Not only Look, but also Listen: Learning Multimodal Violence Detection under Weak Supervision

Peng Wu, Jing Liu⋆, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang

School of Artificial Intelligence, Xidian University, Xi'an, China
xdwupeng@gmail.com,yjsun@stu.xidian.edu.cn,
{neouma,shiyujiaaaa,shaofangtao96,15191737495,zwyang97}@163.com

Abstract. Violence detection has been studied in computer vision for years. However, previous work are either superficial, e.g., classification of short-clips, and the single scenario, or undersupplied, e.g., the single modality, and hand-crafted features based multimodality. To address this problem, in this work we first release a large-scale and multi-scene dataset named XD-Violence with a total duration of 217 hours, containing 4754 untrimmed videos with audio signals and weak labels. Then we propose a neural network containing three parallel branches to capture different relations among video snippets and integrate features, where holistic branch captures long-range dependencies using similarity prior, localized branch captures local positional relation using proximity prior, and score branch dynamically captures the closeness of predicted score. Besides, our method also includes an approximator to meet the needs of online detection. Our method outperforms other state-of-the-art methods on our released dataset and other existing benchmark. Moreover, extensive experimental results also show the positive effect of multimodal input and modeling relationships. The code and dataset will be released in https://roc-ng.github.io/XD-Violence/.

Keywords: Violence Detection, Multimodality, Weak Supervision, Relation Networks

1 Introduction

Everyone hopes for peaceful life, and it is our duty to safeguard peace and oppose violence. Violence detection [24,15,4,23,34,35,13,29] in videos has been studied in computer vision community for years. However, due to limited application and challenging nature, this specific task received far less attention than other popular tasks, e.g., video classification [41], action recognition [11], and temporal action detection [12], in the past decades. Along with the advance in video technology in recent years, the application of violence detection is becoming

---

⋆ Corresponding author: neouma@163.com

more and more extensive. For example, violence detection is not only used for real-world scenarios, e.g., intelligent surveillance, but also used for Internet, e.g., video content review (VCR). Violence detection aims to timely locate the start and the end of violent events with minimum human resource cost.

The earliest task of violence detection [24,15] can be considered as video classification. Within this context, most methods assume well-trimmed videos, where violent events last for nearly the entire video. However, such solutions restrict their scope to short clips and cannot be generalized to locate violent events in untrimmed videos, therefore they render a limited use in practice. A small step towards addressing violence detection is to develop algorithms to focus on untrimmed videos. Such as the violent scene detection (VSD) task on MediaEval [4], and the Fighting detector [29]. However, assigning frame-level annotations to videos is a time-consuming procedure which is adverse to building large-scale datasets.

Recently, several research [35,50] focus on weakly supervised violence detection, where only video-level labels are available in the training set. Compared with annotating frame-level labels, assigning video-level labels is labor-saving. Thus, forming large-scale datasets of untrimmed videos and training a data-driven and practical system is no longer a difficult challenge. In this paper, we aim to study weakly supervised violence detection.

Furthermore, we leverage multimodal cue to address violence detection, namely, incorporating both visual and audio information. Multimodal input is beneficial for violence detection as compared to unimodal input. In most cases, visual cue can precisely discriminate and locate events. At times, visual signals are ineffective and audio signals can separate visually ambiguous events. For example, it is difficult for visual signals to figure out what happen in the violently shaking video accompanied by the sounds of explosion, rather, audio signals are the prime discriminators in this case. Therefore, audiovisual fusion can make full use of complementary information and become an extensive tendency in computer vision and speech recognition communities [1,2,31,25,10,19]. We are certainly not the first to attempt to detect violence by multimodal signals, there were precedents for multimodal violence detection before, such as, [48,22,7,28]. However, the above methods have several drawbacks, e.g., relying on small-scale datasets, using subjective hand-crafted features, and the single scene, that indicates a pragmatic system with high generalization is still in the cradle. Unlike these, we are intended to design a reliable neural network based algorithm on large-scale data.

To support research toward leveraging multimodal information (vision and audio) to detect violent events in weakly supervised perspective, we first release a large-scale video violence dataset consisting of 4754 untrimmed videos. Unlike previous datasets [24,15,4,35], our dataset has audio signals and is collected from both movies and in-the-wild scenarios. With the dataset in hands, we then view weakly supervised violence detection as a multiple instance learning (MIL) task; that is, a video is cast as a bag, which consists of several instances (snippets), and instance-level annotations are learned via bag-level labels. Based on

this, we attempt to learn more powerful representations to remedy weak labels. Therefore, we propose a holistic and localized network (HL-Net) that explicitly exploits relations of snippets and learns powerful representations based on these relations, where holistic branch captures long-range dependencies by similarity prior of snippets, and localized branch models short-range interactions within a local neighborhood. In addition, we introduce a holistic and localized cue (HLC) approximator for online violence detection since HL-Net need the whole video to compute relations of snippets. The HLC approximator only processes a local neighborhood and learns precise prediction guided by HL-Net. Even better, HLC approximator brings a dynamic score branch paralleling to holistic branch and localized branch, which computes the response at a position by a weighted sum of all features, and weights depend on predicted scores.

