DETECTING ANOMALIES IN VIDEOS USING MEMORY-AUGMENTED AUTOENCODER WITH KEY FRAME SELECTION

Anh Le[1], *Quang Uy Nguyen*[1,*], *Thi Huong Chu*[1], *Hai-Hong Phan*[1]

Abstract

In this article, we propose a novel method to train a memory-augmented autoencoder in supervised mode by generating pseudo abnormal videos based on key frame selection techniques. Most video anomaly detection methods employ a machine learning model to learn patterns of normal videos. Any video where the patterns significantly deviate from the learnt patterns is considered an anomaly. However, developing an effective machine learning model for video anomaly detection is a challenging task due to the deficiency of anomalies. Specifically, abnormal samples are often much rarer and harder to collect than normal samples. To address this problem, we propose a novel approach using key frame selection techniques to generate pseudo anomalies. The generated pseudo anomalies are then combined with normal data to create the augmented dataset. The memory-augmented autoencoder is then trained on the augmented datasets. The experimental results show that the AUC scores of the proposed solution are higher than those of the base network architecture from 0.20% to 1.31% on three well-known datasets for video anomaly detection.

Index terms

Video anomaly detection, autoencoder, pseudo anomaly generator, key frame selection.

1. Introduction

Security surveillance cameras are often used to monitor public activities by recording and analyzing videos of observed areas. These systems are increasingly deployed in public areas to assist the government in social activities management. Videos captured by security cameras are typically processed using artificial intelligence software, aiding administrators in quickly identifying law violations [1]. In such systems, rapidly detecting anomalous objects and unusual behaviors plays a crucial role in safeguarding public environments. Subsequently, anomaly detection in videos has garnered significant attention from the research community [1]–[3].

[1]Institute of Information and Communication Technology, Le Quy Don Technical University
[*]Corresponding author, email: quanguyhn@lqdtu.edu.vn DOI: 10.56651/lqdtu.jst.v13.n01.820.ict

There have been numerous methods [4]–[6] for video anomaly detection. Based on the availability of anomalous and normal data during the training process, these methods are classified into two main groups: unsupervised and supervised methods. Unsupervised methods refer to the scenario where the model is trained only on normal samples. These methods are more popular in video anomaly detection since they do not require collecting the abnormal data. In reality, abnormal videos are often much rarer than normal ones. Thus, gathering and labeling anomalous videos is usually time-consuming and expensive. Moreover, owing to privacy and security concerns, collecting real-world videos is very difficult.

Among unsupervised methods for video anomaly detection, autoencoder-based methods [7]–[12] are widely used and they have achieved great success. An autoencoder allows to transform the input data into a new and more compact representation expressing the underlining characteristics of the original data. Thus, these models are very useful for anomaly detection in videos.

There are two typical approaches to applying autoencoders for video anomaly detection. The first approach uses the reconstruction error (RE) as the indication for the anomaly [4], [12]–[15]. This approach assumes that the trained autoencoder model reconstructs well on normal samples but does not perform well on abnormal samples. The model is trained on only normal samples using the reconstruction loss. During the testing, any sample that has an RE value greater than a pre-defined threshold is considered an anomaly. The second approach uses the distance between the predicted frame and the ground truth as the indication [10], [16]–[18]. A sequence of T frames is input to the autoencoder model to predict the next frame, i.e. the frame at time step $T + 1$. The loss function commonly used is mean square error (also called prediction error) between the predicted frame and the ground truth frame. During the testing, any frame that has the prediction error greater than a threshold is treated as an anomaly.

Recently, several memory-augmented autoencoders [5], [13], [18]–[20] have been developed for video anomaly detection. These models use an external memory to store features of the previous frames. The memory is updated during the training process to save the most important features as well as eliminate the less important ones. In the testing, the features in the memory are then used to predict the output for the input frames. Specifically, the features of a new input frame are calculated based on its similarity with the features stored in the memory.

Although memory-augmented autoencoders have achieved remarkable success in video anomaly detection [5], [20], these models are often trained in unsupervised mode using only the normal samples. However, when being trained in the unsupervised mode, the models are usually over-fitted to the normal data resulting in high value of mis-detection rate. In other words, these models tend to predict the abnormal samples to be the normal ones.

