ABSTRACT: Automatic detection of violent videos has broad application prospects in many fields such as video surveillance and movie grading. However, most existing violent video detection models based on multimodal feature fusion ignore the fact that the audio-visual data in the same violent video may not semantically correspond. Blindly fusing non-corresponding features is not beneficial even potentially harmful to models. In this paper, we propose a novel violent video detection model based on semantic correspondence between audio-visual data from the same video. Deep neural networks are used to extract features of three different modalities: appearance, motion, and audio. After that, we choose the feature-level fusion strategy to fuse these multimodal features via shared subspace learning. Semantic correspondence is used to guide this process through multitask learning and semantic embedding learning. To evaluate the effectiveness of our model, we conduct experiments on several public datasets and our self-built dataset: Violence Correspondence Detection. The results show that our model achieves quite competitive results on both.

INDEX TERMS: Multimodal feature extraction, semantic correspondence, shared subspace learning, violent video detection.

I. INTRODUCTION

Violent videos not only do harm to the harmony and stability of society, but also greatly jeopardize the physical and mental health of teenagers. It is of great significance to detect violence in videos automatically. We take the definition of violent videos proposed in the MediaEval affective task: “violent videos are those one would not let an 8 years old child see because of their physical violence.” [1]. The word “violence” itself is a very subjective semantic concept. The scenes in violent videos are also very complicated, including bloodshed, fighting, war, riots and many other scenes, as shown in Fig. 1. These all make it difficult to automatically detect violent videos.

In the field of violent video detection, appearance and motion features are crucial and complementary. As most videos contain information of both visual and auditory modalities, more and more researchers adapt audio features to assist detection of violent videos. Early works mainly focus on hand-crafted features. The most commonly used appearance descriptors include: scale invariant feature transform (SIFT) [2], histograms of oriented gradients (HOG) [3], etc.
Space-time interest points (STIP) [4] and improved dense trajectories(iDT) [5] are commonly used motion features. As for auditory feature, Mel-scale frequency cepstral coefficient (MFCC) is the most commonly used descriptor. After feature extraction, a classifier is attached to get the final score.

In recent years, some researchers apply deep neural networks to violent video detection, which is used for not only feature extraction but also feature fusion and classification. There are two most common structures in visual feature extraction: two-stream networks based on 2D ConvNet [6] and networks based on 3D ConvNet [7]. Besides, long short-term memory network (LSTM) [8] is used to capture long-term dependency in the time domain sometimes.

However, we find a fact that many existing violent video detection algorithms have ignored: there exists a situation where the semantic information of audio-visual data in the same violent video does not correspond. For example, some video shooters add artistic charm to videos through the contrast of audio-visual semantics, such as playing soothing music in a fighting scene. Although such videos are still classified as violent, there is a huge contradiction between the semantics of different modalities since the visual signal is violent whereas the audio signal is non-violent. Only if multimodal features hold the same semantics, the model can make full use of complementarity between them through fusing. In the above case, direct fusion of multimodal features is not conducive.

Since the audio and visual signals are data of two different modalities, there may exist a problem called heterogeneity gap, which will hinder the full application of multimodal data [9]. The mainstream method to deal with it is shared subspace learning, which aims to embed data of different modalities into an intermediate common space, where the heterogeneity can be regarded as having been eliminated [10]. In this process, the model may implicitly learn some relevant knowledge about correspondence between audio-visual data. However, we believe that explicit introduction of correspondence knowledge in the training process is more conducive to the model, especially when dealing with semantically non-corresponding data. In this paper, we attempt to explicitly introduce correspondence knowledge through multitask learning and semantic embedding learning.

The main contribution of this paper is the proposal of a violent video detection model based on semantic correspondence. The proposed scheme fuses multimodal features through shared subspace learning. Semantic correspondence information is added to further improve the performance. The experimental results show that after semantic correspondence information is added, the performance of our model is further improved on both the public Violent Scene Detection 2015 dataset (VSD2015) [11] and our self-built Violence Correspondence Detection dataset (VCD).

The paper is organized as follows. Section II presents related works in violent video detection. Section III explains our violent video detection methods, including multimodal feature extraction and shared subspace learning with semantic correspondence. The experiment results are shown in Section IV. Finally, Section V concludes the paper and discusses possible future work.

II. RELATED WORKS

There are mainly two ideas for violent video detection: traditional methods based on hand-crafted features and deep learning methods [12].