To summarize, contributions of this paper are threefold,

We release a multimodal violence dataset termed XD-Violence, which consists of 4754 untrimmed videos and covers six common types of violence. To our knowledge, XD-Violence is by far the largest scale violence dataset, with a total of 217 hours. Unlike previous datasets, the videos of XD-Violence are captured from multi scenarios, e.g. movies and YouTube.

We introduce a HL-Net to simultaneously capture long-range relations and local distance relations, of which these two relations are based on similarity prior and proximity prior, respectively. In addition, we also propose an HLC approximator for online detection. Based on this, we use a score branch to dynamically obtain an additional holistic relation.

We conducted extensive experiments to verify the effectiveness of our proposed method, and our method shows clear advantage over existing baselines on two benchmarks, i.e., XD-Violence (Ours), and UCF-Crime. Furthermore, experimental results also demonstrate the superiority of multimodal information as compared with the unimodality.

2   Related Work

Violence detection. In the last years, many researchers proposed different methods for violence detection. For instance, Bermejo *et al.* [24] released two well-known fighting datasets. Gao *et al.* [15] proposed violent flow descriptors to detect violence in crowded videos. Mohammadi *et al.* [23] proposed a new behavior heuristic based approach to classify violent and non-violent videos. Most of prior work utilized hand-crafted features to detect violence on small-scale datasets. Common features include, scale-invariant feature transform (SIFT), spatial-temporal interest point (STIP), histogram of oriented gradient (HOG), histograms of oriented optical flow (HOF), motion intensity, and so on.

With the rise of deep convolutional neural networks (CNNs), many work have looked into designing effective deep convolutional neural networks for violence detection. For example, Sudhakaran and Lanz [34] used a convolutional long short-term memory (LSTM) network for the purpose of recognizing violent videos. Similarly, Hanson *et al.* [13] built a bidirectional convolutional LSTM

architecture for violence detection in videos. Peixoto *et al.* [27] used two deep neural network frameworks to learn the spatial-temporal information under different scenarios, then aggregated them by training a shallow neural network to describe violence. Recently, an interesting research [33] was proposed, which used a scatter net hybrid deep learning network for violence detection in drone surveillance videos.

There are several attempts to detect violence with multimodality or audio [48,22,7,28,8,9]. To our knowledge, vast majority of methods use hand-crafted features to extract audio information, e.g., spectrogram, energy entropy, audio energy, chroma, Mel-scale frequency cepstral coefficients (MFCC), zero-crossing rate (ZCR), pitch, etc. Hand-crafted features are easy to extract but are low-level and not robust. Unlike them, our method uses a CNN based model to extract high-level features.

Relation networks. Several work apply the graph neural networks (GCNs) [20,38] over the graph to model relations among different nodes and learn powerful representations for computer vision. For instance, GCN are used for temporal action localization [49], video classification [36,40], anomaly detection [50], skeleton-based action recognition [32,44], point cloud semantic segmentation [21], image captioning [45], and so on. Besides GCN, temporal relation networks [51], designed to learn and reason about temporal dependencies between video frames, are proposed to address video classification. Recently, self-attention networks [39,46,5,18] have been successfully applied in vision problems. An attention operation can affect an individual element by aggregating information from a set of elements, where the aggregation weights are automatically learned.

3   XD-Violence Dataset

3.1   Selecting Violence Categories

The World Health Organization (WHO) defines violence as "the intentional use of physical force or power, threatened or actual, against oneself, another person, or against a group or community, which either results in or has a high likelihood of resulting in injury, death, psychological harm, maldevelopment, or deprivation." Because of the multiple facets of violence, no common and generic enough definition for violent events was ever proposed, even when restricting ourselves to physical violence. However, establishing a clear definition of violence is a key issue because human annotators can rely on a ground truth reference to reduce ambiguity. To mitigate the problem, we consider six physically violent classes, namely, *Abuse*,*Car Accident*, *Explosion*, *Fighting*, *Riot*, and *Shooting*. We take these violence into account due to clear definition, frequent occurrence, widespread use [4,35], and adverse impact to safety.

3.2   Collection and Annotation

Video collection. Previous violence datasets are collected from either movies or in-the-wild scenes, almost no dataset is collected from both. Unlike them, our
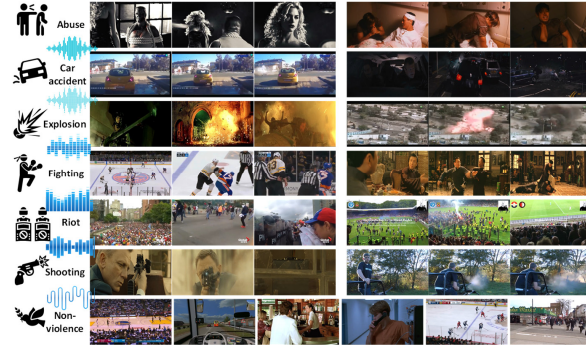
Fig. 1. Sample videos from the XD-Violence dataset.

dataset is collected from both movies and YouTube (in-the-wild scenes). There is a total of 91 movies, of which violent movies are used to collect both violent and non-violent events, and non-violent movies are only used to collect non-violent events. We also collect in-the-wild videos by YouTube. We search and download a mass of video candidates using text search queries. In order to prevent violence detection systems from discriminating violence based on the background of scenarios rather than occurrences, we specifically collect large amounts of non-violent videos whose background is consistent with that of violent videos. A dataset with video-level labels is completed after elaborate efforts of several months. Several example videos from each category are shown in Fig. 1. More details are given in the supplementary material.

