TEVAD: Improved video anomaly detection with captions

Weiling Chen, Keng Teck Ma, Zi Jian Yew, Minhoe Hur, David Aik-Aun Khoo
Hyundai Motor Group Innovation Center in Singapore
2 Bulim Link, Singapore 649674
weiling.chen,kengteck.ma,zijian.yew, david.khoo@hmgics.com

Abstract

Video surveillance systems are used to enhance the public safety and private assets. Automatic anomaly detection is vital in such surveillance systems to reduce the human labor and its associated costs. Previous works only consider spatial-temporal features. In many complex real-world scenarios, such visual features are unable to capture the semantic meanings required to further improve accuracy. To deal with such issues, we propose a novel framework: Text Empowered Video Anomaly Detection (TEVAD) which utilizes both visual and text features. Text features complement the visual features as they are semantically rich. Specifically, we compute text features based on the captions of the videos to capture the semantic meanings of abnormal events and thus improve the overall performance of video anomaly detection. Extensive experiments demonstrate that our proposed framework achieves state-of-the-art results on four benchmark datasets (i.e. ShanghaiTech, UCF-Crime, XD-Violence, and UCSD-Pedestrians) and achieves improved robustness. We further analyze the captions to provide additional explainability for the anomalous videos identified by our proposed algorithm. Our codes are available at https://github.com/coranholmes/TEVAD.

1. Introduction

Video anomaly detection has many practical applications. In manufacturing, it can detect abnormal behavior (e.g. workers tripping) and irregular operations in the production process. In healthcare, intelligent video surveillance systems can reduce the workload of nurses, monitor the conditions of patients and automatically trigger the alarm if an incident occurs to ensure the timely assistance delivered to patients. In public safety domain, anomaly detection can be used to detect illegal behaviors such as fights and shootings to ensure the police officers can be dispatched timely and reduce personal and property losses [2,37].

Despite the wide range of application scenarios, video anomaly detection is a challenging task because such training data are very unbalanced between positive and negative classes, i.e. there are usually fewer positive examples (abnormal events) than negative examples (regular events). In addition, the large diversity of abnormal events mean that the training set often do not contain every possible type of anomalies, hindering the applicability of traditional supervised learning methods for detecting video anomalies. Furthermore, abnormal events in video are vaguely defined due to their ambiguous nature and may cover a wide variety of human activities. Such typical uncertainties of anomalies further complicate the video anomaly detection tasks.

Since video anomaly detection can be used in many scenarios, there have been many attempts on this research topic. Most of the previous models use the spatial-temporal visual features like Temporal Segment Networks (TSN) [55], 3D ConvNet (C3D) [51] or Inflated 3D ConvNet (I3D) [7] to represent the video frames or snippets and perform the video anomaly detection using these visual features.

However, such methods do not consider the high-level semantic meanings of the videos making it difficult to de-
tect certain abnormal events and generalize the models to complex scenarios. Moreover, the actual detection is done based on the anomaly scores generated by the models which are obscure to the front-end surveillance systems users.

On the other hand, video captioning models are trained in a supervised manner using text-video pairs, and learn symbolic representations (words) that are grounded with the visual elements (e.g., people, objects, actions). Recently, through the use of advanced techniques such as transformer [52], the semantically-rich features can be effectively embedded into video captioning models [30, 46, 63]. As a result, such models are able to interpret the input videos with semantically meaningful captions. Such semantic meanings are often absent or extremely difficult to extract solely from the visual features. Inspired by these works, we propose a novel approach to interpret the deep and rich semantic meanings through the use of video-to-text process to improve both accuracy and robustness of weakly supervised video anomaly detection problem.

Specifically, we divide the videos into short snippets and generate the dense captions for these snippets. These features are fused with the visual features to compute the anomaly scores and perform the video anomaly detection. Experimental results show that captions help improve the performance of video anomaly detection. The use of caption has the additional benefit of providing explainability to our model. An example is shown in Figure 1, where high predicted anomaly score of the video snippet is largely due to the “skating” action.

Our contributions of this work are:

• We propose a framework, TEVAD, which exploits both visual and text features for video anomaly detection with different multi-modal fusion methods.

• We extend multi-scale temporal learning to text features to better capture the dependencies between snippet features.

• Our proposed framework outperforms the state-of-the-arts (SOTA) methods on four benchmark datasets and achieves improved robustness.