To alleviate this problem, we propose a novel method to train the memory-augmented autoencoder [18] in supervised mode by generating pseudo

abnormal videos. These pseudo anomalous videos are generated based on key frame selection techniques. Specifically, for a sequence of normal frames, the important or key frames are identified and the less important or non-key frames are eliminated. The resulting sequence is considered anomalous, as the movement of the objects and events is not smooth and continuous due to the removal of some frames. The reason for the usage of key frame selection is that this technique helps to create motion anomalies and these types of anomalies are often popular in camera surveillance [4]. Moreover, key frame selection techniques help to eliminate less important frames while maintaining the essential frames in video sequences. Thus, these techniques potentially help to improve the effectiveness of the memory-augmented autoencoder.

After the pseudo abnormal videos are generated, they are combined with the original benign data to form the augmented dataset. The memory-augmented autoencoder is then trained in supervised mode using the augmented dataset for the prediction task. In other words, the memory-augmented autoencoder is trained to predict the next frame given that a sequence of T previous frames is input. Experimental results evidence the advantage of our solution compared to some recent models for video anomaly detection on three popular benchmarking datasets. The main contributions of the article are:

- Three new key frame selection techniques are proposed that are subsequently used to synthesize anomalous data samples to augment the training dataset for video anomaly detection.
- The integration of the pseudo abnormal data into a memory-augmented autoencoder model is proposed to improve the accuracy of anomaly detection in videos.
- Extensive experiments are conducted on three popular benchmarking datasets to demonstrate the effectiveness of the proposed methods.

The rest of the article is organized as follows. The background and related work of our proposed method are provided in section 2 and section 3, respectively. Detailed proposed methods are given in section 4. section 5 presents the datasets and parameter settings. Next, the experimental results are presented and discussed in section 6. Finally, section 7 summarizes the article and highlights the future research directions.

2. Background

This section presents the memory-augmented autoencoder and the key frame selection techniques forming the foundation of our proposed solutions.

2.1. Memory-augmented autoencoder

A memory-augmented autoencoder called Memory-guided Normality for Anomaly Detection (MNAD) is proposed by Park et al. in [18]. The idea of MNAD is to learn the patterns in the training normal data and store these patterns in the memory module. The architecture of MNAD consists of an encoder, a decoder, and a memory module (Figure 1). The encoder projects the input frame into a new feature at the latent space.

The memory module stores the patterns in the latent space and the decoder maps the feature of the current frame and the retrieved feature from the memory module to reconstruct the input at the output.

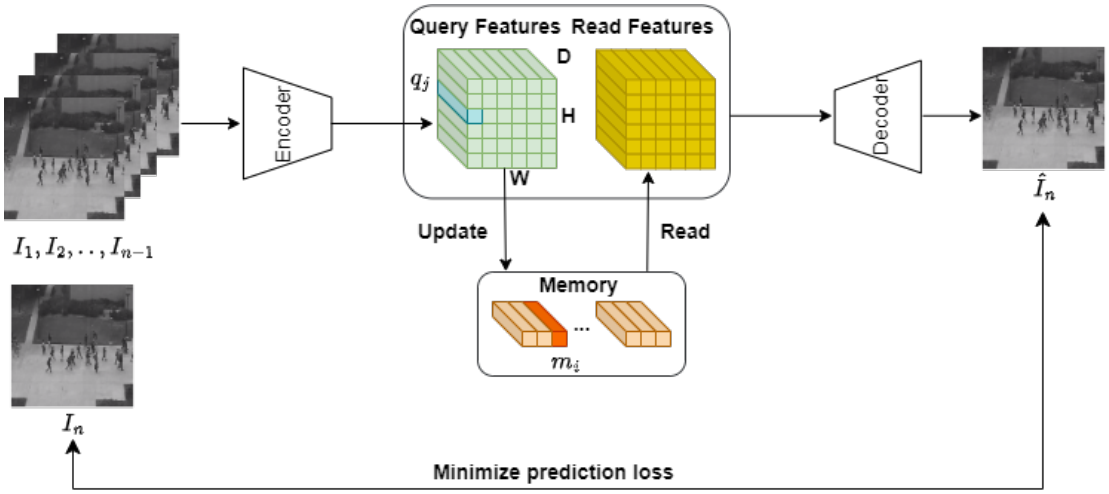

Fig. 1. The architecture of Memory-guided Normality for Anomaly Detection model - (MNAD).