Video annotation. Our dataset has a total of 4754 videos, which consists of 2405 violent videos and 2349 non-violent videos. We split it into two parts: the training set containing 3954 videos and the test set including 800 videos, where the test set consists of 500 violent videos and 300 non-violent videos. To evaluate the performance of violence detection methods, we need to make frame-level (temporal) annotations for test videos. To be specific, for each violent video of the test set, we mark the start and ending frames of violent events. As [35], we also assign the same videos to multiple annotators to label the temporal extent of each violence and average annotations of different annotators to make final temporal annotations more precise. Both training and test sets contain all 6 kinds of violence at various temporal locations in the videos.

3.3   Dataset Statistics

Multi-scenario includes but not limited to the following sources: movies, cartoons, sports, games, music, fitness, news, live scenes, captured by CCTV cameras, captured by hand-held cameras, captured by car driving recorders, etc. (some of them may be overlapped.) We also assign multi violent labels ($1 \leq \#$labels ≤ 3) to each violent video owing to the co-occurrence of violent events. The order of
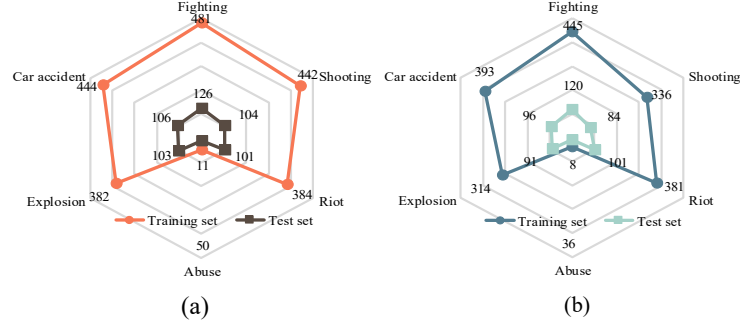
Fig. 2. Dataset Statistics. (a) Distribution of the number of videos belonging to each category according to multi-label. (b) Distribution of the number of videos belonging to each category according to the first label.
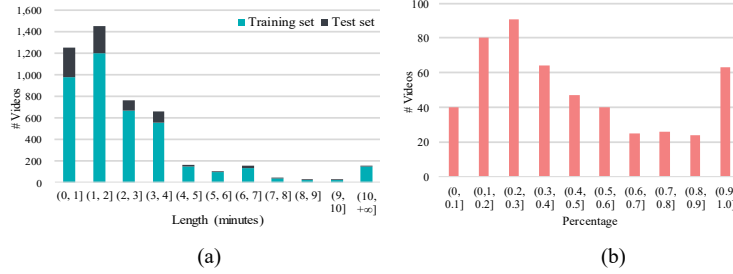

Fig. 3. Dataset Statistics. (a) Distribution of videos according to length (minutes). (b) Distribution of violent videos according to percentage of violence (in each video) in test set.

labels of each video is based on the importance of different violent events occurring in the video. The distribution of the violent videos in terms of the number of labels is shown in Fig. 2. In addition, our dataset consists of untrimmed videos, therefore, we show the distribution of videos in terms of length in Fig. 3(a). We also present the percentage of violence in each test video in Fig. 3(b).

3.4   Dataset Comparisons

In order to highlight the traits of our dataset, we compare our dataset with other widely-used datasets for violence detection. These datasets can be split into three types: small scale, medium scale, and large scale, of which Hockey [24], Movie [24], and Violent-Flows [15] are small-scale, VSD [4] and CCTV-Fights [29] are medium-scale, and the remaining UCF-Crime [35] is large-scale. Table 1 compares several characteristics of these datasets. Our dataset is by far the largest dataset, which is more than 300 times than the total of small datasets, 4

Table 1. Comparisons of different violence datasets.*means quite a few videos are silent or only contain background music.

Dataset	#Videos	Length	Source of scenarios	#Violence types	Audio
Hockey [24]	1000	27 min	Ice hockey	1	No
Movie [24]	200	6 min	Movies and sports	1	No
Violent-Flows [15]	246	15 min	Streets, school, and sports	1	Yes*
CCTV-Fights [29]	1000	18 hours	CCTV and mobile cameras	1	Yes*
VSD [4]	25	35 hours	Movies	8	Yes
UCF-Crime [35]	1900	128 hours	CCTV camera	9	No
XD-Violence (Ours)	4754	217 hours	Movies, sports, games, hand-held cameras, CCTV, car cameras, etc.	6	Yes

times than the total of medium datasets, and almost 2 times than the UCF-Crime dataset. Besides, variations in scenes of previous datasets are also limited, by contrast, our dataset embraces a wide variety of scenarios. In addition, Hockey, Movie, Violent-Flows and CCTV-Fights are only used for fighting detection. Intriguingly, though Violent-Flows and CCTV-Fights contain audio signals, in fact, there are quite a few videos that are silent or only contain background music, which is inoperative or even harmful for training detection methods taking multimodality as input. The UCF-Crime dataset is used to detect violence in surveillance videos, but lacks audio.

