machine learning & knowledge extraction

MDPI

Article

Leveraging Multi-Modality and Enhanced Temporal Networks for Robust Violence Detection

Gwangho Na [1], Jaepil Ko [2] and Kyungjoo Cheoi [1,*]

[1] Department of Computer Science, Chungbuk National University, Cheongju 28644, Republic of Korea; gh369ho@chungbuk.ac.kr
[2] Department of Computer Engineering, Kumoh National Institute of Technology, Gumi 39177, Republic of Korea; nonezero@kumoh.ac.kr
* Correspondence: kjcheoi@chungbuk.ac.kr

Abstract: In this paper, we present a novel model that enhances performance by extending the dual-modality TEVAD model—originally leveraging visual and textual information—into a multi-modal framework that integrates visual, audio, and textual data. Additionally, we refine the multi-scale temporal network (MTN) to improve feature extraction across multiple temporal scales between video snippets. Using the XD-Violence dataset, which includes audio data for violence detection, we conduct experiments to evaluate various feature fusion methods. The proposed model achieves an average precision (AP) of 83.9%, surpassing the performance of single-modality approaches (visual: 73.9%, audio: 67.1%, textual: 29.9%) and dual-modality approaches (visual + audio: 78.8%, visual + textual: 78.5%). These findings demonstrate that the proposed model outperforms models based on the original MTN and reaffirm the efficacy of multi-modal approaches in enhancing violence detection compared to single- or dual-modality methods.

Keywords: multi-modality; violence detection; feature fusion; enhanced MTN

Citation: Na, G.; Ko, J.; Cheoi, K. Leveraging Multi-Modality and Enhanced Temporal Networks for Robust Violence Detection. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 2422–2434. https://doi.org/10.3390/make6040119

Academic Editors: Antonio Fernández-Caballero and Byung-Gyu Kim

Received: 11 September 2024
Revised: 10 October 2024
Accepted: 25 October 2024
Published: 28 October 2024

1. Introduction

In recent years, the exponential growth in video content consumption has driven increasing demand for automated technologies capable of detecting violent scenes in videos. These technologies play a critical role in providing filtering and warning systems to protect users and regulate content on social media platforms and streaming services. Additionally, they are invaluable in surveillance systems, enabling real-time monitoring and alerting authorities to violent behavior for a rapid response.

Previous research on video violence detection has predominantly focused on visual information [1–6]. While effective in some cases, this approach faces limitations in detecting violence in complex, real-world scenarios. To achieve more accurate detection, it is essential to adopt multi-modal approaches that integrate not only visual data but also other modalities. Multi-modal techniques enhance video interpretation by incorporating semantic information that visual data alone may fail to capture. For instance, audio information can detect violence through sounds—such as screams or breaking glass—that visual cues may not convey. Even in the absence of explicit violent imagery, audio data can help identify threatening situations. Similarly, captions generated by video captioning models (textual information) provide a natural language interpretation of the video, offering context that enhances understanding. For example, a caption like "Two men are fighting on the street" can directly indicate a violent scenario.

Building on these observations, this research seeks to address the following question: Can a multi-modal approach that integrates visual, audio, and textual data enhance the accuracy of violent scene detection in videos compared to traditional methods that rely solely on visual data? We hypothesize that incorporating audio and textual information,

alongside visual data, will improve the accuracy of violence detection in videos by capturing semantic meanings that are challenging to extract from visual features alone.

In this paper, we propose an integrated approach that extends the text-empowered video anomaly detection (TEVAD) model [7] by incorporating audio information and leveraging textual data generated via video captioning models. This approach captures semantic meanings that are difficult to extract using visual features alone. The key differences from the original TEVAD [7] include the addition of audio data and modifications to both the feature fusion method and the timing of fusion. Additionally, we introduce an enhanced multi-scale temporal network (enhanced MTN) to improve the extraction of temporal features across multiple time scales. These enhancements significantly boost the performance of violence detection in video content.

The structure of this paper is as follows. Section 2 provides an overview of prior research on video violence detection and the original TEVAD model. Section 3 presents the architecture of the proposed system, including the feature fusion strategy and the enhanced MTN. In Section 4, we describe the experimental setup, outline the multi-modal expansion methodology, and analyze the results of experiments conducted to validate the effectiveness of the multi-modal approach and enhanced MTN. Finally, in Section 5, we summarize the key findings and discuss potential avenues for future research.

2. Related Work

Video violence detection aims to identify violent events within video content. The traditional systems for violence detection typically rely on large-scale labeled datasets; however, labeling such datasets is costly and time-consuming. To address this challenge, recent research has increasingly adopted weakly supervised learning [1], which trains models using partially or indirectly labeled data, thereby reducing the burden of manual labeling.

Sultani et al. [1] were the first to introduce weakly supervised learning for video anomaly detection and proposed the UCF-Crime dataset, a large-scale dataset for this purpose. Most weakly supervised learning methods are based on the multiple instance learning (MIL) framework. In MIL, labels are assigned to entire video clips rather than individual frames or segments, allowing the model to infer labels for specific video portions. Zhong et al. [2] improved anomaly detection by incorporating graph convolutional networks (GCNs) to refine label noise, thereby enhancing the action classification models' performance. Tian et al. [3] advanced anomaly detection further by introducing robust temporal feature magnitude learning (RTFM), which leverages dilated convolutions and self-attention mechanisms to model temporal relationships in visual features, improving the differentiation between normal and anomalous video clips.

