
[55] M.M. Soliman, M.H. Kamal, M.A.E.-M. Nashed, Y.M. Mostafa, B.S. Chawky, D. Khattab, Violence recognition from videos using deep learning techniques, in: 2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS), IEEE, 2019, pp. 80–85.
[56] M. Sjöberg, Y. Baveye, H. Wang, V.L. Quang, B. Ionescu, E. Dellandréa, M. Schedl, C.-H. Demarty, L. Chen, The mediaeval 2015 affective impact of movies task, MediaEval (2015).
[57] B.M. Peixoto, B. Lavi, Z. Dias, A. Rocha, Harnessing high-level concepts, visual, and auditory features for violence detection in videos, J. Vis. Commun. Image Represent. 103174 (2021).
[58] Z. Zheng, W. Zhong, L. Ye, L. Fang, Q. Zhang, Violent scene detection of film videos based on multi-task learning of temporal-spatial features, in: 2021 IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR), IEEE, 2021, pp. 360–365.
[59] Y. Yun, H. Wang, B. Zhang, Y. Jian, Mic-tju in mediaeval 2015 affective impact of movies task, in: Mediaeval Workshop, 2015.
[60] M.G. Constantin, L.D. Stefan, B. Ionescu, C.-H. Demarty, M. Sjoberg, M. Schedl, G. Gravier, Affect in multimedia: Benchmarking violent scenes detection, IEEE Trans. Affect. Comput.
[61] X. Li, Y. Huo, Q. Jin, J. Xu, Detecting violence in video using subclasses, in: Proceedings of the 24th ACM international conference on Multimedia, 2016, pp. 586–590.
[62] A. Jain, D.K. Vishwakarma, Deep neuralnet for violence detection using motion features from dynamic images, in: 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), IEEE, 2020, pp. 826–831.
[63] M.M. Moaaz, E.H. Mohamed, Violence detection in surveillance videos using deep learning, Informatics Bulletin, Helwan University 2 (2) (2020) 1–6.
[64] A.R. Abdali, Data efficient video transformer for violence detection, in: 2021 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT), 2021, pp. 195–199.
[65] Y. Tian, J. Shi, B. Li, Z. Duan, C. Xu, Audio-visual event localization in unconstrained videos, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 247–263.

Yujiang Pu received the B.E. degree from Communication University of China, Beijing, China in 2020, where he is currently pursuing the M.S. degree with the School of Information and Communication Engineering, Communication University of China. His research interests include action recognition, video anomaly detection, and computer vision.

Xiaoyu Wu received the B.E. degree from Jilin University, Changchun, China, in 2004 and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences (CAS), Beijing, China, in 2009. She was also a Visiting Scholar with the North Carolina State University, from 2017 to 2018. She is currently a Professor with the School of Information and Communication Engineering, Communication University of China, Beijing, China. Her research interests include image processing, video understanding, and computer vision.

Shengjin Wang received his B.E. degree from Tsinghua University, Beijing, China, in 1985 and the Ph.D. degree from Tokyo Institute of Technology, Tokyo, Japan, in 1997. From May 1997 to August 2003, he was a Member of the Senior Research Staff in the Internet System Research Laboratories, NEC Corporation, Nara, Japan. Since September 2003, he has been a Professor with the Department of Electronic Engineering, Tsinghua University. He has published more than 100 papers and possessed more than 20 patents. His current research interests include computer vision, pattern recognition, object detection, advanced driving assistant system, and deep learning. Dr. Wang is a director of the Research Center for Media Big-data Cognitive Computing at Tsinghua University. He is also a senior member of IEEE.

Yuming Huang received the B.S. degree in applied physics from the University of Science and Technology Beijing, Beijing, China, in 2013. He received the Ph.D. degree in physics from North Carolina State University, Raleigh, NC, USA, in 2020. He is currently a postdoc associate in University of Massachusetts Chan Medical School, Worcester, MA, USA. His research interests include computational neuroscience and machine learning.

Zihao Liu is currently pursuing the B.E. degree in the School of Information and Communication Engineering from Communication University of China, Beijing, China. His research interests include video understanding, computer vision, and video captioning.

Chaonan Gu received the B.S. and M.S. degrees in the School of Information and Communication Engineering from Communication University of China, Beijing, China in 2018 and 2021, respectively. Her research interests include machine learning and video content understanding.