• We further conduct additional analysis to provide explainability for the anomalous videos identified through the use of a word-masking protocol.

2. Related work

2.1. Image anomaly detection using captions

To the best of our knowledge, we are the first to propose to incorporate captions in video anomaly detection tasks. Nevertheless, a few prior works use captions to perform image anomaly detection. In one of the works [20], the authors use a DenseCap [23] module to generate the regions of interest and their captions. Image based features are extracted using CNN networks on the detected regions. Caption based features are calculated using Word2Vec [36]. Then they concatenate the embeddings and image based features together and perform unsupervised anomaly detection using clustering. Another work [14] exploits more state-of-the-art CLIP [42] model and performs experiments on CIFAR-10 dataset [27]. For experimental setting, they treat one category as abnormal while the others as normal. Their proposed method basically follows the zero-shot classifier described in the original CLIP paper with limited adaptation. However, the assumption that the normal and abnormal category are well defined is not practical in the real-world scenarios.

2.2. Video anomaly detection using visual features

The mainstream methods for anomaly detection in videos can be divided into several categories, depending on the amount of supervision during training.

Earlier efforts focused on the unsupervised learning scenario, where only normal data are available during training. With the emergence of generative models, many approaches proposed such networks to learn the representation of normal data [11, 19, 32, 38, 40, 43, 44, 53, 54]. The basic assumption is that such models only learn the normal representation thus would be unable to reconstruct the abnormal data. However, this assumption does not always hold in many scenarios due to the absence of prior knowledge of abnormal data, resulting in inferior performance. To address this issue, some researchers [3, 17, 64] proposed to generate the pseudo anomalies and perform pseudo-supervised training on the normal and pseudo anomalous data.

Since then, leveraging some abnormal samples have shown more potential compared to unsupervised learning methods. However, frame level annotations on video datasets are especially expensive. Recently, weakly supervised methods has gradually attracted more attention in terms of video anomaly detection tasks. This is because weakly supervised models can be trained on binary video-level labels while being able to predict frame-level labels. Most of the weakly supervised methods [8, 13, 35, 39, 45, 48, 50, 59, 68] are based on Multiple Instance Learning (MIL) framework. These work mainly propose different aggregation functions to process features or anomaly scores so that video-level labels can be used to indirectly supervise instance-level learning. We design our framework which supports the weakly supervised learning as well.

2.3. Video captioning

Video captioning is an important task in video understanding [47]. Several works [12, 21, 42, 49] focused on exploring different 2D/3D video representations to facilitate video captioning tasks. Moreover, many efforts have been made to learn object-level representations [22, 65, 66]...
to further improve the performance of video captioning.

More recently, with the success of transformer models [10, 52] in natural language processing fields, the computer vision community has tried to apply the ideas on different downstream tasks and achieved promising results [5, 18, 28, 29, 33, 67]. Specifically, [30, 46, 63] have proposed end-to-end vision transformer based models to perform video captioning and achieved significantly improved performance.

3. Our method

Figure 2 shows an overview of our Text Empowered Video Anomaly Detection (TEVAD) framework. Given training videos \( \mathcal{V} \), TEVAD first splits each input video \( v \in \mathcal{V} \) into \( T \) snippets. Afterwards, two separate branches extract visual and text features for each snippet in parallel. The text branch generates dense captions (Section 3.1.1) before transforming them to sentence embeddings (Section 3.1.2), while the visual branch extracts visual I3D [7] features. Multi-scale temporal networks are included in both branches to better capture multi-scale temporal dependencies (Section 3.3). The resulting multi-scale visual \( F_{\text{vis}} \in \mathbb{R}^{d_{\text{vis}}} \) and text features \( F_{\text{txt}} \in \mathbb{R}^{d_{\text{txt}}} \) are fused together (Section 3.4), and used to calculate the feature magnitude of snippets. Top-K largest feature magnitudes from normal and abnormal videos are passed to train a binary snippet classifier. During interference phase, the trained snippet classifier is able to predict the snippet level predictions which are propagated to the individual frames within each snippet to obtain frame level predictions (Section 3.5).

3.1. Generating text features for videos

3.1.1 Generating dense captions for videos

Although there are some research works [26, 56] featuring generating dense captions for videos, the performance of such models is often not satisfying enough compared to single caption generation models. Particularly, it is challenging for dense caption models to determine the number of “important events” in the video sequences which is essential in video anomaly detection.