A. TRADITIONAL METHODS

In some early works, researchers mainly focused on visual information in videos. The extracted visual features (e.g., SIFT, HOG) are sent to an encoder (e.g., Fisher Vector, Bag-of-Words), followed by a classifier (e.g., SVM) to classify whether the input video is violent or not. Some of these works pay a lot of attention to motion feature extraction, and propose many hand-crafted descriptors and their variant. Nievas et al. [13] combined two motion descriptors: STIP and motion scale invariant feature transform (MoSIFT) for fight detection. Gao et al. [14] presented a descriptor called Oriented Violent Flows (OVIF) to extract motion features. A new motion feature descriptor called Optical flow Magnitude and Orientation (HOMO) was devised in [15]. Zhang et al. [16] proposed motion improved weber local descriptor (MoWLD) with dictionary learning.

Later, some researchers began to take advantage of audio features as assistance to visual features. In [17], multimodal features including appearance, motion, and audio features were used to describe the input video, and fused in a late fusion strategy. In general, the performance of the model is improved after audio features are added, because the model has learned more knowledge for violent video detection.

B. DEEP LEARNING METHODS

In recent years, deep learning methods have achieved remarkable results in computer vision. Deep neural networks have also been utilized for violent video detection.

In the Mediaeval affective task 2015, Fudan-Huawei [18] designed a violent video detection system consisting of two-stream networks and LSTM networks. Some conventional motion and audio features were used as complementary information. The system fused appearance, motion, and audio features in a late fusion fashion and achieved the best results that year. Zhou et al. [19] devised a model called FightNet to detect the complicated visual violence interaction based on the classic action recognition model temporal segment networks [20]. Since 3D ConvNet has been widely used in video content understanding, some researchers started to use it for violent video detection. In [21], 3D ConvNet was used to extract spatiotemporal features, whereas Song et al. [22] built an end-to-end violent video detection system based on 3D ConvNet.

Although the above methods have achieved relatively good results, there is still a lot of space left for improving the performance of existing violent video detection methods. In this paper, we propose a violent video detection model
based on semantic correspondence. Multimodal features are extracted using deep learning methods. After that, semantic correspondence information is exploited to guide shared subspace learning in the feature fusion process through two ways: multitask learning and semantic embedding learning. The experimental results demonstrate the effectiveness of our model both qualitatively and quantitatively. It has achieved quite good results on our self-built violent video dataset VCD and exceeds other violent video detection methods by a large margin on the public dataset VSD2015.

III. PROPOSED METHOD

The structure of our model is shown in Fig. 2. First, three kinds of typical features are extracted from the video: appearance, motion, and audio features. Then, we devise a feature fusion baseline based on shared subspace learning to fuse those three features. Finally, the semantic correspondence information is added into the fusion network through multitask learning and semantic embedding learning.

A. MULTIMODAL FEATURE EXTRACTION

Violent videos generally contain the following information: (i) appearance information, such as objects (e.g., firearms, cold arms), scenes (e.g., gory scene, people-lying-down), (ii) motion information, such as human actions or behaviors (e.g., fighting, chasing, shooting), and (iii) audio information. Typical violent videos are often accompanied by audios with screams, explosions, gunshots in. According to the above analysis, we select three typical features to represent violent videos: appearance, motion, and audio features.

1) APPEARANCE FEATURE EXTRACTION

Currently, the methods for appearance feature extraction have been relatively mature. 3D ConvNet is a commonly used frame sequence processing model which could capture spatiotemporal information better than 2D ConvNet. We make use of the pseudo-3D model (P3D) proposed in [23] to extract short-term spatiotemporal features of input video. P3D model consists of pseudo 3D blocks, which are used to replace original 3D ConvNet kernel to simplify calculation. We attach a LSTM network after P3D network to extract long-term features of the video.

As shown in Fig. 2, the input video of length \( N \) is cut into a number of clips \( C \) with a size of \( n \times l \times h \times w \), where \( n \) is the number of channels of each frame, \( l \) is the number of frames, \( h \) and \( w \) are the height and width of the frame, respectively. In appearance feature extraction, \( n \) is equal to 3 for RGB frames. Let \( C^i_k \) be the \( i_{th} \) frame of \( k_{th} \) clip, which is constructed as:

\[
C^i_k = V^{(k-1)l+s+i}, \quad i \in [1, l], k \in [1, \lfloor N/l \rfloor], s \in [0, l/2] \tag{1}
\]

where, \( V \) is the input video, and \( s \) is a random integer which works as a time jittering trick for data augmentation.

All clips are sequentially sent to the P3D network. The output of last average pooling layer in the P3D network is chosen to represent feature of the current clip, which can be seen as the local feature of the input video in temporal domain. All the local features are sent to a LSTM network to get global appearance feature \( F_{ap} \) in time domain.