Let a training sample consists of n frames: $(I_1, I_2, ..., I_{n-1}, I_n)$, in which the first $n-1$ frames $(I_1, I_2, ..., I_{n-1})$ are used as the input and the last frame I_n is used as the target frame. The training process involves updating features from normal data into memory and reading stored features in memory to predict the next frame. Specifically, the input frames are encoded into a feature map of size H x W x D, which is then decomposed into H x W queries (q_j, j = 1,.., H x W). Each query has dimensions of 1 x 1 x D. The features are updated into the memory based on the matching between them and the items m_i in the memory. Moreover, the features are read from the memory based on the matching between the query and the items in the memory.

The loss function of MNAD includes three terms as in equation 1. The first term is the intensity loss, which measures the difference between the input and the predicted frame using the L2 distance. The second term is the feature compactness loss that encourages the memory items of the same objects to be similar. The last term is the feature separateness loss that is used to prevent memory items of different objects from getting too close to each other.

$$L_{\text{n}} = L_{\text{intensity}} + \sigma_c L_{\text{compact}} + \sigma_s L_{\text{separate}} \tag{1}$$

with

$$L_{\text{intensity}} = \left\| \hat{I}_n - I_n \right\|_2 \tag{2}$$

$$L_{\text{compact}} = \sum_{j=1}^{H \times W} \|q_j - m_i\|_2 \tag{3}$$

$$L_{\text{separate}} = \max\{0, \sum_{j=1}^{H\times W} \left(\|q_j - m_i\|_2 - \|q_j - m_k\|_2 + r\right)\} \tag{4}$$

where σ_c and σ_s are scale parameters, m_i and m_k is the nearest item and the second one of the query q_j in the memory, respectively, and r is a margin.

During the inference stage, a frame I_n is identified as normal or abnormal based on its anomaly score A_n, which is the sum of the normalized Peak Signal to Noise Ratio $f(P_n)$ between the predicted and ground truth frame, and the normalized L2 distance $f(D_n)$ between the queries and the nearest memory item, as follows:

$$A_n = \alpha(1 - f(P_n(\hat{I}_n - I_n))) + (1 - \alpha)f(D_n(q, m)) \tag{5}$$

with

$$P_n(\hat{I}_n, I_n) = 10\log_{10}\frac{\max\left(\hat{I}_n\right)}{\left\|\hat{I}_n - I_n\right\|_2^2/M} \tag{6}$$

$$D_n(q, m) = \frac{1}{H \times W}\sum_{j=1}^{H\times W}\|q_j - m_i\|_2 \tag{7}$$

where α is a ratio parameter, $f(.)$ is the min-max normalization [17] over whole video frames, M is the total number of pixels in the predicted frame $\hat{I}_n$, and m_i is the nearest memory item of the query q_j.

2.2. Key frame selection

Key frames are widely used in video compression, video summarization and detection, video retrieval and extraction, video analysis and understanding [21]–[23]. Generally, key frames are understood as frames selected from a sequence of frames in a video that represent the main or significant content of the original video. The criterion for assessing the importance of a frame is often based on significant changes in the content of that frame compared to neighboring frames or the occurrence of important events such as new objects or actions in the frame.

Formally, given a list of original frames in a video $N = [I_1, I_2, .., I_n]$, key frames are defined as the smallest list of frames $N_k = [K_{k_1}, K_{k_2}, .., K_{k_m}]$ where $k_1, k_2, .., k_m \in \{1, 2, .., n\}$ such that the original frame set can be recovered from the key frame list [24].