While these methods have primarily focused on single-modal approaches using visual data, relying solely on one modality is often insufficient for detecting complex and diverse violent situations in videos. To overcome this limitation, Wu et al. [8] proposed a multi-modal approach that combines weakly supervised learning with both visual and audio information. In their method, entire video clips are labeled as violent or non-violent, and the model subsequently identifies violent moments within the clips. They introduced the XD-Violence dataset, a large-scale multi-modal dataset designed for weakly supervised violence detection, which has catalyzed further research in multi-modal approaches. Wu et al. [9] also proposed a method that fuses visual and audio features, improving violence detection accuracy by modeling temporal continuity and capturing similarities between video clips using techniques such as temporal modeling and graph neural networks (GNNs).

Zhang et al. [10] introduced a strategy to enhance model performance by evaluating the completeness and uncertainty of pseudo-labels during training. Their method extracts visual features on a frame-by-frame basis using CNN-based models and learns temporal relationships between frames using RNN-based models. The extracted features are then fused to detect anomalies, with the completeness and uncertainty of pseudo-labels incorporated into the learning process. Similarly, Pang et al. [11] proposed an efficient method for extracting and integrating visual and audio features, demonstrating superior

performance on the XD-Violence dataset. Zhou et al. [12] introduced uncertainty regulated dual memory units (UR-DMUs), which manage visual and temporal features by assigning lower weights to video clips with high uncertainty, thus reducing their impact on training. Conversely, clips with low uncertainty are given higher weights, improving the model's overall performance.

Chen et al. [7] extended RTFM [3] by proposing text-empowered video anomaly detection (TEVAD), a text-based video anomaly detection system that integrates generated text features with visual features to capture semantic information that visual features alone might miss. The overall structure of TEVAD is illustrated in Figure 1. In this approach, input videos are divided into T snippets, with each snippet processed by separate text and visual branches. In the text branch, captions generated by the video captioning model SwinBERT [13] are converted into sentence embeddings using SimCSE [14] to create text features. In the visual branch, visual features are extracted using a pre-trained visual feature extractor (I3D) [15]. Both sets of features are processed by a multi-scale temporal network (MTN) to capture temporal features across multiple time scales; afterward, the features are fused for anomaly detection. The anomaly detection classifier generates anomaly scores for each snippet, which are used to predict frame-level anomaly scores. TEVAD demonstrated improved performance over traditional methods by leveraging text information to capture the semantic meaning of videos.

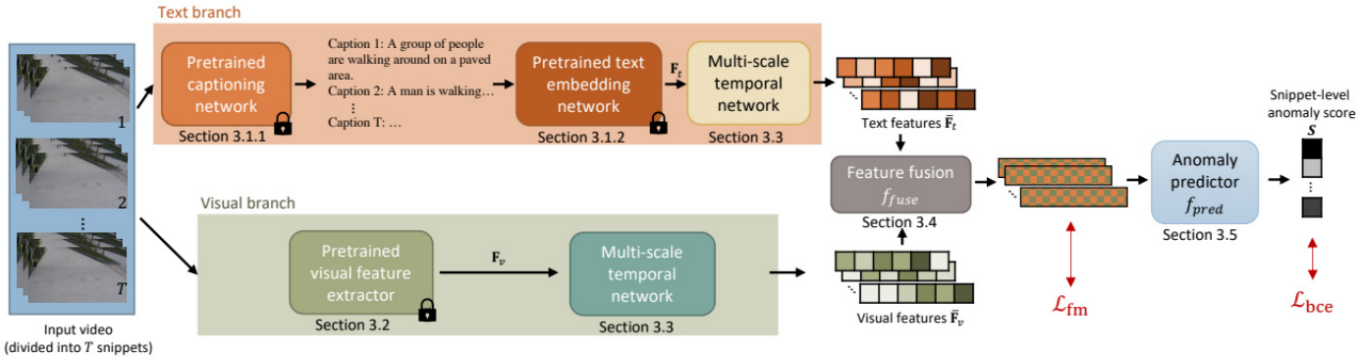

Figure 1. Overall structure of TEVAD [7].

Previous research on video violence detection has predominantly focused on single-modal approaches leveraging visual data or dual-modal approaches combining visual and audio data, or visual and textual information. In contrast, this paper introduces a multi-modal approach that integrates audio, visual, and textual data. Our experimental results demonstrate that fusing multi-modal features yields superior performance in video violence detection compared to both single-modal and dual-modal approaches. Additionally, we propose an enhanced multi-scale temporal network (enhanced MTN, E-MTN), which improves the extraction of temporal features across multiple time scales, further boosting performance.

3. Methodology

Figure 2 illustrates the overall structure of the proposed system. The input video is divided into T snippets, and for each snippet, the visual features (F_v), text features (F_t), and audio features (F_a) are extracted (Section 3.1). These extracted features are then fused into a single feature, F_{vta} (Section 3.2). The fused features are subsequently passed through the enhanced multi-scale temporal network (E-MTN) to generate multi-scale temporal features, $F_{E\text{-}MTN}$ (Section 3.3). The generated $F_{E\text{-}MTN}$ is used to calculate the feature magnitude of each snippet. The top k largest feature magnitudes from normal and violent videos are passed to the violence detection classifier to train a binary classifier for snippet-level violence detection (Section 3.4).