2) MOTION FEATURE EXTRACTION

The frameworks used for appearance and motion feature extraction respectively are quite similar. The difference lies in that the former operates on video frames, whereas the latter operates on the stacked optical flow displacement fields between consecutive frames. Since optical flow expresses motion information more explicitly than video frames, it is more suitable for motion feature extraction.

Optical flow in horizontal direction \( d_x \) and vertical directions \( d_y \) is extracted respectively based on the method in [6]. The \( i_{th} \) frame of \( k_{th} \) optical flow clip which is denoted
as \( \hat{C}_k^i \) is constructed as:

\[
\hat{C}_k^i = (d_x^{(k-1)i+s+i}, d_y^{(k-1)i+s+i}),
\]

\[i \in [1, l], k \in [1, \lfloor N/l \rfloor], s \in [0, l/2]\] (2)

To capture the motion feature across a sequence of frames, we first stack the corresponding horizontal and vertical optical flow together and then combine the optical flow of each non-overlap \( l \) consecutive frames as motion clips. The number of channels of each frame in motion clips is 2. After the P3D and LSTM networks, we can get the motion feature \( F_{mo} \) of the input video. More details can be found in our previous work [24].

3) AUDIO FEATURE EXTRACTION

As for audio feature extraction, the most commonly used VGGish [25] network is exploited. The VGGish network is modified from the classic VGG network. It outperforms traditional sound process methods, and proves its effectiveness on large voice datasets such as AudioSet [26]. We have made some modifications to the original VGGish network. The last three fully connected layers are replaced with a global average pooling layer to eliminate overfitting.

The audio signal extracted from the video is processed to obtain a mel-spectrogram with size of \( 96 \times 64 \). The spectrogram is then sent to the VGGish network to obtain a 128-dimensional feature \( F_{au} \), which represents the audio feature of the video.

B. FEATURE FUSION BASELINE

Compared to late fusion which operates on the decision-level scores, fusion in feature-level can see more information, thus get better result. Since the audio-visual data are media of two different modalities that may exist heterogeneity gap, shared subspace learning is a widely-used method to deal with it. The key of shared subspace learning is to eliminate the heterogeneity of different modal features through projection transformation, and exploit the complementarity of multimodal features. In the follows, we will introduce our shared subspace learning approach.

The feature fusion baseline is shown in Fig. 3. The appearance feature \( F_{ap} \) and motion feature \( F_{mo} \) are concatenated to be the visual feature \( F_v \). The audio feature \( F_{au} \) and visual feature \( F_v \) are then respectively embedded into a shared subspace. Then, the embedded visual feature \( F_v \) and audio feature \( F_a \) are concatenated to be the ultimate feature of the video \( F \). We can obtain the violence classification score \( p_{cls} \) using two fully connected layers as a classifier. Binary cross entropy loss \( L_{cls}(p_{cls}) = -\log(p_{cls}) \) is used.

In our feature fusion baseline, visual and audio features are projected onto the shared feature space respectively through similar structures, which can be denoted as:

\[
F_v = f_v(F_v) + g_v(F_v)
\] (3)

\[
F_a = f_a(F_{au}) + g_a(F_{au})
\] (4)

where, function \( f_v \) and \( f_a \) are functions to do linear transformation. Function \( g_v \) and \( g_a \) are used to represent group convolution operations.

These two parallel branches are set to learn the subspace with complimentary information and discriminating capacity simultaneously incorporated. The first branch is simply a fully connected layer. Since the fully connected layer has global receptive field, it could capture the connections of different feature points. The other branch is a group convolutional layer. It is exploited to enhance the ability to focus on local feature points. Here we set the number of groups to half of the input feature dimension, so that each convolution kernel only focuses on a very small local area in the input feature. These two parallel branches enable the fusion network to process features from different scales, so the original features can be fully integrated and expressed. Besides, this design increases the width of the network, which can help eliminate overfitting.

C. SEMANTIC CORRESPONDENCE

This section explains how semantic correspondence information is introduced to help optimize shared subspace learning. In addition to the violence classification label, a new label called “correspondence” is added to the datasets we use. This label is used to describe whether the audio-visual data of the same video contain same semantic information. Video data containing blood, weapons, physical violence, etc. are considered to be visual violent. Audio data containing
gunshots, screams, explosions, etc. are considered to be auditory violent. The audio and visual data are labeled separately to prevent interference between each other (e.g., the McGurk effect). If the visual violent label of a video is the same as the audio violent label, the video is considered to have semantic correspondence, otherwise not.