Overall, our dataset has three good traits: 1) large scale, which is beneficial for training generalizable methods for violence detection; 2) diversity of scenarios, so that violence detection methods actively respond to complicated and diverse environments and are more robust; 3) containing audio signals, making algorithms leverage multimodal information and more confidence.

4 Methodology

4.1 Multimodal Fusion

Our proposed method is summarized in Fig. 4. Consider that we have a training set of videos, we denote an untrimmed video and the corresponding label as V and y, where $y \in \{0, 1\}, y = 1$ denotes V covers violent events. With a video V in hands, we use feature extractors F^V and F^A to extract visual and audio feature matrix X^V and X^A using the sliding window mechanism, where $X^V \in \mathbb{R}^{T' \times d^V}, X^A \in \mathbb{R}^{T' \times d^A}$, T' is the length of feature matrix, x_i^V and x_i^A are visual and audio features of the ith snippet, respectively.

Quite a few fusion manners have been proposed for multimodal input, we there opt to the simple yet effective concatenation fusion. More precisely, X^V and X^A are first concatenated in the channels, then the concatenation passes through
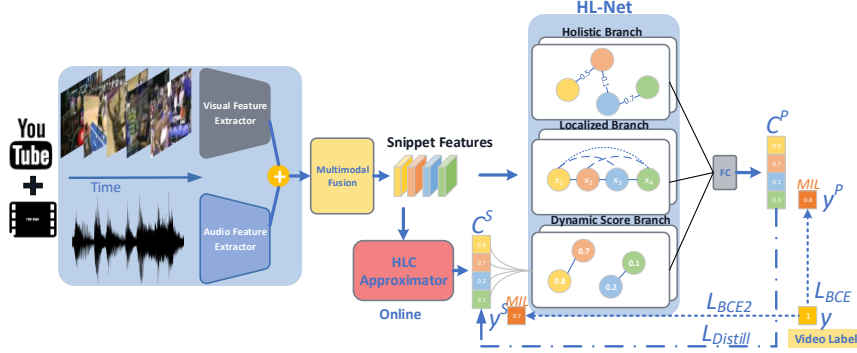
Fig. 4. The pipeline of our proposed method.

two stacked fully connected layers (FC) with 512 and 128 nodes respectively, where each FC layer is followed by ReLU and dropout. Finally, the output is taken as the fusion features, which is denoted by X^F.

4.2 Holistic and Localized Networks

Revisit relations. We first recap the long-range dependencies of neural networks, which can be acquired by two prevalent types of networks, namely, GCNs and non-local networks (NL-Net, self-attention networks).

A general graph convolution operation can be formulated as follows [21],

$$X_{l+1} = Update\left(Aggregate(X_l, W_l^{agg}), W_l^{update}\right) \tag{1}$$

which contains two essential operations, aggregation and update, and corresponding learnable weights, where the aggregation operation is used to compile information from the global vertices (long-range dependencies), while update functions perform a non-linear transform to compute new representations.

An instantiated non-local operation can be formulated as follows [39,46],

$$X_{l+1} = \left[softmax(X_l^T W_\theta^T W_\phi X_l)(W_g X_l)\right] W_\psi \tag{2}$$

Although they have different original intentions (GCN is mainly used to address problems with non-Euclidean data, capturing long-range dependencies is an avocation), they are similar on capturing long-range dependencies. Because the term within the outer brackets in Eq. (2) can be viewed as an aggregation operation based on feature similarity, which is followed by the update operation. **Holistic branch implementations.** Inspired by the GCN for video understanding [50,40,49], we define the holistic relation matrix by feature similarity prior, as shown below:

$$A_{ij}^H = g\left(f(x_i, x_j)\right) \tag{3}$$

where $A^H \in T' \times T'$, A_{ij}^H measures the feature similarity between the ith and j^{th} features. g is the normalization function, and function f computes the similarity of a pair of features, we define f as follows,

$$f(x_i, x_j) = \frac{x_i^T x_j}{\|x_i\|_2 \cdot \|x_j\|_2} \tag{4}$$

we also define other versions of f, and present them on the supplementary material. f bounds the similarity to the range of (0, 1] for the sake of the thresholding that filters weak relations and strengthens correlations of more similar pairs. The thresholding operation can be defined as follows,

$$f(x_i, x_j) = \begin{cases} f(x_i, x_j) & f(x_i, x_j) > \tau \\ 0 & f(x_i, x_j) \le \tau \end{cases} \tag{5}$$

where τ is the threshold. After that, we adopt the softmax as the normalization function g to make sure the sum of each row of A is 1, as shown in Eq. (6):

$$A_{ij}^H = \frac{exp(A_{ij}^H)}{\sum_{k=1}^{T'} exp(A_{ik}^H)} \tag{6}$$

We emphasize that X used in Eq. (3) is the concatenation of raw features to capture the original feature prior.