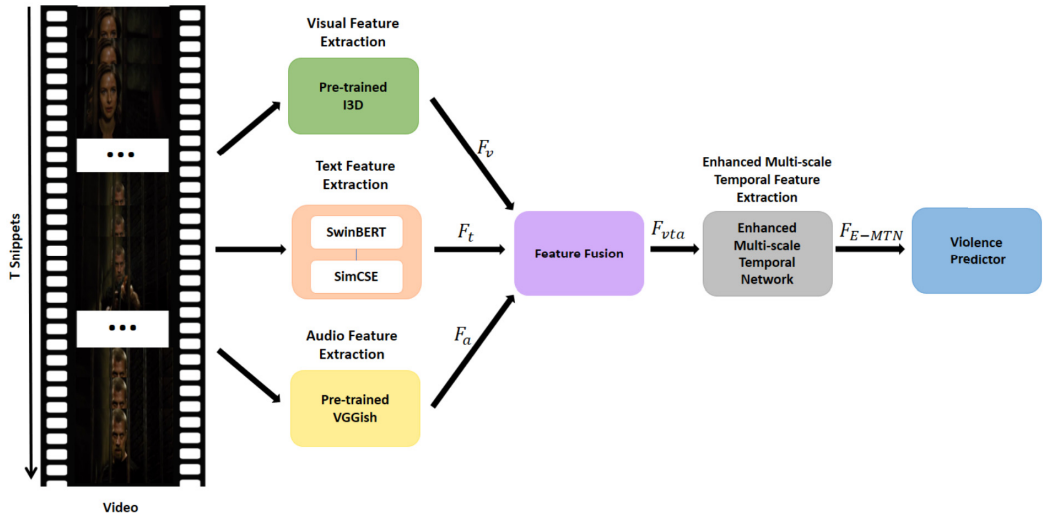

Figure 2. Overall structure of the proposed system.

3.1. Feature Extraction

Visual features (F_v) are extracted using an Inflated 3D ConvNet (I3D) [15], pre-trained on the Kinetics-400 dataset [15]. I3D extends 2D convolutions to 3D, allowing the model to capture both temporal and spatial features simultaneously. Following the TEVAD approach, we apply the five-crop augmentation technique to the XD-Violence dataset. This technique generates image crops from five positions—four corners and the center—enhancing the diversity of the training data.

Text features (F_t) are generated using SwinBERT, a model pre-trained on the Video and Text EXplanations (VATEX) dataset [16]. VATEX is derived from the Kinetics-600 dataset [17] and comprises 41,250 video clips, each accompanied by natural language descriptions (captions) in both English and Chinese. SwinBERT is an end-to-end video captioning model that integrates VidSwin [18] for feature extraction and a multi-modal transformer encoder for caption generation. VidSwin, a Swin Transformer-based model for video understanding, efficiently captures both spatial and temporal information, making it suitable for tasks such as video classification, object detection, and action recognition. As SwinBERT incorporates VidSwin, it benefits from extracting visual features that differ from those captured by I3D. Furthermore, due to its training on various human actions, SwinBERT can generate meaningful captions for both normal and violent video scenarios. To convert these generated captions into text features, simple contrastive sentence embedding (SimCSE) [14] is employed. SimCSE performs sentence embedding by transforming natural language sentences into fixed-size vector representations, leveraging contrastive learning to position similar sentences closer in the embedding space while separating dissimilar sentences.

Finally, audio features (F_a) are extracted using VGGish [19], an audio feature extraction model developed by Google. VGGish is based on the VGG network architecture and segments audio signals into 1 s intervals, converting them into log-mel spectrograms for input. A log-mel spectrogram represents the frequency content of an audio signal over time, capturing human auditory characteristics by mapping the frequency axis to the mel scale and applying a logarithmic transformation. VGGish is pre-trained on the YouTube-100 M dataset [19], which consists of 70 million training videos, 10 million evaluation videos, and 20 million validation videos. This extensive pre-training makes VGGish a highly effective audio feature extractor for a wide range of sound recognition and classification tasks. In this study, we use VGGish, pre-trained on a large YouTube dataset [20], to extract audio features.

3.2. Feature Fusion

The visual, audio, and text features extracted from the three modalities are fused before being input into the enhanced MTN. We investigate four fusion methods, described below. Since the visual features undergo five-crop augmentation, we apply five-fold tiling to the audio and text features, repeating them to match the number of frames in the visual data.