The correspondence label is exploited in two ways to help optimize the violent video classification model: (i) multitask learning, (ii) semantic embedding learning. In the first way, a semantic correspondence classification task is added and a new loss \( L_{crp} \) is introduced. Based on that, the second way introduces another loss \( L_{cos} \) to make the shared subspace semantic retentive. Finally, we minimize a total loss defined as:

\[
L(p_{cls}, p_{crp}, F_A, F_V, y_{crp}) = L_{cls}(p_{cls}) + L_{crp}(p_{crp}) + L_{cos}(F_A, F_V, y_{crp}) \quad (5)
\]

1) MULTITASK LEARNING
The first way to exploit correspondence information comes from hard parameter sharing in multitask learning. As shown in Fig. 2, a correspondence classification task is added to our feature fusion baseline, parallel to violent video classification task. Similar to violent video classification task, two new fully connected layers are attached after the concatenation of the embedded audio and visual features \( F \). Except for the last two fully connected layers in each task, all the remaining layers are shared by these two tasks. A new binary cross entropy loss \( L_{crp}(p_{crp}) = -\log(p_{crp}) \) is added, where \( p_{crp} \) is the predicted probability of the video semantic correspondence.

In our fusion network baseline, only the violent video classification loss \( L_{cls} \) is used to optimize the model. The model may implicitly learn some correspondence-related information. However, we believe that by explicitly introducing correspondence classification loss \( L_{crp} \), the model can better distinguish whether the audio and visual information is semantic complementary. When the visual and audio data in a video are considered to have semantic correspondence, the network will not only focus on the audio and visual features themselves, but also take advantage of the complementarity between them. In the opposite case, the network will pay more attention to the audio and visual features themselves, thus reducing the interference of concatenating features blindly. The experimental results in Section IV confirm that the introduction of correspondence classification task does contribute to the improvement of model performance.

2) SEMANTIC EMBEDDING LEARNING
The second way to introduce semantic correspondence information is semantic embedding learning. We hope the feature embedding in shared subspace is semantically retentive. When audio and visual data have same semantics, the distance between the vectors they embedded in shared subspace should be small. Conversely, the distance between the two vectors should reach at least a set threshold.

We use cosine distance to measure the similarity of vectors the audio and visual features are embedded to. Cosine similarity loss \( L_{cos} \) defined as follow is added to learn semantic embedding:

\[
L_{cos}(F_V, F_A, y_{crp}) = \begin{cases} 
1 - \cos(F_V, F_A), & y_{crp} = 1 \\
\max(0, \cos(F_V, F_A) - \alpha), & y_{crp} = -1
\end{cases}
\]

where, \( y_{crp} \) is the correspondence label of the input video. If the audio-visual data are considered to have semantic correspondence, \( y_{crp} \) is set to 1 otherwise -1. When the audio-visual data do not semantically correspond, \( \alpha \) is the minimum of distance between two vectors in the shared subspace. Through semantic embedding learning, the ability of the network to discriminate semantic correspondence of violent videos can be enhanced, which is more conducive to eliminating interference between non-corresponding features. Besides, the semantic embedding learning could be considered as a form of regularization, which is helpful to enhance the generalization ability of the model and prevent overfitting.

The semantic correspondence information can be regarded as an external knowledge. The purpose of adding it is to learn better embedding for the shared subspace. The semantic correspondence classification task is considered as an indirect way of exploiting it, which does not directly affect the feature embedding process. The network itself learns better embedding that is beneficial to all the tasks. In contrast, semantic embedding learning is a direct way to exploit semantic correspondence information. It measures the quality of the embedding results through a manually designed loss function, the effect of which is direct on the feature embedding. The combination of these two methods takes advantage of both human knowledge and the powerful autonomous learning capabilities of neural networks to make full use of the semantic correspondence information.

IV. EXPERIMENTS
In this part, we introduce the experiments conducted on public and our self-built datasets. We study the effect of three different modal features to the task of violent video detection. Then, we compare our feature fusion baseline with several common-used late fusion approaches. Finally, we explore the impact of semantic correspondence on our violent video detection model.

A. DATASETS
Due to the difficulty of collecting violent data, currently, there lacks of public large-scale violent video datasets. A total of three public violent video datasets are used in our experiments: Hockey Fight [13], Violent Flow [27], and VSD2015. The video duration in these datasets is about 2 to 10 seconds. The scene and semantic information in each video are basically unchanged. In these datasets, violent video detection
task is actually turned into a violent-non-violent binary classification problem.