In order to capture long-range dependencies, we follow the GCN paradigm and design the holistic layer as,

$$X_{l+1}^H = Dropout\left(ReLU(A^H X_l^H W_l^H)\right) \tag{7}$$

which allows us to compute the response of a position defined by the similarity prior based on the global filed rather than its neighbors.

Localized branch implementations. Non-local operations capture long-range dependencies directly by computing interactions between any two positions, regardless of their positional distance [39]. However, positional distance has positive effects on temporal events detection [50], and to retain it, we devise the local relation matrix based on proximity prior as,

$$A_{ij}^L = exp\left(\frac{-|i - j|^\gamma}{\sigma}\right) \tag{8}$$

which only depends on temporal positions of the ith and j^{th} features, and where γ and σ are hyper-parameters to control the range of influence of distance relation. Likewise, X_{l+1}^L is the output of the $(l+1)^{th}$ localized layer.

4.3   Online Detection

As we mentioned, a violence detection system is not only applied for offline detection (Internet VCR), but also online detection (surveillance system). However, online detection by the above HL-Net is impeded by a major obstacle: HL-Net

needs the whole video to obtain long-range dependencies. To jump out of the dilemma, we propose an HLC approximator, only taking previous video snippets as input, to generate precise predictions guided by HL-Net. Two stacked FC layers followed by ReLU and a 1D causal convolution layer constitute HLC approximator. The 1D causal convolution layer has kernel size 5 in time with stride 1, sliding convolutional filters over time. The 1D causal convolution layer also acts as the classifier, whose output is the violent activation denoted as C^S of shape T'. Even better, this operation introduces an additional branch named dynamic score branch to extend HL-Net, which depends on C^S. The main role of this branch is to compute the response at a position as a weighted sum of the features at all positions, where weights depend on the closeness of the scores. Different from the relation matrices of holistic and localized branches, the relation matrix of score branch is updated in each iteration, and depends on predicted scores rather than the prior. Formally, the relation matrix of score branch is devised as follows,

$$A_{ij}^S = \rho \left(1 - \left|s(C_i^S) - s(C_j^S)\right|\right) \tag{9}$$

$$\rho(x) = \frac{1}{1 + exp(-\frac{x-0.5}{0.1})} \tag{10}$$

where function s is sigmoid, and function ρ is used to enhance (and weak) the pairwise relation where the closeness of the scores is greater (and less) than 0.5, and softmax is also used for the normalization.

Analogously, X_{l+1}^S is the output of the $(l+1)^{th}$ score layer, where $X_0^S = (X_0^H = X_0^L) = X^F$.

4.4   Training Based on MIL

We use an FC layer with 1 node to project the concatenation representations to the label space (1D space), and the violent activations we obtain after this projection can be represented as follows,

$$C^P = (X^H \| X^L \| X^S)W \tag{11}$$

where $\|$ denotes the concatenation operation, and $C^P \in \mathbb{R}^{T'}$ denotes the violent activations.

Following the principles of MIL [35,26], we use the average of K-max activation (C^P and C^S) over the temporal dimension rather than the whole activations to compute y^P and y^S, and K is defined as $\left\lfloor \frac{T'}{q} + 1 \right\rfloor$. The instances corresponding to the K-max activation in the positive bag is most likely to be true positive instances (violence). The instances corresponding to the K-max activation in the negative bag is hard instances. We expect these two types of instances to be as far as possible.

We define the classification loss, L_{BCE} and L_{BCE2}, as the binary cross-entropy between the predicted labels (y^P and y^S) and ground truth y. In addition, we also use the knowledge distillation loss to encourage the output of the HLC approximator to approximate the output of HL-Net.

$$L_{DISTILL} = \sum_{j=1}^{N} \left(- \sum_{i} s(C_i^P) log\left(s(C_i^S)\right) \right)_j \tag{12}$$

where N is the batch size. Finally, the total loss is the weighted sum of the above three loss, which is shown as follows,

$$L_{TOTAL} = L_{BCE} + L_{BCE2} + \lambda L_{DISTILL} \tag{13}$$

4.5 Inference

Aiming at different requests, our method can choose offline or online manners to efficiently detect violent events. Sigmoid functions follow the violent activations C^P and C^S and generate the violence confidence (score) that is bounded in the range of [0, 1]. Note that, in the online inference, only HLC approximator works, and HL-Net can be removed.

5 Experiments

5.1 Evaluation Metric

we utilize the frame-level precision-recall curve (PRC) and corresponding area under the curve (average precision, AP) [29] rather than receiver operating characteristic curve (ROC) and corresponding area under the curve (AUC) [43,42] since AUC usually shows an optimistic result when dealing with class-imbalanced data, and PRC and AP focus on positive samples (violence).