There have been several techniques proposed for extracting key frames from videos. For example, Wei el al. [25] introduced a key frame extraction algorithm that utilizes affective saliency estimation to enhance emotion recognition in video. The key frames are selected to minimize the impact of emotion-irrelevant frames on recognition

outcomes. Abraham et al. [26] applied Differential Evolution (DE) algorithm to selecting the key frames. Each individual is a set of candidate key frames, and an initial population is generated randomly. The fitness function is set by one of three metrics: Structural Similarity Index Method (SSIM), Euclidean Distance, or Entropy Difference. After a number of generations, the best individual is selected as a set of key frames. Mangai el al. [27] proposed to select key frames by applying the HSV histogram and K-means algorithm. The temporal features from the key frames are then extracted by a convolutional neural network, and a concatenated ConvLSTM is employed to identify the activities in a frame.

The main advantage of key frame selection is that it removes unnecessary frames, thereby saving processing time or computational costs [28]. Therefore, models can focus only on learning the most important frames or the critical events in a video [27]. However, key frame selection techniques may also have potential drawbacks. Long videos can make some key frame extraction algorithms inefficient due to their containing complex motions that occur over extended periods [29]. Moreover, there is a risk of losing important frames during the extraction process, as the selection criteria may not capture all important features of the original video. To leverage the strengths of key frame selection techniques, we propose simple yet efficient algorithms that extract key frames from normal videos, and the key frames are then used to generate pseudo anomalous samples representing the anomalous motion of objects. The experimental results show that our proposed method could help improve the performance of a video anomaly detection model. The details of the proposed method are presented in section 4. In the next section, some previous will be analyzed research on generating pseudo anomalies for video anomaly detection.

3. Related work

This section briefly reviews the related work on generating pseudo anomaly data for videos. Due to the lack of anomalous data to supervise models, autoencoders for video anomaly detection are often trained in an unsupervised manner. However, when trained in an unsupervised mode, an autoencoder is often overfitted to the normal data, resulting in low RE for abnormal input frames. In other words, the trained autoencoder also reconstructs well for abnormal input. Thus, it is challenging to detect abnormal events using these models. Subsequently, generating pseudo anomaly data is a potential direction to improve the effectiveness of autoencoders in video anomaly detection.

There have been a number of studies focusing on generating pseudo anomaly data for video anomaly detection. For example, Ionescu et al. [9] addressed the challenge of insufficient truly abnormal training samples by clustering normal training samples using K-means and then sampling from other clusters to form dummy abnormal samples. Zaheer et al. proposed a pseudo-anomaly module in OGNet [14], where any pair of normal frames is passed through a weak generator called G^{old}. Subsequently, they are mixed together and passed through a strong generator called G to generate a pseudo

anomaly frame. The pseudo anomaly and real normal frames are then trained by two generators and a discriminator in an adversarial model. Park et al. [16] adopted two data transformation techniques named Spatial Rotation Transformation (SRT) and Temporal Mixing Transformation (TMT). They applied these techniques to randomly selected patches from the training set to generate abnormal appearance and motion, respectively. The objective is to encourage the model to learn spatially and temporally invariant features of normal frames by attempting to predict the next occurring frame.

In a recent study [4], Strid et al. utilized various data transformations, including skip-frames, repeat-frames, patches, and fusion on normal frames to generate different types of abnormal data. The experimental results have demonstrated the effectiveness of these transformations in enhancing the accuracy of the anomaly detection model. Huang et al. [30] introduced a process where random masks are applied to an original video frame to create the corresponding masked pseudo anomaly video frame. A transformer encoder is then applied to extract latent features, which are used to predict the next frames by a linear decoder. Ristea et al. [31] constructed a set of synthetic anomalies by cropping out abnormal objects from the UBnormal dataset and pasting them onto normal samples to train a masked auto-encoder [32]. This research opens up a direction to generate pseudo anomaly data by combining data from virtual environments with data collected in the real world.

4. Proposed method

This section presents in detail three proposed techniques for extracting key frames and a model for video anomaly detection that combines the generated pseudo anomaly and the memory-augmented autoencoder.