In view of this, we propose to use single caption models to generate captions needed for producing text features. To fuse the text features with visual features in the next step, a caption needs to be generated for each snippet. However, each snippet usually only includes too few frames for generating meaningful video captions. To circumvent this problem, we employ a sliding window strategy and compute the caption for a consecutive 64 frames for every 16 frame. Although this sliding window strategy results in redundant information being encoded, it has the advantage of minimizing information loss and preserving important events.

In this work, we use one of the state-of-the-art video captioning model SwinBERT [30] to generate the descriptions of video snippets. Apart from the performance, another reason we choose SwinBERT is that it uses a Video Swin Transformer (VidSwin) [34] to extract visual features instead of I3D features used in the visual branch of TEVAD. The different network architectures encourages the learning of different representations so as to improve the anomaly detection performance.

To generate the captions, we use pre-trained models on several different video captioning datasets (i.e., MSVD [9], VATEX [58], TVC [29]) instead of training on datasets used for experiments described in Section 4. This is because the anomaly detection datasets do not contain the necessary captions to train the captioning model. As a result, the captions do not always reflect the video contents accurately. Despite this, as we show in the results in Section 4, these inaccurate captions are still highly beneficial for anomaly detection.
3.1.2 Generating sentence embeddings for videos

To compute the text features from generated video captions, we use SimCSE [15] to generate sentence embeddings. SimCSE is a framework using contrastive learning methods to learn sentence embeddings by using dropout noises and incorporating annotated pairs from natural language inference datasets. It uses “entailment” pairs as positives and “contradiction” pairs as hard negatives to train the framework and achieve good results.

Notably, the proposed TEVAD framework is quite flexible in terms of each individual component and SimCSE can be replaced by other state-of-the-art sentence embedding models with minimum adaptations.

3.2. Generating visual features for videos

In this work, we extract I3D [7] features using a ResNet-50 [21] as backbone. Following previous works [13, 50], we perform ten-crop or five-crop augmentation on datasets to obtain better performance. For five-crop, we crop the given frame into four corners and the central crop. For ten-crop, we further include the horizontal flipped version of five-crop.

C3D, TSN or other feature extractors can also be used to replace the I3D feature extractor used in the proposed framework. Previous experiments [8, 50] show that I3D achieves the best performance among other feature extractors for similar tasks, thus we use I3D features for the following experiments.

3.3. Multi-scale temporal feature learning

Multi-scale Temporal Network (MTN) was firstly proposed in [50] to capture the long and short range temporal dependencies between visual features of snippets. In this work, we extend MTN to process the text features and then fuse them with visual features. The performance improves significantly after adding MTN to process text features (see Section 4).

Similar to the visual MTN, the text MTN also includes a 3-layer pyramid dilated convolutions (PDC) [31] block and a non-local block (NLB) [57]. The PDC over time span is used to learn multi-scale representation of video snippets while the NLB is used to learn the global temporal dependencies between video snippets. More details are introduced in Section A of the supplementary materials.

The outputs from the two blocks are concatenated and added to the original features to produce the final output of text MTN denoted as \( \overline{\mathbf{F}}_{\text{txt}} = f_{\text{MTN}}(\overline{\mathbf{F}}_{\text{txt}}; \theta) \), where \( \overline{\mathbf{F}}_{\text{txt}} \in \mathbb{R}^{d_{\text{txt}}} \) and \( \theta \) comprises the weights for all convolution functions described in this section. Both visual and text features go through the similar process thus we have \( \overline{\mathbf{F}}_{\text{vis}} = f_{\text{MTN}}(\overline{\mathbf{F}}_{\text{vis}}; \theta) \), where \( \overline{\mathbf{F}}_{\text{vis}} \in \mathbb{R}^{d_{\text{vis}}} \). By applying MTN to process both visual and text features, TEVAD is able to learn the temporal dependencies between video snippets in both modalities.

3.4. Multi-modal feature fusion

After obtaining the output from MTN, we employ the late fusion scheme [4] to fuse the features together. We investigate three different fusion methods: concatenation, addition and product. Since visual features are five/ten-cropped, the text features are tiled for five/ten times to be consistent with visual features.