5.2 Implementation Details

Visual features. We utilize two mainstream networks as our visual feature extractor F^V, namely, C3D [37] and I3D [3] networks. We extract *fc6* features from C3D that is pretrained on the Sports-1M dataset, and extract *global_pool* features from I3D pre-trained on Kinetics-400 dataset. I3D is a two-stream model, therefore, the visual feature has two versions, RGB and optical flow. We use the GPU implementation of TV-L1 [47] to compute the optical flow. We fix the frame rate to 24 FPS for all videos, and set the length of sliding window as 16 frames.

Audio features. we leverage the VGGish [6,17] network pretrained on a large YouTube dataset as the audio feature extractor F^A due to its remarkable performance on audio classification. The audio is divided into overlapped 960-ms segments, of which each segment has the only corresponding video snippet with
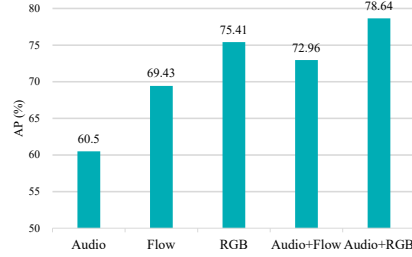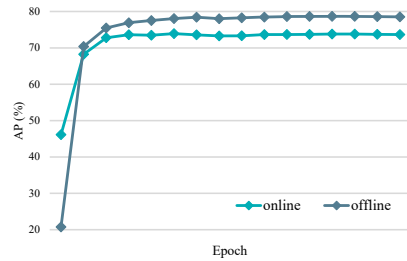
Fig. 5. AP comparison of different modalities.

Fig. 6. AP changing curves using models saved in different epoch.

aligned end time. The log-mel spectrogram patches of 96×64 bins, computed from segments, form the input to VGGish. Unless otherwise stated, we use the fusion of RGB feature of I3D and VGGish feature by default.

HL-Net architecture. Generalized HL-Net is formed from holistic branch, localized branch, and an additional score branch. Each branch is a stack of two layers, where the number of output channels for each layer is 32. Furthermore, taking inspiration from [16,32,21], we add a residual connection for each layer, which enables GCN to reliably converge in the training stage.

Training details. We implement the network based on PyTorch. For hyperparameters, without otherwise stated, we set τ as 0.7, γ and σ of A^L as 1, q as 16, dropout rate as 0.7, and λ as 5. For network optimization, Adam is used as the optimizer. The initial learning rate is set as 10^{-3}, and is divided by 10 at the 10th epoch and 30th epoch. The network is trained for 50 epochs in total, and the batch size is 128.

5.3   Ablation Studies

The effect of modality. Most of work still focus on the unimodality for video event detection, and multimodal input is one advantage of our work. Therefore, we conduct experiments on the XD-Violence dataset to verify the superiority of multimodality. We try five different inputs, i.e., audio modality, optical flow (Flow) modality, RGB modality, the fusion of audio and Flow modalities, the fusion of audio and RGB modalities, (For the unimodality, the two stacked FC layers of multimodal fusion module are still retained.) and report the results in Fig. 5. For the unimodality, a key observation is that visual modality is significantly superior to audio, which is not hard to understand: visual signals contain more rich information, which make algorithms see farther and more. More importantly, the fusion of multimodality shows clear advantage over the single modality, to be specific, the fusion of audio and Flow achieves clear improvements against the Flow-only input by 3.5% on AP, and the fusion of audio and RGB outperforms the RGB-only input by 3.2% on AP. This matches the

expectation that audio provides complementary information for visual signals.

The effect of holistic, localized and score branches. As depicted in Section 4, there are three parallel branches in our HL-Net. We manually delete one or two of the branches and show their performance in Table 2. We observe that: 1) three separate branches achieve similar performance; 2) removing any one of the three branches will harm the performance; 3) HL-Net achieves the best performance with all three branches work together. This demonstrates that all branches play an irreplaceable role in our HL-Net.

Online detection vs Offline detection. We show the performance comparison between online detection and offline detection in Table 3. It is observed that offline detection outperforms online detection by 5% on AP. The performance improvement benefits from the powerful ability of HL-Net, i.e., computing the response by three different types of feature aggregation. We also use the models saved in different training stages to inference, and show the performance changing curves. We can see from Fig. 6 that offline detection outperforms online detection except for in the initial training stage, we argue that this is because HLC approximator is a lightweight module that is easy to train and can find a good solution in the early stage, and HL-Net has relatively more parameters, needing more time to train.