3.2.1. Product

This method multiplies the visual, audio, and text features element-wise to create a single feature vector. To achieve this, the visual and text features are first transformed to match the dimensions of the audio features using fully connected layers, as shown in Equation (1).

$$F'_v = W_v F_v + b_v \; F'_t = W_t F_t + b_t \tag{1}$$

In Equation (1), W_v and W_t represent the weight matrices for transforming the visual features and text features into the dimensions of the audio features, respectively, while b_v and b_t represent the corresponding biases for each transformation. The transformed features are then combined through an element-wise product, as shown in Equation (2).

$$F_{vta} = F'_v \otimes F'_t \otimes F_a \tag{2}$$

3.2.2. Addition

This method adds the visual, audio, and text features element-wise to create a single feature vector. Similar to the product method, the features are first transformed to the same dimensions before the addition, as shown in Equation (3).

$$F_{vta} = F'_v + F'_t + F_a \tag{3}$$

3.2.3. Concatenation

This method concatenates the visual, audio, and text features along their respective dimensions to create a single feature vector. For example, if the vectors [1–6] are concatenated, the result is [1–6], as represented in Equation (4).

$$F_{vta} = F_v \oplus F_t \oplus F_a \tag{4}$$

In Equation (4), $\oplus$ denotes the concatenation of vectors, and the size of concatenated feature vector F_{vta} is the sum of the sizes of the visual, text, and audio features. This concatenation operation connects the vectors along each feature dimension, creating a single integrated vector incorporating visual, text, and audio features.

3.2.4. Projected Concatenation

In this method, the visual, audio, and text features are first linearly transformed to the same dimensional space before concatenation. After the transformation, the feature vectors are concatenated into a single vector, as shown in Equation (5).

$$F_{vta} = F'_v \oplus F'_t \oplus F'_a \tag{5}$$

3.3. Extraction of Multi-Scale Temporal Features between Video Snippets

The fused feature F_{vta} obtained from the feature fusion process is passed through the enhanced multi-temporal network (E-MTN) to extract multi-scale temporal features $F_{E\text{-}MTN}$ across video snippets. The original MTN comprises pyramid dilated convolution (PDC) blocks [21] and non-local blocks (NLBs) [22]. In this paper, we propose an enhanced MTN model, which incorporates the squeeze and excitation (SE) module [23] and replaces the non-local block with a transformer encoder block (TEB) [24].

Figure 3 illustrates the overall architecture of the enhanced MTN. The pyramid dilated convolution (PDC) [21] is organized in a pyramid structure, typically using three or more

different dilation rates to generate feature maps with varying receptive fields. This design captures multi-scale information at different levels of granularity. The SE module [23], initially introduced to enhance convolutional neural networks (CNNs), dynamically adjusts the importance of each feature channel by learning inter-channel relationships. This mechanism enables the model to focus on the most relevant features, facilitating the extraction of meaningful representations.

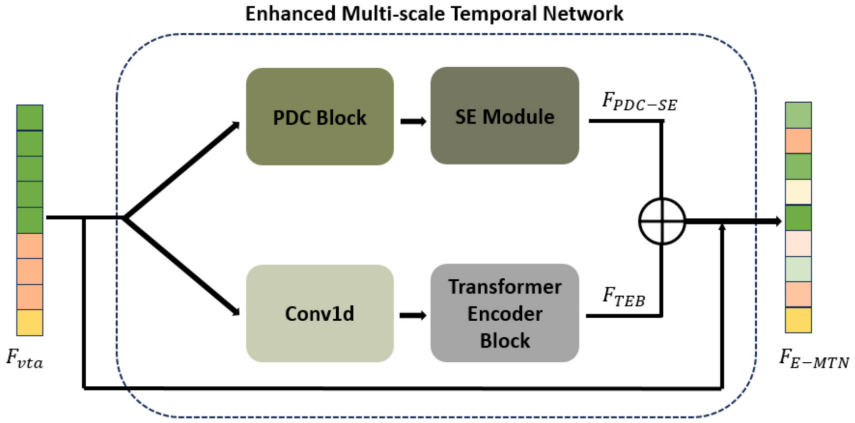

Figure 3. Overall structure of the enhanced multi-scale temporal network.

The SE module consists of two main phases: the squeeze phase and the excitation phase. During the squeeze phase, global average pooling is applied to compress the inter-channel dependency information across the input feature maps, allowing global context extraction for each channel. Following this, the excitation phase employs two fully connected layers and a non-linear activation function to compute the weights that determine the relative importance of each channel, thereby emphasizing significant channels and suppressing less important ones. In this architecture, the PDC captures multi-scale features from video snippets, while the SE module adjusts the features based on channel importance, enhancing the extraction of key information.

The transformer encoder block (TEB) [24], a critical component of the transformer model widely used in natural language processing (NLP) and computer vision tasks, processes input data by learning interactions between words in a sentence or patches in an image. The TEB employs a multi-head self-attention mechanism to capture temporal relationships, facilitating the comprehension of interactions across multiple frames. In the proposed system, the TEB captures global temporal dependencies between video snippets, enhancing the system's ability to model long-range interactions.

The process for extracting multi-temporal scale features between video snippets is as follows. First, the fused feature F_{vta} passes through the PDC block and SE module, resulting in the feature $F_{PDC\text{-}SE}$. Simultaneously, F_{vta} undergoes a 1D convolution and then passes through the TEB, generating the feature F_{TEB}. Next, FPDC-SE and F_{TEB} are concatenated, and the original feature F_{vta} is added (through addition) to produce the final output $F_{\text{E-MTN}}$.

$$F_{E-MTN} = (F_{PDC-SE} \oplus F_{TED}) + F_{vta} \tag{6}$$

The output $F_{E\text{-}MTN}$ is then used to calculate the feature magnitude of the video snippets in the next step. This step helps distinguish between normal and violent videos by reflecting the critical information contained in each snippet.