(a) concatenation: We direct concatenate \( \overline{\mathbf{F}}_{\text{vis}} \) and \( \overline{\mathbf{F}}_{\text{txt}} \) given by: \( \mathbf{X} = \{ \overline{\mathbf{F}}_{\text{vis}} \overline{\mathbf{F}}_{\text{txt}} \} \) where \( \mathbf{X} \in \mathbb{R}^{d_{\text{vis}}+d_{\text{txt}}} \).

(b) addition: We employ an element-wise addition between the visual and text embedding features. However, since \( d_{\text{vis}} > d_{\text{txt}} \), we add a fully connected layer to reduce the dimension of visual features to the same as the text features and fuse the two by \( \mathbf{X} = f_{\text{FC}}(\overline{\mathbf{F}}_{\text{vis}}; \delta) + \overline{\mathbf{F}}_{\text{txt}} \), where \( \mathbf{X} \in \mathbb{R}^{d_{\text{vis}}} \) and \( \delta \) comprises all the weights of the full connected layers described in this section.

(b) product: We employ a Hadamard product between the visual and text embedding features. Similar to addition, a fully connected layer is added to reduce the dimension of visual features and the fused features are calculated by \( \mathbf{X} = f_{\text{FC}}(\overline{\mathbf{F}}_{\text{vis}}; \delta) \odot \overline{\mathbf{F}}_{\text{txt}} \), where \( \mathbf{X} \in \mathbb{R}^{d_{\text{vis}}} \).

Overall, we use \( \mathbf{X} = f_{\text{fuse}}(\overline{\mathbf{F}}_{\text{vis}}; \overline{\mathbf{F}}_{\text{txt}}; \delta) \) to denote the fused features in the following sections. Three fully connected layers are added to calculate the anomaly scores given by \( s = f_{\text{pred}}(\mathbf{X}; \delta) \). Additionally, \( \mathbf{S} = \{ s_i \}^T \) denotes the anomaly scores of snippets in one video \( v = \{ \mathbf{X}_i \}^T \).

3.5. Model training

During the training phase, the model only has access to video level labels. According to [50], abnormal snippets have larger feature magnitude than normal ones. We follow the same work and use \( l_2 \) norm to compute the feature magnitude. \( \text{topK}(v; k) \) is used to denote such a subset which includes \( k \) snippets with the highest magnitude among the \( T \) snippets in a video. The feature magnitude of a video \( v \) is computed as:

\[
f_{FM}(v; k) = \frac{1}{k} \sum_{k \in \text{topK}(v; k)} \| \mathbf{X}_i \|_2
g(1)
\]

The purpose of training is to maximise the difference between the anomaly score of normal videos and abnormal videos. Thus the total training loss of the normal and abnormal videos in one batch are denoted as:

\[
\mathcal{L}_{fm} = \begin{cases} 
\sum_{j=1}^{V} (c - f_{FM}(v_j; k)), & \text{if } y_j = 1 \\
\sum_{j=1}^{V} f_{FM}(v_j; k), & \text{if } y_j = 0 
\end{cases}
g(2)
\]
where $c$ is a pre-defined constant and $|V|$ is the number of videos in the training set.

Similarly, the average of the selected $k$ snippets’ anomaly scores is calculated to represent the anomaly score of the whole video as:

$$f_s(v; k) = \frac{1}{k} \sum_{X_i \in \text{topK}(v;k)} f_{pred}(X_i; \delta) \quad (3)$$

For the actual anomaly detection, we train a simple binary classifier by using a binary cross entropy loss:

$$\mathcal{L}_{bce} = -\frac{1}{|V|} \sum_{j=1}^{|V|} (y_j \log(f_s(v_j; k))) + (1 - y_j) \log(1 - f_s(v_j; k))) \quad (4)$$

Overall, the loss function is given as below where $\alpha$ is a hyper-parameter to adjust the weights of the loss components.

$$\mathcal{L} = \alpha \mathcal{L}_{fm} + \mathcal{L}_{bce} \quad (5)$$

### 4. Experimental results

#### 4.1. Datasets and evaluation metrics

We present the results of TEVAD on four different datasets, namely UCSD Ped2 [62], ShanghaiTech [32], UCF-Crime [48], and XD-Violence [61]. Among the four datasets, UCF-Crime and XD-Violence are designed for the weakly supervised video anomaly detection task while the other two are originally designed for unsupervised or semi-supervised video anomaly detection tasks. More detailed introduction of these datasets are provided in Section B of the supplementary materials.