5.4   Comparisons with State-of-the-Arts

We compare our method with several baselines on XD-Violence, and show the results in Table 3. It is obvious that our method can outperform current state-of-the-art methods. We also show the performance of the fusion of C3D and audio features in Table 3, and observe that C3D is inferior to I3D by a large margin in our violence detection task. Note that all baselines in Table 3 take the fusion of the RGB features of I3D and VGGish features as input. PRC on the XD-Violence and results on the UCF-Crime dataset are given in the supplementary material.
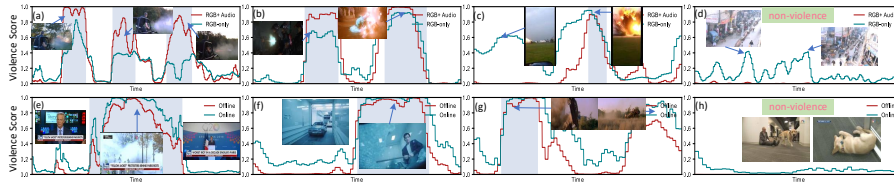
5.5   Qualitative Results

We present several qualitative examples in Fig. 7. (a)-(c) and (e)-(g) are violent videos, and our method successfully detects violent events. (d) and (h) are non-violent videos, and our method generates very low violence scores. As can be seen from the 1st row, multimodal input can localize violence more precisely than the RGB-only input with lower false positives and false negatives, especially in videos with audio. For instance, in (c), Audio+RGB detects the explosion according to not just fire, but explosive sound. From the 2nd row it is evident that online detection is slightly worse than offline detection, and with higher false alarms, this is because of the lack of contextual information. Besides, an interesting finding is that our method considers there are violent events in the second half of (g), but this part is not marked as violence. After watching the video, we find

Table 2. AP comparison of different branches.

Holisitic branch	Localized branch	Score branch	AP(%)
✓			75.44
	✓		76.60
		✓	75.40
✓	✓		77.23
✓		✓	77.05
	✓	✓	77.70
✓	✓	✓	78.64

Table 3. AP comparison on the XD-Violence dataset.

Method	AP(%)
SVM baseline	50.78
OCSVM [30]	27.25
Hasan *et al.* [14]	30.77
Sultani *et al.* [35]	73.20
Ours (C3D)	67.19
Ours (Online)	73.67
Ours	78.64

Fig. 7. Qualitative results of our method on test videos. Colored window shows the ground truth of violent regions. [Best viewed in color.]

out the reason: the accident process is finished, but there is still an overturned car. More qualitative results are provided in the supplementary material.

6   Conclusions

In this paper, we study the large-scale violence detection with multimodality under weak supervision. Due to lack of applicable datasets, we first release a large-scale violence dataset to fill the gap. Then we propose a method to explicitly modeling relationships among video snippets and learn powerful representations. Extensive experiments show, 1) our dataset is applicable; 2) multimodality significantly improves the performance; 3) explicitly exploiting relations is highly effective. In the further work, our research will naturally extend to multi-class violence detection as XD-Violence is a multi-label dataset. In addition, more powerful online detection is yet to be further explored.

References

1. Aytar, Y., Vondrick, C., Torralba, A.: Soundnet: Learning sound representations from unlabeled video. In: Advances in neural information processing systems (NIPS). pp. 892–900 (2016)
2. Aytar, Y., Vondrick, C., Torralba, A.: See, hear, and read: Deep aligned representations. arXiv preprint arXiv:1706.00932 (2017)
3. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6299–6308 (2017)
4. Demarty, C.H., Penet, C., Soleymani, M., Gravier, G.: Vsd, a public dataset for the detection of violent scenes in movies: design, annotation, analysis and evaluation. Multimedia Tools and Applications **74**(17), 7379–7404 (2015)
5. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3146–3154 (2019)
6. Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 776–780. IEEE (2017)
7. Giannakopoulos, T., Makris, A., Kosmopoulos, D., Perantonis, S., Theodoridis, S.: Audio-visual fusion for detecting violent scenes in videos. In: Hellenic conference on artificial intelligence. pp. 91–100. Springer (2010)
8. Giannakopoulos, T., Pikrakis, A., Theodoridis, S.: A multi-class audio classification method with respect to violent content in movies using bayesian networks. In: IEEE 9th Workshop on Multimedia Signal Processing. pp. 90–93. IEEE (2007)
9. Giannakopoulos, T., Pikrakis, A., Theodoridis, S.: A multimodal approach to violence detection in video sharing sites. In: 20th International Conference on Pattern Recognition (ICPR). pp. 3244–3247. IEEE (2010)
10. Ginosar, S., Bar, A., Kohavi, G., Chan, C., Owens, A., Malik, J.: Learning individual styles of conversational gesture. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3497–3506 (2019)
11. Girdhar, R., Carreira, J., Doersch, C., Zisserman, A.: Video action transformer network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 244–253 (2019)
12. Gu, C., Sun, C., Ross, D.A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., et al.: Ava: A video dataset of spatio-temporally localized atomic visual actions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6047–6056 (2018)
13. Hanson, A., Pnvr, K., Krishnagopal, S., Davis, L.: Bidirectional convolutional lstm for the detection of violence in videos. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 0–0 (2018)
14. Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K., Davis, L.S.: Learning temporal regularity in video sequences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 733–742 (2016)
15. Hassner, T., Itcher, Y., Kliper-Gross, O.: Violent flows: Real-time detection of violent crowd behavior. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. pp. 1–6. IEEE (2012)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)

17. Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., et al.: Cnn architectures for large-scale audio classification. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 131–135. IEEE (2017)
18. Hu, H., Gu, J., Zhang, Z., Dai, J., Wei, Y.: Relation networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3588–3597 (2018)
19. Kazakos, E., Nagrani, A., Zisserman, A., Damen, D.: Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 5492–5501 (2019)
20. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
21. Li, G., Muller, M., Thabet, A., Ghanem, B.: Deepgcns: Can gcns go as deep as cnns? In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 9267–9276 (2019)
22. Lin, J., Wang, W.: Weakly-supervised violence detection in movies with audio and video based co-training. In: Pacific-Rim Conference on Multimedia. pp. 930–935. Springer (2009)
23. Mohammadi, S., Perina, A., Kiani, H., Murino, V.: Angry crowds: Detecting violent events in videos. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 3–18. Springer (2016)
24. Nievas, E.B., Suarez, O.D., García, G.B., Sukthankar, R.: Violence detection in video using computer vision techniques. In: International Conference on Computer Analysis of Images and Patterns. pp. 332–339. Springer (2011)
25. Oh, T.H., Dekel, T., Kim, C., Mosseri, I., Freeman, W.T., Rubinstein, M., Matusik, W.: Speech2face: Learning the face behind a voice. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7539–7548 (2019)
26. Paul, S., Roy, S., Roy-Chowdhury, A.K.: W-talc: Weakly-supervised temporal activity localization and classification. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 563–579 (2018)
27. Peixoto, B., Lavi, B., Martin, J.P.P., Avila, S., Dias, Z., Rocha, A.: Toward subjective violence detection in videos. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 8276–8280. IEEE (2019)
28. Penet, C., Demarty, C.H., Gravier, G., Gros, P.: Multimodal information fusion and temporal integration for violence detection in movies. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2393–2396. IEEE (2012)
29. Perez, M., Kot, A.C., Rocha, A.: Detection of real-world fights in surveillance videos. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2662–2666. IEEE (2019)
30. Schölkopf, B., Williamson, R.C., Smola, A.J., Shawe-Taylor, J., Platt, J.C.: Support vector method for novelty detection. In: Advances in Neural Information Processing Systems (NIPS). pp. 582–588 (2000)
31. Senocak, A., Oh, T.H., Kim, J., Yang, M.H., So Kweon, I.: Learning to localize sound source in visual scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4358–4366 (2018)
32. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12026–12035 (2019)

33. Singh, A., Patil, D., Omkar, S.: Eye in the sky: Real-time drone surveillance system (dss) for violent individuals identification using scatternet hybrid deep learning network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1629–1637 (2018)
34. Sudhakaran, S., Lanz, O.: Learning to detect violent videos using convolutional long short-term memory. In: IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). pp. 1–6. IEEE (2017)
35. Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6479–6488 (2018)
36. Sun, C., Shrivastava, A., Vondrick, C., Murphy, K., Sukthankar, R., Schmid, C.: Actor-centric relation network. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 318–334 (2018)
37. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 4489–4497 (2015)
38. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. arXiv preprint arXiv:1710.10903 (2017)
39. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7794–7803 (2018)
40. Wang, X., Gupta, A.: Videos as space-time region graphs. In: Proceedings of the European conference on computer vision (ECCV). pp. 399–417 (2018)
41. Wu, C.Y., Feichtenhofer, C., Fan, H., He, K., Krahenbuhl, P., Girshick, R.: Long-term feature banks for detailed video understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 284–293 (2019)
42. Wu, P., Liu, J., Li, M., Sun, Y., Shen, F.: Fast sparse coding networks for anomaly detection in videos. Pattern Recognition p. 107515 (2020)
43. Wu, P., Liu, J., Shen, F.: A deep one-class neural network for anomalous event detection in complex scenes. IEEE Transactions on Neural Networks and Learning Systems **31**(7), 2609–2622 (2020)
44. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Thirty-second AAAI Conference on Artificial Intelligence (AAAI) (2018)
45. Yao, T., Pan, Y., Li, Y., Mei, T.: Exploring visual relationship for image captioning. In: Proceedings of the European conference on computer vision (ECCV). pp. 684–699 (2018)
46. Yue, K., Sun, M., Yuan, Y., Zhou, F., Ding, E., Xu, F.: Compact generalized non-local network. In: Advances in Neural Information Processing Systems (NIPS). pp. 6510–6519 (2018)
47. Zach, C., Pock, T., Bischof, H.: A duality based approach for realtime tv-l 1 optical flow. In: Joint Pattern Recognition Symposium. pp. 214–223. Springer (2007)
48. Zajdel, W., Krijnders, J.D., Andringa, T., Gavrila, D.M.: Cassandra: audio-video sensor fusion for aggression detection. In: IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). pp. 200–205. IEEE (2007)
49. Zeng, R., Huang, W., Tan, M., Rong, Y., Zhao, P., Huang, J., Gan, C.: Graph convolutional networks for temporal action localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 7094–7103 (2019)

50. Zhong, J.X., Li, N., Kong, W., Liu, S., Li, T.H., Li, G.: Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1237–1246 (2019)
51. Zhou, B., Andonian, A., Oliva, A., Torralba, A.: Temporal relational reasoning in videos. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 803–818 (2018)