3.4. Training for Anomaly Detection

The objective of anomaly detection training is to accurately distinguish between normal and violent videos. As observed by Tian et al. [3], violent videos typically exhibit larger feature magnitude values compared to normal ones. To maximize this distinction, two loss functions are utilized.

3.4.1. Feature Magnitude Loss

This loss function is designed to increase the difference in feature magnitudes between normal and violent videos. The feature magnitude is computed as the average of the *L*2 norms of the top *k* snippets with the highest magnitudes, as illustrated in Equation (7).

$$L_{fm=}\begin{cases}\sum_{i=1}^{N}\left(c - F_{fm}\right) \; if \; violence \; video \\ \sum_{i=1}^{N} F_{fm} \; if \; normal \; video\end{cases} \tag{7}$$

In Equation (7), the feature magnitude value, denoted as F_{fm}, is calculated as the average of the *L2* norms of the top k snippets with the largest feature magnitudes out of the T snippets (=$F_{E\text{-}MTN}$) in the video. The *L2* norm is computed by squaring each element of the vector, summing these squares, and then taking the square root of the sum. This metric is useful for measuring the size or similarity of vectors. Since larger differences are amplified, it effectively highlights the distinction in feature magnitudes between normal and violent videos. Here, c is a pre-defined constant, and N represents the number of videos in the training set.

3.4.2. Binary Cross-Entropy Loss

This loss function minimizes the difference between predicted probabilities and actual labels, as expressed in Equation (8). The anomaly score for each snippet is computed through three fully connected layers, and the anomaly score F_s for the entire video is determined by averaging the anomaly scores of the top k snippets.

$$L_{bce} = -\frac{1}{N}\sum_{i=1}^{N} L_{bce} log F_s + (1 - y_i)\log(1 - F_s) \tag{8}$$

The final loss function L is a combination of these two loss terms, weighted by the hyperparameter α, as shown in Equation (9).

$$L = \alpha L_{fm} + L_{bce} \tag{9}$$

4. Experiments and Results

4.1. Datasets and Evaluation Metrics

To assess the performance of the proposed system, experiments were conducted using the XD-Violence dataset [8], which provides multi-modal data. XD-Violence is currently the only large-scale dataset that provides both visual and audio data related to violent incidents. The commonly used UCF-Crime dataset [1] does not include audio signals. While the Violent-Flows [25] and CCTV-Fights [26] datasets contain audio signals, these are often silent or feature only background music, and they also suffer from insufficient data. In contrast, the XD-Violence dataset consists of 4754 unedited videos, totaling 217 h, with both audio signals and weak labels. Additionally, this dataset encompasses a diverse range of scenarios gathered from both films and real-world environments. The videos vary in length and depict a wide array of violent events, such as fights, shootings, explosions, and accidents, which greatly contribute to the dataset's substantial size and overall diversity. For this reason, we evaluated the proposed method using the XD-Violence dataset, excluding all datasets that lacked audio components from the evaluation.

The XD-Violence dataset is explicitly divided into separate training and test sets. The training set comprises 4754 videos, with 2405 labeled as violent and 2349 as non-violent, while the test set consists of 800 videos, 500 of which are labeled as violent and 300 as non-violent.

For performance evaluation, we utilize average precision (AP) [26], as used in previous studies [7–12]. AP is computed by calculating the area under the curve (AUC) of the

precision–recall curve, with higher AP values indicating better performance. Given that the XD-Violence dataset is extremely large in scale, its video content is highly diverse, and it is explicitly divided into separate training and test sets, the application of cross-validation is generally unnecessary. We can independently use the training and test data, and the dataset's size is sufficient to mitigate concerns regarding overfitting or poor generalization. Consequently, performance evaluation can be reliably conducted without the need for cross-validation.

4.2. Implementation Details

4.2.1. Implementation Details of Feature Extraction

Visual features are extracted using I3D, pre-trained on the Kinetics-400 dataset. The frame rate of all videos is fixed at 24 fps, and a sliding window method is used to divide the videos into non-overlapping 16-frame segments.

Text features are generated using captions created by SwinBERT, pre-trained on the VATEX dataset. A sliding window method with a window size of 64 frames is applied, and captions are generated every 16 frames. To extract text features from these captions, sentence embeddings are generated using supervised SimCSE, pre-trained on the BERT model (bert-base-uncased) [27]. The dimensionality of the text features is 768.

Audio features are extracted using VGGish, pre-trained on a large YouTube dataset. The audio is segmented into overlapping 960 ms intervals, and a 96×64 bin log-mel spectrogram is computed for each segment. The dimensionality of the visual features is 1024, while the audio features have a dimensionality of 128.

4.2.2. Implementation Details of Anomaly Detection Training

For comparative experiments on multi-modal approaches, the basic hyperparameter settings are consistent with those used in TEVAD. The number of T snippets for the input video is set to 32, and the dilation parameters for the three-layer pyramid dilated convolution (PDC) are set to 1, 2, and 4, respectively. The number of heads for the multi-head attention in the transformer encoder block is set to 8, and the size of the feed-forward network is configured to 1024. To compute the anomaly score for each snippet, a fully connected layer with 512, 128, and 1 nodes is employed. In the loss function of Equation (7), c is set to 100; k is set to 3; and α in Equation (9) is set to 0.0005.