To evaluate the performance of TEVAD, we consider Area Under the ROC curve (AUC) which is widely used for evaluation in video anomaly detection fields. We adopt the micro-averaged AUC by concatenating all frames then computing the AUC scores on UCF-Crime, ShanghaiTech and UCSD Ped2 datasets. For XD-Violence, since most of the previous work used Average Precision (AP), we use it as the evaluation metric to make the results comparable. Similarly, we adopt the micro-averaged AP by concatenating all frames.

#### 4.2. Implementation details

**Visual Feature extraction** Given a video, we split it into non-overlapping 16-frame snippets. For UCF-Crime, ShanghaiTech and UCSD Ped2 datasets, we use an 3D feature extractor with a ResNet50 backbone pretrained on Kinetic-400 [24] to extract the visual features of snippets with a dimension $d_{vis} = 2048$ from Mixed-5c layer. We use the 3D features provided by the author of XD-Violence directly with $d_{vis} = 1024$.

**Text feature extraction** We use the default setting for SwinBERT pretrained on VATEX dataset [58] to generate captions. As described in Section 3, the caption of each snippet is generated based on the current and the following three snippets with a total number of 64 frames. To extract the sentence embeddings of the captions, we use the default setting of supervised SimCSE pretrained on bert-base-uncased. The dimension of text features for each snippet is $d_{txt} = 768$.

**Multi-scale temporal feature learning** For 3-layer pyramid dilated convolutions in MTN, we set the dilation parameter as 1, 2, 4 respectively following [50]. We set $\alpha = 0.0001$ in Equation (5).

**Training details** We train our model on a single V100 GPU using Pytorch [41]. The model is trained with a batch size of 64 using an Adam [25] optimiser with a learning rate of 0.001 and weight decay of 0.005.

#### 4.3. Results on benchmark datasets

We divide previous models or frameworks for video anomaly detection into supervised and unsupervised methods and show the results from Tabs. 1 to 4. For comparisons, we use the published results of other methods.

<table>
<thead>
<tr>
<th>Type</th>
<th>Source</th>
<th>Method</th>
<th>AUC (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CVPR’18</td>
<td>Liu et al. [32]</td>
<td>95.4</td>
</tr>
<tr>
<td></td>
<td>WACV’22</td>
<td>FastAno [40]</td>
<td>96.3</td>
</tr>
<tr>
<td></td>
<td>CVPR’21</td>
<td>SSMTL [16]</td>
<td>97.5</td>
</tr>
<tr>
<td></td>
<td>TPAMI’21</td>
<td>Georgescu et al. [17]</td>
<td>98.7</td>
</tr>
<tr>
<td></td>
<td>CVPR’19</td>
<td>GCN-Anomaly [68]</td>
<td>93.2</td>
</tr>
<tr>
<td></td>
<td>CVPR’18</td>
<td>Sultani et al. [48]</td>
<td>92.3</td>
</tr>
<tr>
<td></td>
<td>ICCV’21</td>
<td>RTFM [50]</td>
<td>98.6</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>TEVAD</td>
<td>98.7</td>
</tr>
</tbody>
</table>

Table 1. Frame-level AUC results on UCSD Ped2 dataset.

**Results on UCSD Ped2**: The frame-level micro AUC results on UCSD Ped2 dataset are presented in Tab. 1. This dataset is relatively old and small-scaled thus over studied. Nevertheless, our proposed model still performs best compared to the SOTA unsupervised and supervised methods.

**Results on ShanghaiTech**: The frame-level micro AUC results on ShanghaiTech dataset are presented in Tab. 2. This dataset has been well studied but our proposed framework managed to outperform the SOTA unsupervised methods and supervised methods by a minimum of 14.9% and 1.2% respectively. [17] achieves similar performance as ours on this dataset but much worse on UCF-Crime dataset which indicates that their method can perform well on detecting anomalies in daily settings but is not adaptive in terms of detecting rarer anomalies like crime related events.
Table 2. Frame-level AUC results on ShanghaiTech dataset.