4.1. Key frame selection

The motivation for our proposed key frame selection techniques is inspired by the skip-frames technique in generating pseudo-anomalous samples. Skip-frames technique helps to create fast-moving objects and these objects are one of the typical anomalies in videos. Previous studies on video anomaly detection [4], [15], [33] have utilized the skip-frames technique and shown that it helps improve the accuracy of anomaly detection models. Similar to skip-frames technique, our techniques focus on synthesizing pseudo anomalous samples related to the abnormal motion of objects in videos. However, unlike the skip-frames technique where the jump between two frames is uniform, the jump between our selected frames is not uniform because key frames are extracted depending on the content of actions in each video.

The three proposed key frame selection techniques include key frame with Threshold (KF-TH), key frame with Mean Distance (KF-MD), and key frame with Non-consecutive Pair (KF-NP). The KF-TH technique considers a pair of consecutive frames that are significantly different as the key frames. Specifically, KF-TH calculates the distance d between a pair of consecutive frames and compares d with a threshold. Any pair with d

greater than the threshold is added to the key frame list. Algorithm 1 describes KF-TH in detail.

Algorithm 1: Key frame with Threshold - (KF-TH)

Input: All frames in a video: V, a threshold: T
Output: List of key frames, K

$N = len(V)$
for $i = 1$ ***to*** $N - 1$ **do**
  $d = distance(V[i], V[i + 1])$
  if $d \geq T$ **then**
    $K = K \cup V[i]$
    $K = K \cup V[i + 1]$
  end
end
return K

In Algorithm 1, the function $len(V)$ returns the number frames N in the video V, and the function $distance(V[i], V[i + 1])$ calculates the distance (the entropy) between frame $V[i]$ and $V[i + 1]$. Afterward, the distance of the frame pair is compared to a threshold T, and they are added to the key frame list if $d \geq T$.

Algorithm 2: Key frame with Mean Distance - (KF-MD)

Input: All frames in a video: V
Output: List of key frames, K

$N = len(V)$
for $i = 1$ ***to*** $N - 1$ **do**
  $d[i] = distance(V[i], V[i + 1])$
end
$T = mean(d)$
for $i = 1$ ***to*** $N - 1$ **do**
  if $d[i] \geq T$ **then**
    $K = K \cup V[i]$
    $K = K \cup V[i + 1]$
  end
end
return K

The second key frame selection technique (KF-MD) aims to eliminate the need to calibrate the threshold T in KF-TH, where the average distance of two consecutive frames is used as the threshold. KF-MD is described in detail in Algorithm 2. Similar to KF-TH, KF-MD differs only in that the threshold T is set to the average distance of two consecutive frames in the sequence using the function $mean(d)$.

The third key frame selection technique (KF-NP) differs from the two previous techniques in that it considers pairs of non-consecutive frames as candidates for key frames. Algorithm 3 presents this technique in detail. The first loop calculates the distance between each pair of two consecutive frames to determine the threshold T, similar to Algorithm2. After that, KF-NP iterates over pairs of non-consecutive frames. If the distance of this pair is greater than T, then the start frame is added to the key frame list. The process of selecting key frames continues by assigning the start frame to be the last key frame. In case the distance of the pair is less than T, the end frame is assigned to the next frame.

Algorithm 3: Key frame with Non-consecutive Pair - (KF-NP)

Input: All frames in a video: V
Output: List of key frames, K

$N = len(V)$
for $i = 1$ ***to*** $N - 1$ **do**
  $d[i] = distance(V[i], V[i+1])$
end
$T = mean(d)$
$start = 1$
$end = start + 1$
while $(end \leq N)$ **do**
  if $distance(V[start], V[end]) \geq T$ **then**
    $K = K \cup V[start]$
    $start = end$
    $end = start + 1$
  end
  else
    $end = end + 1$
  end
end
return K

One advantage of our key frame selection algorithms is that they are simpler than other key frame selection techniques such as DE-Entropy [34] and K-means [35]. In fact, the complexity of our three algorithms is $O(N)$ while the complexity of K-mean and DE-Entropy are $O(N * K * D)$ and $O(G * C * N)$, respectively. In the above equations, N is the number of frames in a video, K and D are the number of clusters and the number of iterations in K-mean, G and C are the number of generations and the number of individuals in DE-Entropy, respectively. Moreover, our key frame selection techniques are then used to generate abnormal data for augmenting the training data of memory-augmented autoencoder. The experimental results show the effectiveness of our three algorithms in enhancing the accuracy of memory-augmented autoencoder.