The proposed model was trained with a batch size of 64 using the Adam optimizer [28], with a learning rate of 0.005 and a weight decay of 0.0005. All experiments were conducted on a single A6000 GPU.

4.3. Experimental Results and Discussion

4.3.1. Performance According to Multi-Modal Extension Methods

In this study, we extended the existing TEVAD framework to propose a violence detection system that integrates three modalities: visual, audio, and text. To evaluate the performance of the proposed method, various extension approaches were tested, and the results were analyzed. For consistency, the early fusion method (before MTN input) was fixed to concatenation, while the late fusion method (after passing through the MTN) was set to addition. Concatenation was chosen as the early fusion method because it is the most basic combination technique, allowing for the creation of an integrated vector while preserving the distinct characteristics of each modality. Experimental results confirmed that concatenation performed well in terms of overall performance. Addition was selected as the late fusion method because it was validated as the best-performing method in the original TEVAD framework.

Table 1 presents the performance results based on the multi-modal extension methods. In [Table 1], the "+" symbol indicates that features were fused before passing through the MTN, while the "/" symbol denotes fusion after passing through different branches of the MTN. The number of branches represents how many paths each modality follows during processing. For example, "V/T" in [Table 1] indicates that the visual and text modalities are

processed independently along separate paths, with fusion occurring after each one passes through the MTN. This approach aligns with the original TEVAD framework. On the other hand, "V + A + T" means that the visual, audio, and text modalities are fused into a single feature before being processed through the MTN. In the case of "V/T/A", the fusion method follows the same process as TEVAD, but with the addition of audio information.

Table 1. Comparison of performance based on multi-modal extension methods.

Multi-Modal Extension	Early Fusion	Number of Branches	Late Fusion	MTN	AP(%)
V/T [7]	-	2	Addition	Original	79.8
V/A/T	-	3	Addition	Original	80.4
				Enhanced	80.8(+0.4)
V/T + A	Concatenation	2	Addition	Original	81.0
				Enhanced	82.4(+1.4)
T/V + A	Concatenation	2	Addition	Original	81.7
				Enhanced	82.8(+1.1)
A/V + T	Concatenation	2	Addition	Original	77.5
				Enhanced	81.7(+4.2)
V + A + T	Concatenation	1	-	Original	82.2
				Enhanced	83.9(+1.7)

The experimental results showed that the "V + A + T" method, where all modalities are fused via concatenation before being passed through a single MTN to extract multi-scale temporal features, achieved the best performance. This is likely because fusing features before extracting multi-scale temporal features preserves the distinct characteristics of each modality while generating an integrated representation. This integrated feature allows for richer information to be utilized as it passes through the MTN, which is crucial for subsequent processing. Furthermore, in all experiments, the proposed enhanced MTN consistently outperformed the original MTN, demonstrating improved performance across various tasks.

4.3.2. Performance According to Early Fusion Methods

In the previous experiments, we confirmed that fusing individual features before inputting them into the MTN contributes to improved performance. To further investigate which fusion method is the most effective, we conducted additional experiments comparing various fusion methods. Table 2 summarizes the performance results based on these fusion methods.

Table 2. Performance comparison based on fusion methods.

Fusion Method	MTN	AP (%)
Product	Original	76.0
	Enhanced	79.4 (+3.4)
Addition	Original	80.5
	Enhanced	81.1 (+0.6)
Projected Concatenation	Original	80.6
	Enhanced	81.5 (+0.9)
Concatenation	Original	82.2
	Enhanced	83.9 (+1.7)

The experimental results demonstrated that the concatenation method achieved the highest performance, with an average precision (AP) of 82.2% when using the original MTN and 83.9% when using the enhanced MTN. Concatenation combines information from each modality while preserving as much of the original data as possible. During the process of extracting multi-scale temporal features via the MTN, an attention mechanism is employed. This suggests that concatenating the features, while retaining the distinct information from each modality and minimizing information loss, is more effective than applying transformations during feature fusion.

The addition and projected concatenation methods exhibited slightly lower performance compared to concatenation. This is likely due to their reliance on simple addition or linear transformation, which may not adequately capture the complex interactions between modalities. The product method resulted in the lowest performance, likely because element-wise multiplication of feature vectors from different modalities can cause the overall vector values to shrink or approach zero when the values of a specific modality are near zero, leading to information loss.

Additionally, in most experiments, the proposed enhanced MTN demonstrated superior performance compared to the original MTN. This improvement can be attributed to the multi-head self-attention mechanism, which is more effective than the traditional attention mechanism in learning complex interactions between features across multiple temporal scales with greater precision and efficiency.

In conclusion, we confirmed that using the concatenation method as the feature fusion technique for each modality effectively preserves and integrates information, resulting in the best performance. This finding supports the idea that retaining as much information from various modalities as possible contributes to enhanced performance in multi-modal learning.

4.3.3. Analysis of the Effectiveness of Multi-Modality

To confirm that multi-modal approaches achieve higher performance compared to single-modal or dual-modal approaches, we conducted six experiments. Three experiments were performed using single-modality input (one each for visual, audio, and text); two experiments used dual-modality input (visual + audio, visual + text); and one experiment incorporated all three modalities (visual + audio + text).