<table>
<thead>
<tr>
<th>Type</th>
<th>Source</th>
<th>Method</th>
<th>AUC (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TPAMI’21</td>
<td>Georgescu et al. [17]</td>
<td>82.7</td>
</tr>
<tr>
<td></td>
<td>CVPR’22</td>
<td>SSPCAB [43]</td>
<td>83.6</td>
</tr>
<tr>
<td></td>
<td>CVPR’22</td>
<td>SSMTL [1]</td>
<td>83.7</td>
</tr>
<tr>
<td></td>
<td>CVPR 2019</td>
<td>GCN-Anomaly [68]</td>
<td>84.4</td>
</tr>
<tr>
<td>Sup</td>
<td>IEEE Trans Multimedia’21</td>
<td>AR-Net [53]</td>
<td>91.2</td>
</tr>
<tr>
<td></td>
<td>CVPR’21</td>
<td>MIST [13]</td>
<td>94.8</td>
</tr>
<tr>
<td></td>
<td>CVPR’22</td>
<td>BN-SVP [45]</td>
<td>96.0</td>
</tr>
<tr>
<td></td>
<td>ICCV’21</td>
<td>RTFM [50]</td>
<td>97.2</td>
</tr>
<tr>
<td></td>
<td>TIP’21</td>
<td>Wu et al. [59]</td>
<td>97.5</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>TEVAD</td>
<td>98.1</td>
</tr>
</tbody>
</table>

Table 3. Frame-level AUC results on UCF-Crime dataset.

<table>
<thead>
<tr>
<th>Type</th>
<th>Source</th>
<th>Method</th>
<th>AUC (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unsup</td>
<td>ICCV’19</td>
<td>BODS [54]</td>
<td>68.3</td>
</tr>
<tr>
<td></td>
<td>ICCV’19</td>
<td>GODS [54]</td>
<td>70.5</td>
</tr>
<tr>
<td></td>
<td>Patter Recog’20</td>
<td>FSCN [60]</td>
<td>70.6</td>
</tr>
<tr>
<td>Sup</td>
<td>CVPR’18</td>
<td>Sultani et al. [48]</td>
<td>75.4</td>
</tr>
<tr>
<td></td>
<td>CVPR’19</td>
<td>GCN-Anomaly [68]</td>
<td>82.1</td>
</tr>
<tr>
<td></td>
<td>CVPR’21</td>
<td>MIST [13]</td>
<td>82.3</td>
</tr>
<tr>
<td></td>
<td>CVPR’22</td>
<td>BN-SVP [45]</td>
<td>83.4</td>
</tr>
<tr>
<td></td>
<td>ICCV’21</td>
<td>RTFM [50]</td>
<td>84.3</td>
</tr>
<tr>
<td></td>
<td>ICCV Trans Multimedia’21</td>
<td>Chang et al. [8]</td>
<td>84.6</td>
</tr>
<tr>
<td></td>
<td>TIP’21</td>
<td>Wu et al. [59]</td>
<td>84.9</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>TEVAD</td>
<td>84.9</td>
</tr>
</tbody>
</table>

Table 4. Frame-level AP results on XD-Violence dataset.

<table>
<thead>
<tr>
<th>Type</th>
<th>Source</th>
<th>Method</th>
<th>AP (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sup</td>
<td>arXiv’22</td>
<td>CSL-TAL [39]</td>
<td>71.7</td>
</tr>
<tr>
<td></td>
<td>CVPR’18</td>
<td>Sultani et al. [48]</td>
<td>75.7</td>
</tr>
<tr>
<td></td>
<td>TIP’21</td>
<td>Wu et al. [59]</td>
<td>75.9</td>
</tr>
<tr>
<td></td>
<td>ICCV’21</td>
<td>RTFM [50]</td>
<td>77.8</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>TEVAD</td>
<td>79.8</td>
</tr>
</tbody>
</table>

4.4. Ablation studies

4.4.1 Effectiveness of main components

We perform an ablation study on different datasets to demonstrate the effectiveness of the main components in TEVAD and the results are shown in percentage format in Tab. 5. To be consistent, we show the AUC results for UCSD Ped2, ShanghaiTech and UCF-Crime dataset and AP results for XD-Violence dataset. It can be observed from the table that all four datasets show a consistent improvement in performance by adding text features. In addition, the performance can be further boosted if the text features are processed using MTN. To sum up, TEVAD’s performance increases by 14.88%, 3.93%, 1.8% and 2.82% on UCSD Ped2, ShanghaiTech, UCF-Crime and XD-Violence datasets respectively compared to using visual features alone.