Hockey Fight: The scene in this dataset is relatively simple. It not only contains one particular violent scene: fight, but also lacks audio data. So, it is only used to validate the effect of appearance and motion features.

Violent Flow: According to our investigation, almost all audio data in this dataset do not contain violent audio events such as explosions and gunshots, which means that almost all audio data are what we call auditory non-violent. The visual violence label of each video in this dataset is basically the same as the video violence label. In this case, the violent labels and semantic correspondence labels in this dataset will show a strong correlation, which makes semantic correspondence information meaningless. Therefore, experiments about semantic correspondence are not conducted on this dataset.

VSD2015: This dataset comes from the Mediaeval2015 Emotional Challenge and is also one of the most widely used violent video datasets. Although it contains a large number of non-violent videos, it also lacks violent videos (252 in development set, 230 in test set). Besides, this will lead to serious data imbalance problems which may do harm to the model convergence in the training process. Therefore, during the training phase, we expand the violent videos in VSD2015 by jittering, mirroring and rotating, etc. Semantic correspondence labels are manually added to the dev sets.

In addition, to solve the problem of insufficient violent videos and further verify our algorithm, we download about a hundred of movies and short videos from YouTube and build our own violent video dataset called VCD which contains 877 violent videos and 832 non-violent videos. Each video is manually labeled with violence and correspondence attributes.

B. IMPLEMENTATION DETAILS

1) FEATURE EXTRACTION
Three typical video features: appearance, motion, and audio are extracted separately using deep learning methods. In appearance feature extraction, we first set all frames extracted from the input video to 224 × 224 which are randomly cropped from the resized 240 × 320 video frames. Each non-overlap 16 frames in the frame sequence are combined to be a clip, which is sent to the P3D199 model pre-trained on the Kenetics-400 [28], to get temporal local feature with 2048-dimensions. All temporal local features are sent to a LSTM network, and the last output of the LSTM network which has 512-dimensions is considered to be the temporal global feature of the input.

As for motion feature extraction, optical flow is extracted as done in [6]. The corresponding horizontal and vertical optical flows are stacked together. And then the optical flows of each non-overlap 16 consecutive frames are combined as motion clips. The network structure used for motion feature extraction is basically the same as that used for appearance feature extraction. The P3D model is pretrained on Kinetics-400 optical flows. A LSTM network is also attached after it.

In order to study the effect of our appearance and motion feature extraction architecture, the following experiments are designed. To demonstrate the effect of the P3D model, a classifier with two fully connected layers is trained to predict violence score for each clip and the video-level prediction score is calculated by averaging all clip-level scores. The label of each clip is inherited from its parent video. Similarly, another classifier with two fully connected layers is also attached to find out the impact of the LSTM network. The above experiments are conducted both on our appearance and motion feature extraction structure.

For audio feature extraction, the ffmpeg tool is used to extract audio files from videos. After short-time Fourier transform and log operation, a mel-spectrogram with size of 96 × 64 is obtained. Next, the spectrogram is sent to the VGGish network pretrained on the AudioSet to extract audio features with 128-dimensions. A classifier consisting of two fully connected layers is also added to verify the effect of audio features. Violent labels for audio files are inherited from video labels.

2) FEATURE FUSION BASELINE

We try to evaluate our feature fusion baseline and compare it with some wide-used late fusion methods. In order to verify the merit of our baseline, three shared subspace learning methods are designed. The first one is a fully connected layer to carry out linear transformation. A group convolutional layer is used as the second subspace learning method. The number of groups is set to half of the input feature dimension, which is 512 for visual feature and 64 for audio feature, so the network will focus more on the local areas of the feature. Then, we combine the fully connected layer and group convolutional layer together in parallel, which is our feature fusion baseline shown in Fig. 3. In the training process of the above models, \( L_{cls} \) is exploited with violent video classification labels as ground truth.

We further compare our feature fusion baseline with some wide-used late fusion methods which fuse the decision scores of multiple modalities. We choose a rule-based late fusion method: decision scores average and three classifier-learning-based methods: naive Bayes, linear kernel SVM, sigmoid kernel SVM.

3) SEMANTIC CORRESPONDENCE

To verify the effectiveness of semantic correspondence, the following experiments based on multitask learning and semantic embedding learning are conducted. Semantic information is only added in the training process.