Figure 4 presents the performance results for single modalities (visual, audio, and text), dual modalities combined with visual input, and the multi-modal approach using all three modalities. The experimental results demonstrated that the proposed multi-modal method achieved the highest performance (AP: 83.9%), confirming that the information provided by each modality is complementary, which enhances the accuracy of violence detection.

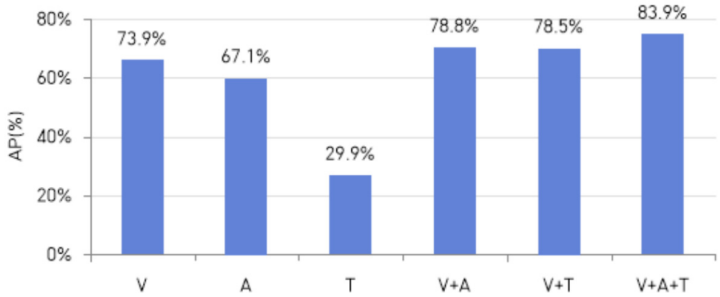

Figure 4. Performance comparison based on different modalities (V: Visual, A: Audio, T: Text).

Among the single modalities, visual information showed the highest performance (AP: 73.9%). While audio information provides valuable cues for violence detection, its effectiveness is somewhat limited due to noise interference. Furthermore, the relatively low performance of the text-only modality is likely attributed to the captioning model being trained on general situations, which may not accurately reflect violent scenarios.

In conclusion, these results demonstrate that the multi-modal approach significantly improves video violence detection performance compared to single- and dual-modality methods. The combination of unique information from each modality, along with their complementary characteristics, contributes to a more accurate and robust violence detection system.

4.3.4. Quantitative Evaluation

Table 3 presents the results of comparing frame-level AP performance on the XD-Violence dataset. To assess the effectiveness of the proposed multi-modal approach, we compare its performance against existing single-modal and dual-modal models. The single-modal model utilizes only visual information, while the dual-modal models combine visual information with either audio or text.

Table 3. Frame-level AP performance on XD-Violence dataset.

Method	Modality	AP(%)
CSL-TAL (2022) [4]	V	71.7
Sultani et al. (2018) [1]	V	75.7
Wu et al. (2021) [5]	V	75.9
Chang et al. (2021) [6]	V	76.9
RTFM (2021) [3]	V	77.8
Wu et al. (2020) [8]	V, A	78.6
TEVAD (2023) [7]	V, T	79.8
Zhang et al. (2023) [10]	V, A	81.4
Pang et al. (2021) [11]	V, A	81.7
UR-DMU (2023) [12]	V, A	81.7
Ours (Original MTN)	V, T, A	82.2
Ours (Enhanced MTN)	V, T, A	83.9

Among the single-modal models, RTFM [3] achieved the highest performance with an AP of 77.8%, demonstrating that high performance can be attained using only visual information. However, single-modal models are limited in complex scenarios because they cannot leverage additional information from audio or text. Among the dual-modal models, the approaches by Zhang et al. [10], Pang et al. [11], and UR-DMU [12] recorded AP performances in the 81% range, highlighting the benefits of combining visual and audio information. These results show that incorporating additional modalities beyond visual data improves detection performance compared to single-modal models.

The proposed multi-modal approach achieved an AP of 83.9% by integrating visual, audio, and text information. This performance surpasses that of existing single-modal and dual-modal approaches, indicating that combining multiple modalities allows for more accurate recognition of violent situations. Notably, these results confirm that effectively integrating the unique information and complementary characteristics of each modality leads to a more robust and precise violence detection system.

4.3.5. Training Time Comparison

To provide a more comprehensive evaluation of the proposed multi-modal approach, we also compared the average training time per iteration across different system configurations. The proposed multi-modal system, which integrates visual, audio, and textual modalities, required an average of 6.54 s per iteration. In comparison, the dual-modality system (visual + text) took 6.20 s per iteration, while the single-modality system (visual only) required 4.28 s per iteration.

Although the multi-modal approach incurs slightly longer training times due to the additional complexity of integrating multiple modalities, this increase is marginal. The significant improvement in accuracy, as demonstrated by the AP scores, more than compensates for the additional computational cost.

5. Conclusions and Future Work

In this study, we proposed a multi-modal video violence detection system that extends TEVAD by incorporating visual, audio, and text information. Through extensive experiments, we confirmed that fusing the features from all modalities into a single integrated feature and processing it through the MTN yielded the best performance. Among the various feature fusion methods, we demonstrated that the concatenation method, which simply connects each feature, proved to be the most effective. Furthermore, we confirmed that the multi-modal approach significantly outperforms both single- and dual-modal approaches. In addition, the proposed enhanced MTN consistently showed improved performance compared to the original MTN across most experiments.

This study demonstrated that effectively utilizing visual, audio, and text modalities can significantly enhance video violence detection performance. However, one limitation of this study is that the captioning model used was primarily trained on general situations, which may lead to inaccurate captions in violent scenarios. As shown in Figure 4, the text-only modality achieved an AP of only 29.9%, indicating its limitations in violence detection. To address this, future work could involve fine-tuning the captioning model on violence-specific datasets, such as the UCA dataset [29]. This would enable the model to better understand and describe violent scenes, ultimately improving the overall performance of text-based violence detection.