4.4.2 Impact of captions quality

Since the anomaly detection datasets do not contain the necessary captions to train the captioning model, we use the pre-trained models trained on other video captioning datasets. To understand the impact of different pretrained models (i.e., caption quality), we perform additional experiments on UCF-Crime dataset as it is the most challenging. It can be observed from Tab. 6 that VATEX pre-trained models perform better than the other two. These results are intuitive as MSVD [9] is a relatively small video captioning dataset and does not contain enough crime or violence related video content. In addition, although TVC [29] is relatively large, videos in this dataset are collected from TV programs and are significantly different from the surveillance contexts in crime dataset. On the other hand, VATEX contains a large number of videos covering 600 human activities which follows the Kinetics-600 [6] taxonomy. Hu-
Table 5. Ablation study results.

<table>
<thead>
<tr>
<th>Visual</th>
<th>Text</th>
<th>Fusion</th>
<th>Ped2 (%)</th>
<th>Shanghai (%)</th>
<th>Crime (%)</th>
<th>Violence (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>83.81</td>
<td>94.17</td>
<td>83.1</td>
<td>76.94</td>
</tr>
<tr>
<td>✓</td>
<td>Vanilla</td>
<td>concat</td>
<td>93.17</td>
<td>97.85</td>
<td>83.18</td>
<td>77.91</td>
</tr>
<tr>
<td>✓</td>
<td>MTN</td>
<td>concat</td>
<td>96.71</td>
<td>97.86</td>
<td>84.9</td>
<td>79.3</td>
</tr>
<tr>
<td>✓</td>
<td>MTN</td>
<td>add</td>
<td>98.69</td>
<td>98.1</td>
<td>84.13</td>
<td>79.76</td>
</tr>
<tr>
<td>✓</td>
<td>MTN</td>
<td>product</td>
<td>94.12</td>
<td>97.2</td>
<td>83.83</td>
<td>78.49</td>
</tr>
</tbody>
</table>

Table 6. Experimental results using different SwinBERT pre-trained models.

<table>
<thead>
<tr>
<th>Fusion</th>
<th>Pre-trained</th>
<th>AUC (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>add</td>
<td>MSVD</td>
<td>82.9</td>
</tr>
<tr>
<td>concat</td>
<td>MSVD</td>
<td>83.8</td>
</tr>
<tr>
<td>add</td>
<td>TVC</td>
<td>82.3</td>
</tr>
<tr>
<td>concat</td>
<td>TVC</td>
<td>82.6</td>
</tr>
<tr>
<td>add</td>
<td>VATEX</td>
<td>84.1</td>
</tr>
<tr>
<td>concat</td>
<td>VATEX</td>
<td>84.9</td>
</tr>
</tbody>
</table>

Figure 3. Example results from (a) ShanghaiTech (riding a bike), (b) XD-Violence (riot) , and (c) UCF-Crime (vandalism) datasets. The top row shows predicted anomaly scores and the groundtruth labels. For frames labeled with green or red arrows, we also show the image frames and their associated generated captions in the bottom row.

4.5. Robustness comparisons

Another advantage of TEVAD is that it is more robust by considering both visual and text modalities. We run 1,000 epochs for both RTFM and our method and evaluate every 5 epochs after training for 50 epochs. The standard deviations of AUC/AP are presented in Tab. 7.

It can be concluded from the experimental results that multi-modality features help improve the robustness of the model. TEVAD shows a more robust results on Ped2, ShanghaiTech and Crime datasets when the text features are added. In addition, the framework achieves the lowest standard deviation in terms of AUC/AP on all four datasets when MTN is applied to process the text features.

4.6. Qualitative analysis

We provide some qualitative results from different datasets in Figure 3. In terms of anomaly scores, our TEVAD can effectively predict a small score for normal snippets and a large score for abnormal snippets regardless of the different background scenes and the types of abnormal events. Additionally, our model is able to detect multiple abnormal events (Figure 3 (c)), which makes it applicable to real-world scenarios. Moreover, the margins between normal and abnormal snippets are relatively clear.