4.2. Memory-augmented autoencoder with key frames

This subsection presents the proposed model for anomaly detection that uses key frames as pseudo anomalies to supervise the memory-augmented autoencoder. The proposed model is called Pseudo Anomalies-Memory-augmented autoencoder with key frames and abbreviated as PAMAE-KF.

Figure 2 describes PAMAE-KF in detail. PAMAE-KF includes two main components: a Data Generator (DG) and a Memory-augmented autoencoder (MAE). The DG component aims to generate training data for the MAE component. The training data includes both the normal frames and the pseudo anomaly ones. The normal data is drawn directly from the original videos. In other words, input videos are split into sequences of frames and each training sample includes n consecutive frames $(I_1, I_2, ..., I_{n-1}, I_n)$ where the first $n-1$ frames $(I_1, I_2, ..., I_{n-1})$ are input to MAE, and the last frame I_n is used as the label, i.e., the target frame. The anomalous training samples are taken from the key frame list. In other words, one of the three key frame selection techniques is used to extract the key frames from the original videos and an anomalous training sample also includes n consecutive key frames, $(K_1, K_2, ..., K_{n-1}, K_n)$ in the key frame list.

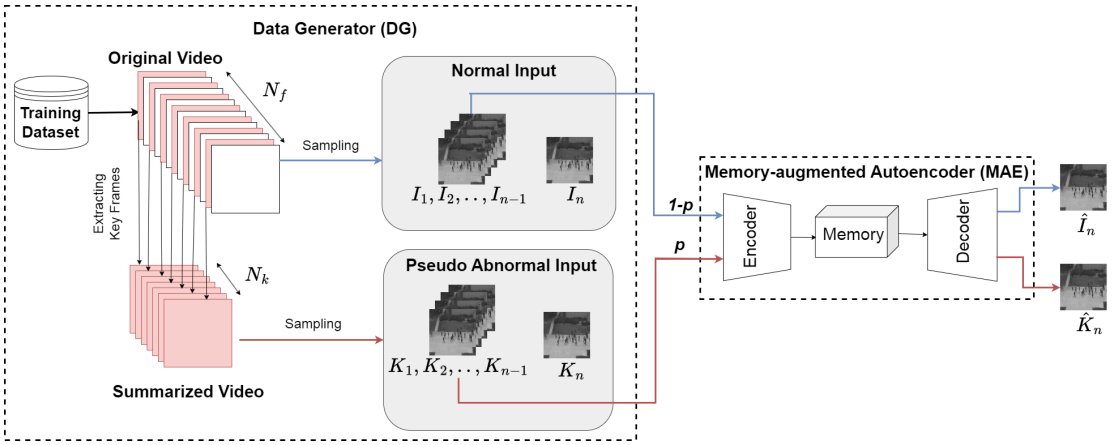

Fig. 2. Pseudo Anomalies-Memory-augmented autoencoder with key frames - (PAMAE-KF).

The second component in PAMAE-KF is a memory-augmented autoencoder. That is similar to the model in [18]. The only difference is that this model is trained on both normal and pseudo abnormal data while the model in [18] is only trained on the normal data. During the training, the pseudo abnormal data is selected with the probability p and the original data is selected with probability $(1-p)$. Adding the pseudo abnormal data to the training process of PAMAE-KF helps to leverage its accuracy in detecting anomalies in videos.

5. Experimental settings

Datasets: The proposed methods are tested on three well-known datasets: Ped2 [36], Avenue [37], and ShanghaiTech [38]. These datasets consist of videos captured by cameras placed at various locations within university campuses. Ped2 and Avenue datasets are single-scene datasets, comprising videos obtained from a single camera, while ShanghaiTech dataset is a multi-scene dataset, consisting of videos captured from multiple cameras. In these three datasets, pedestrians traversing the campus pathways are considered normal, while the appearance of anomalous objects such as bicycles, carts, skateboards, or abnormal actions like running, chasing, fighting, and throwing objects are regarded as abnormal events.