The first way to exploit semantic correspondence information is correspondence classification. Based on the idea of multitask learning, a correspondence classification task is added to the feature fusion baseline in parallel with the violent video classification task. The structure of the network remains untouched, except for two new fully connected layers.
are added as a classifier. Each task is equipped with its own cross-entropy loss, \( L_{crp} \) for correspondence classification task and \( L_{cls} \) for violent video classification task. The final objective function in the current experiment can be written as:

\[
L_1 = L_{cls} + L_{crp}
\]

The second way is semantic embedding learning. In this experiment, cosine similarity loss \( L_{cos} \) defined in Section III B is introduced to learn semantic embedding. We define another loss:

\[
L_2 = L_{cls} + L_{cos}
\]

In the training process, \( L_1 = L_{cls} + L_{crp} \) is used. After that, \( L_1 \) and \( L_2 \) are used interchangeably in different epoch. This training strategy is adopted because we believe that the model parameters are unstable at the beginning of training process. Introducing cosine similarity loss at this time will make the model pay too much attention to the preservation of semantic information, thus affecting learning of classification characteristics of violence.

4) VISUALIZATION

We try to qualitatively measure the effectiveness of semantic correspondence. The test set of VSD2015 dataset is chosen. Since the number of violent videos is much smaller than that of non-violent videos, we selected all 230 violent videos and randomly selected 230 non-violent videos. The ultimate features \( F \) of these videos are extracted and reduced to 64-dimensions by PCA and further reduced to 2-dimensions by tSNE for visualization.

C. EXPERIMENTAL RESULTS

1) EFFECT OF MULTIMODAL FEATURES

First, the effectiveness of our feature extraction methods for multimodal data is verified on all four used violent video datasets. For each dataset, the model is trained and tested respectively.

The results are shown in Table 1. It can be seen that each single-modal feature contributes to the classification of violent videos. It also proves that using LSTM networks to extract temporal global feature of input video gets better results than directly averaging the clip-level scores, which shows the necessity of the LSTM network.

In general, because the appearance feature contains more information, it performs best when single modality feature is used. The motion feature is second, and the audio feature is the worst. Audio signal itself contains less information, and there is a situation in which visual and audio information does not correspond in violent videos. Since violent labels for audio files are inherited from video labels, many audio files that are classified as violent may not contain violent information. This is also one of the reasons leading to bad performance when using audio feature alone for violent video recognition.

2) FEATURE FUSION BASELINE

As can be seen from the experimental results in Table 2, with \( L_{cls} \) as the objective function, our baseline combining fully connected layer and group convolution layer, achieves the best results on all the datasets we use. This is because our feature fusion baseline takes the advantages of the other two methods and can learn features of different scales.

We also compare our baseline with some popular late fusion methods. As can be seen in Table 3, the performance of our feature fusion baseline is better than all those four late fusion methods we use. The results confirm that compared with late fusion methods in which decision-level scores are utilized, feature fusion methods can see more information and make full use of the complementarity between multimodal features.

Besides, we compare our method with some other existing methods on the Violent Flow dataset. The results are shown in Table 4. Our feature fusion baseline obtains the best performance.

3) EFFECT OF SEMANTIC CORRESPONDENCE

It can be seen in Table 5 that after the semantic correspondence classification task is added, the performance of our model on violent video detection task is improved. It is further increased after semantic embedding learning loss is added. The results prove that the addition of semantic
correspondence does contribute to the model performance on violent video classification task. They also prove that multitask learning and semantic embedding learning are two feasible ways to introduce semantic correspondence information.

We also compare our model with some other methods on the public dataset VSD2015. These methods include traditional methods based on hand-crafted features and methods using deep neural networks. The results are shown in Table 6. Our model outperforms all other methods by a large margin.

4) VISUALIZATION AND ANALYSIS
The visualization of videos in ultimate feature space is displayed in Fig. 4. The left and right subfigures in Fig. 4 respectively show the distribution of selected videos in the ultimate feature space before and after semantic correspondence information is added. It must be pointed out that the composition of VSD2015 dataset is complex. Whether the semantic correspondence information is added or not, the data distribution is quite chaotic.

It can be seen that the data distribution in the middle part of the left subfigure is very confusing, which is not good for violent video classification task. However, after the semantic correspondence is added, as shown on the right subfigure, data of same class have clustered, which is more orderly than the left. This helps build a more efficient high-dimensional classifier and proves the validity of semantic correspondence information from a qualitative perspective.

We also investigate the videos, of which the classification results are rectified after semantic correspondence is added. Fig. 5 visualizes some typical examples. Certain scenes, such as dim scenes, someone lying on the ground, etc., can easily interfere with visual violent judgment. At this point, the audio information will have a great impact on the final classification result.