In terms of the usability (i.e. quality of generated captions), TEVAD works well on ShanghaiTech dataset which manly contains day to day activities and can effectively capture the main abnormal event like "riding bikes" (Figure 3 (a)). Figure 3 (b) and (c) present more challenging videos...
Table 7. Robustness of using both modality features.

<table>
<thead>
<tr>
<th>Visual</th>
<th>Text</th>
<th>Fusion</th>
<th>Ped2 (%)</th>
<th>Shanghai (%)</th>
<th>Crime (%)</th>
<th>Violence (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>×</td>
<td></td>
<td>14.77</td>
<td>3.18</td>
<td>1.98</td>
<td>4.63</td>
</tr>
<tr>
<td>✓</td>
<td>Vanilla concat</td>
<td></td>
<td>7.4</td>
<td>1.62</td>
<td>1.86</td>
<td>4.96</td>
</tr>
<tr>
<td>✓</td>
<td>MTN concat</td>
<td></td>
<td><strong>3.43</strong></td>
<td><strong>1.33</strong></td>
<td>1.75</td>
<td>6.92</td>
</tr>
<tr>
<td>✓</td>
<td>MTN add</td>
<td></td>
<td>5.83</td>
<td>1.61</td>
<td><strong>1.48</strong></td>
<td><strong>4.27</strong></td>
</tr>
<tr>
<td>✓</td>
<td>MTN product</td>
<td></td>
<td>4.62</td>
<td>2.09</td>
<td>1.62</td>
<td>4.67</td>
</tr>
</tbody>
</table>

Figure 4. Example results from (a) ShanghaiTech (riding a bike), (c) XD-Violence (shooting), and (b) UCF-Crime (arrest) datasets showing the contribution of each word in the caption to the snippet anomaly score. An image frame of the abnormal event from the snippet is also shown on the right of each caption.

which includes the abnormal event of “riot” and “vandalism” respectively. Notably, though the VATEX dataset used for training the captioning models does not explicitly include such activities, the generated captions capture the similar semantic meaning in the embedding space. For example, “a large crowd of people are gathered” is possibly related to riot while “throws it to the camera” indicates potential vandalism.

4.7. Explainability analysis

Although the generated captions may not be completely accurate in some cases, we conduct additional analysis to demonstrate the explainability of incorporating captions for video anomaly detection tasks. During the inference phase, we iteratively mask each word in the caption of the snippet and calculate the sentence embeddings (i.e., text features) based on the masked captions. The text features are then fused with the visual features and fed into the trained model to predict the anomaly scores for each snippet of the video.

Figure 4 shows the explainability results to understand the contribution of each word in captions of the snippets. The score above each word in the caption is the difference between the anomaly score by masking this word and the original anomaly score without masking. Therefore, a higher score indicates a higher contribution to the predicted anomaly score.

Figure 4 (a) shows the caption and an image of a video snippet from ShanghaiTech dataset. This snippet contains an abnormal event of “riding a bicycle”. Consequently, the word “bikes” contributes the most for identifying this anomalous event comparing to other words in the caption. Similarly in Figure 4 (b), the word “gun” contributes most for identifying the “shooting” scene in this snippet. On the other hand, Figure 4 (c) shows an inaccurate caption for a snippet related to an “arrest” scene from crime dataset. Regardless of the inaccuracy of the caption, the word “fall” which is possibly related to the “arrest” action contributes significantly for identifying the anomalous event.

The observations described in this section and previous Section 4.6 provides the insights that the performance of TEVAD framework can potentially be further improved if some captions of the video anomaly detection datasets are available.

5. Conclusions

Video anomaly detection is a critical yet challenging task in many real-world scenarios. Most of previous works only consider using spatial-temporal visual features to perform video anomaly detection and fail to capture the semantic meaning of complex anomalies in real world contexts. In this work, we have proposed a weakly supervised framework called TEVAD which uses both visual and text modality features to perform video anomaly detection tasks. We extend MTN to process sentence embeddings of captions to learn the dependencies between snippets and further improve the performance. In addition, the generated captions provide explainable results to the surveillance end users. Our proposed TEVAD framework achieves SOTA performance on four different benchmark datasets.
References


[47] Vijeta Sharma, Manjari Gupta, Ajai Kumar, and Deepti Mishra. Video processing using deep learning techniques: A systematic literature review. IEEE Access, 2021. 2