Additionally, while the current fusion method proved to be effective, it remains relatively simple and may not fully capture the complex interactions between different modalities. Future research will focus on developing more advanced feature fusion techniques that better account for these interactions. Specifically, addressing the imbalance among modalities is critical for improving both the accuracy and efficiency of the system. Therefore, we plan to explore new fusion strategies to mitigate this imbalance and further enhance the overall performance of video violence detection systems.

Author Contributions: Conceptualization, K.C. and G.N.; software, data curation, G.N.; methodology, formal analysis, validation, investigation, G.N. and K.C.; supervision, project administration, K.C.; writing—original draft preparation, G.N.; writing—review and editing, K.C. and J.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The XD-Violence dataset presented in this study is openly available in the ROC-NG repository at https://roc-ng.github.io/XD-Violence/ (accessed on 27 October 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. 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, Salt Lake City, UT, USA, 18–22 June 2018.
2. 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/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
3. Tian, Y.; Pang, G.; Chen, Y.; Singh, R.; Verjans, J.W.; Carneiro, G. Weakly-Supervised Video Anomaly Detection with Robust Temporal Feature Magnitude Learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021.
4. Panariello, A.; Porrello, A.; Calderara, S.; Cucchiara, R. Consistency-Based Self-Supervised Learning for Temporal Anomaly Localization. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022.
5. Wu, P.; Liu, J. Learning Causal Temporal Relation and Feature Discrimination for Anomaly Detection. *IEEE Trans. Image Process.* **2021**, *30*, 3513–3527. [CrossRef] [PubMed]
6. Chang, S.; Li, Y.; Shen, S.; Feng, J.; Zhou, Z. Contrastive attention for video anomaly detection. *IEEE Trans. Multimed.* **2021**, *24*, 4067–4076. [CrossRef]
7. Chen, W.; Ma, K.T.; Yew, Z.J.; Hur, M.; Khoo, D.A.A. TEVAD: Improved video anomaly detection with captions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023.
8. Wu, P.; Liu, J.; Shi, Y.; Sun, Y.; Shao, F.; Wu, Z.; Yang, Z. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020.

9. Wu, P.; Liu, X.; Liu, J. Weakly supervised audio-visual violence detection. *IEEE Trans. Multimed.* **2022**, *25*, 1674–1685. [CrossRef]
10. Zhang, C.; Li, G.; Qi, Y.; Wang, S.; Qing, L.; Huang, Q.; Yang, M.H. Exploiting completeness and uncertainty of pseudo labels for weakly supervised video anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023.
11. Pang, W.F.; He, Q.H.; Hu, Y.J.; Li, Y.X. Violence detection in videos based on fusing visual and audio information. In Proceedings of the 2021 IEEE International Conference on Acoustics, Speech, and Signal Processing, Toronto, ON, Canada, 6–11 June 2021.
12. Zhou, H.; Yu, J.; Yang, W. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023.
13. Lin, K.; Li, L.; Lin, C.C.; Ahmed, F.; Gan, Z.; Liu, Z.; Wang, L. Swinbert: End-to-end transformers with sparse attention for video captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022.
14. Gao, T.; Yao, X.; Chen, D. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In Proceedings of the EMNLP 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 7–11 November 2021.
15. 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, Honolulu, HI, USA, 21–26 July 2017.
16. Wang, X.; Wu, J.; Chen, J.; Li, L.; Wang, Y.F.; Wang, W.Y. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019.
17. Carreira, J.; Noland, E.; Banki-Horvath, A.; Hillier, C.; Zisserman, A. A short note about kinetics-600. *arXiv* **2018**, arXiv:1808.01340.
18. Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; Hu, H. Video swin transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022.
19. Hershey, S.; Chaudhuri, S.; Ellis, D.P.; Gemmeke, J.F.; Jansen, A.; Moore, R.C.; Wilson, K. CNN architectures for large-scale audio classification. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, New Orleans, LA, USA, 5–9 March 2017.
20. Gemmeke, J.F.; Ellis, D.P.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R.C.; Ritter, M. Audio set: An ontology and human-labeled dataset for audio events. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, New Orleans, LA, USA, 5–9 March 2017.
21. Liu, C.; Xu, X.; Zhang, Y. Temporal attention network for action proposal. In Proceedings of the 2018 25th IEEE International Conference on Image Processing, Athens, Greece, 7–10 October 2018.
22. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
23. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018.
24. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is All You Need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
25. Hassner, T.; Itcher, Y.; Kliper-Gross, O. Violent flows: Real-time detection of violent crowd behavior. In Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, 16–21 June 2012.
26. Perez, M.; Kot, A.C.; Rocha, A. Detection of real-world fights in surveillance videos. In Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing, Brighton, UK, 12–17 May 2019.
27. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the NAACL-HLT 2019, Minneapolis, MN, USA, 2–7 June 2019.
28. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
29. Yuan, T.; Zhang, X.; Liu, K.; Liu, B.; Chen, C.; Jin, J.; Jiao, Z. Towards Surveillance Video-and-Language Understanding: New Dataset Baselines and Challenges. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024.