As shown in Fig. 5 (a), when the video has semantic correspondence, the model will make full use of the complementarity between audio and visual information to correct the misidentified violent video. However, as shown in Fig. 5 (b), when the video does not have semantic correspondence, the model will pay more attention to visual information and
eliminate the interference of non-violent audio information, thus correcting the misjudgment. Fig. 5 (c) shows that the addition of semantic correspondence can also help rectify non-violent video judgment results.

V. CONCLUSION AND FUTURE WORK
In this paper, we propose a violent video detection model based on semantic correspondence. First, three typical video features: appearance, motion, and audio, are extracted and fused based on deep learning methods. We further introduce semantic correspondence information in feature fusion stage to eliminate interference of non-corresponding data. The correspondence information is exploited in two ways to help optimize shared subspace learning: (i) multitask learning, (ii) semantic embedding learning. After semantic correspondence information is exploited in two ways to help eliminate interference of non-corresponding data. The correspondence information is added, the performance of our model has been improved. Its effectiveness has been proved both on the public dataset VSD2015 and our self-built dataset VCD. Considering that most violent videos contain obvious emotional information, we are trying to utilize emotion information in videos to assist violent video detection in future work.

REFERENCES
[1] M. Sjøberg, Y. Basevy, H. Wang, V. L. Quang, B. Jönescu, E. Dellandréa, M. Schell, C.-H. Demarty, and L. Chen, “The mediaEval 2015 affective impact of movies task,” in Proc. MediaEval, Wurzen, Germany, 2015, pp. 1–3.
[2] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” Int. J. Comput. Vis., vol. 60, no. 2, pp. 91–110, Nov. 2004. doi: 10.1023/B:VISI.0000029664.99615.94.
[3] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2005, pp. 886–893.
[4] I. Laptev, “On space-time interest points,” Int. J. Comput. Vis., vol. 64, nos. 2–3, pp. 107–123, Sep. 2005. doi: 10.1007/s11263-005-1838-7.
[5] H. Wang and C. Schmid, “Action recognition with improved trajectories,” in Proc. IEEE Int. Conf. Comput. Vis., Dec. 2013, pp. 3551–3558.
[6] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” in Proc. NIPS, Montreal, QC, Canada, 2014, pp. 568–576.
[7] S. Ji, W. Xu, M. Yang, and K. Yu, “3D convolutional neural networks for human action recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 1, pp. 221–231, Jan. 2013. doi: 10.1109/TPAMI.2012.59.
[8] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Comput., vol. 9, no. 8, pp. 1735–1780, Nov. 1997. doi: 10.1162/neco.1997.9.8.1735.
[9] Y. Peng and J. Qi, “CM-GANs: Cross-modal generative adversarial networks for common representation learning,” ACM Trans. Multimedia Comput. Commun. Appl., vol. 15, no. 1, pp. 1–24, Feb. 2019. doi: 10.1145/3284750.
[10] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. G. Lanckriet, R. Levy, and N. Vasconcelos, “A new approach to cross-modal multimedia retrieval,” in Proc. Int. Conf. Multimedia (MM), Firenze, Italy, 2010, pp. 251–260.
[11] C.-H. Demarty, C. Penet, M. Soleymani, and G. Gravier, “VSD, a public dataset for the detection of violent scenes in movies: Design, annotation, analysis and evaluation,” Multimedia Tools Appl., vol. 74, no. 17, pp. 7379–7404, Sep. 2015, doi: 10.1007/s11042-014-1984-4.
[12] M. Ramzan, A. Abid, H. U. Khan, S. M. Awan, A. Ismail, M. Ahmed, M. Ilyas, and A. Mahmood, “A review on state-of-the-art violence detection techniques,” IEEE Access, vol. 7, pp. 107560–107575, Jul. 2019. doi: 10.1109/ACCESS.2019.2932114.
[13] E. B. Nieves, O. D. Suarez, G. B. García, and R. Sukthankar, “Violence detection in video using computer vision techniques,” in Proc. CAIP, Seville, Spain, 2011, pp. 332–339.
[14] Y. Gao, H. Liu, X. Sun, C. Wang, and Y. Liu, “Violence detection using oriented Violent flows,” Image Vis. Comput., vol. 48–49, pp. 37–41, Apr. 2016, doi: 10.1016/j.imavis.2016.01.006.
[15] J. Mahmoodi and A. Salajeghe, “A classification method based on optical flow for violence detection,” Expert Syst. Appl., vol. 127, pp. 121–127, Aug. 2019, doi: 10.1016/j.eswa.2019.02.032.
[16] T. Zhang, W. Jia, X. He, and J. Yang, “Discriminative dictionary learning with motion weber local descriptor for violence detection,” IEEE Trans. Circuits Syst. Video Technol., vol. 27, no. 3, pp. 696–709, Mar. 2017. doi: 10.1109/TCSVT.2016.2589858.
[17] B. Zhang, Y. Yi, H. Wang, and J. Yu, “MIC-TIU at mediaEval violent scenes detection (VSD) 2014,” in Proc. MediaEval, Barcelona, Spain, 2014, pp. 1–2.
[18] Q. Dai, R. W. Zhao, Z. Wu, X. Wang, Z. Gu, W. Wu, and Y. G. Jiang, “Fuidan-Huawei at mediaeval 2015: Detecting violent scenes and affective impact in movies with deep learning,” in Proc. MediaEval, Wurzen, Germany, 2015, pp. 1–3.
[19] P. Zhou, Q. Ding, H. Luo, and X. Hou, “Violent interaction detection in video based on deep learning,” J. Phys., Conf. Ser., vol. 844, Jun. 2017, Art. no. 012044, doi: 10.1088/1742-6596/844/1/012044.
[20] Y. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, “Temporal segment networks: Towards good practices for deep action recognition,” in Proc. ECCV, Amsterdam, The Netherlands, 2016, pp. 20–36.
[21] F. U. M. Ullah, A. Ullah, K. Muhammad, I. U. Haq, and S. W. Baik, “Violence detection using spatiotemporal features with 3D convolutional neural network,” Sensors, vol. 19, no. 11, p. 2472, May 2019, doi: 10.3390/s19112472.
[22] W. Song, D. Zhang, X. Zhao, J. Yu, R. Zheng, and A. Wang, “A novel violent video detection scheme based on modified 3D convolutional neural networks,” IEEE Access, vol. 7, pp. 39172–39179, Mar. 2019. doi: 10.1109/ACCESS.2019.2906275.
[23] Z. Qiu, T. Yao, and T. Mei, “Learning spatio-temporal representation with pseudo-3D residual networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 5534–5542.
[24] X. Xu, X. Wu, G. Wang, and H. Wang, “Violent video classification based on spatial-temporal cues using deep learning,” in Proc. 11th Int. Sympt. Comput. Intell. Design (ISCID), Hangzhou, China, Dec. 2018, pp. 319–322.
[25] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, “CNN architectures for large-scale audio classification,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), New Orleans, LA, USA, Mar. 2017, pp. 131–135.
[26] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, “Audio set: An ontology and human-labeled dataset for audio events,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), New Orleans, LA, USA, Mar. 2017, pp. 776–780.
[27] T. Hassner, Y. Itcher, and O. Kliper-Gross, “Violent flows: Real-time detection of violent crowd behavior,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops, Providence, RI, USA, Jun. 2012, pp. 1–6.

[28] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, “The kinetics human action video dataset,” 2017, arXiv:1705.06950. [Online]. Available: http://arxiv.org/abs/1705.06950

[29] P. Bilinski and F. Bremond, “Human violence recognition and detection in surveillance videos,” in Proc. 13th IEEE Int. Conf. Adv. Video Signal Based Survell. (AVSS), Denver, CO, USA, Aug. 2016, pp. 30–36.

[30] Y. Yi, H. Wang, B. Zhang, and J. Yu, “MIC-TJU in mediaeval 2015 affect impact of movie task,” in Proc. MediaEval, Wurzen, Germany, 2015, pp. 1–2.

[31] X. Li, Y. Huo, Q. Jin, and J. Xu, “Detecting violence in video using subclasses,” in Proc. ACM Multimedia Conf. (MM), Amsterdam, The Netherlands, 2016, pp. 586–590.

CHAONAN GU received the B.S. degree in digital media technology from the Communication University of China, in 2018, where she is currently pursuing the M.S. degree with the School of Information and Communication Engineering. Her research interest includes video content understanding.

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

SHENGJIN WANG (Senior Member, IEEE) received the B.E. degree from Tsinghua University, Beijing, China, in 1985, and the Ph.D. degree from the Tokyo Institute of Technology, Tokyo, Japan, in 1997.

From May 1997 to August 2003, he was a member of the Senior Research Staff with the Internet System Research Laboratories, NEC Corporation, Nara, Japan. Since September 2003, he has been a Professor with the Department of Electronic Engineering, Tsinghua University, where he is currently the Director of the Research Center for Media Big-data Cognitive Computing. He has published more than 100 articles. He holds more than 20 patents. His current research interests include computer vision, pattern recognition, object detection, advanced driving assistant systems, and deep learning.

* * *