The Ped2 dataset is the smallest and has the simplest context among the three datasets. Compared to the Ped2 dataset, the Avenue dataset poses greater challenges due to the higher number of abnormal events, totaling up to 47 events. Moreover, the camera is shaken and the objects are occluded, complicating the anomaly detection problem. The ShanghaiTech dataset presents the biggest challenge among the three datasets, not only due to its large image size and longer duration, but also because it involves images captured from 13 different cameras, containing a total of 130 distinct abnormal events.

Experimental settings: Three proposed key frame selection techniques, including KF-TH, KF-MD and KF-NP are used in PAMAE-KF. These models are shortened as PAMAE-KF-TH, PAMAE-KF-MD and PAMAE-KF-NP, respectively.

In KF-TH, the threshold T is selected so that about 70% of frames are selected into the key frame list. This value is carefully calibrated from the experiments for the good performance of PAMAE-KF-TH in detecting anomalies. If the rate is too low, the number of removed frames will be too high leading to losing important information from the videos. Conversely, if the rate is too high, the number of key frames will nearly match the number of original frames, rendering the process of generating pseudo anomalies from the key frames ineffective. All models are implemented in Python using the PyTorch library, and the experiments are conducted on Nvidia RTX 3060 12GB card.

Most hyper-parameters of PAMAE-KF are kept the same as the parameters in MNAD [18], except for the ratio α in the anomaly score formula. In MNAD [18], α is set at 0.4 for all datasets while in our experiments, the values of α for three dataset Ped2, Avenue, and ShanghaiTech are 0.6, 0.5, and 0.4, respectively. For probability of using pseudo anomaly data in training, denoted as p, we apply the grid search in range of [0.001, 0.002, 0.005, 0.01, 0.02]. The best values of p for the three datasets Ped2, Avenue, and ShanghaiTech are 0.002, 0.002, and 0.01, respectively.

Evaluation metrics: After computing anomaly scores for all frames, the anomaly threshold value is chosen to determine whether those frames are normal or abnormal. Based on this, we obtain True Positive Rate (TPR) against False Positive Rate (FPR) values, resulting in the plotting of the ROC curve. Finally, Area Under the ROC Curve, abbreviated as AUC, is used to evaluate the accuracy of anomaly detection models on the test dataset, and the higher AUC values correspond to better performance outcomes.

6. Results and discussion

This section presents the experimental results and discussion. Four sets of experiments will be presented and analysed in this section. The first set is to compare different models with and without memory module. The second one is to compare our best model with the other models for video anomaly detection. The third set is to compare our key frame selection techniques with two other popular key frame selection techniques. The last set is to visualize the prediction results of the proposed model.

6.1. Comparing with the base network architectures

This subsection compares PAMAE-KF with MNAD [18]. MNAD is the memory-augmented autoencoder for anomaly detection that is the foundation for PAMEA-KF. In other words, PAMEA-KF is an improvement over MNAD by adding a pseudo anomaly data generator to MNAD's input. To compare PAMEA-KF with MNAD, we set up two experimental sets. The first set uses the memory module in both PAMEA-KF and MNAD and the second set removes this module from both models. Table 1 presents the AUC (%) scores of PAMAE-KF and MNAD on the three test datasets.

Table 1. AUC scores of PAMAE-KF and MNAD: boldface indicates better performance than MNAD, and underlining indicates the best result

Methods	Ped2	Avenue	ShanghaiTech
A. **Without memory**			
MNAD w/o Mem [18]	94.30	84.50	66.80
PAMAE-KF-TH w/o Mem	**96.03**	**84.53**	**70.26**
PAMAE-KF-MD w/o Mem	**97.27**	**85.10**	**69.18**
PAMAE-KF-NP w/o Mem	**97.35**	**85.58**	**71.20**
B. **With memory**			
MNAD w/ Mem [18]	97.00	88.50	70.50
PAMAE-KF-TH	**97.91**	88.04	**71.06**
PAMAE-KF-MD	**97.65**	87.76	**70.95**
PAMAE-KF-NP	**98.21**	**88.70**	**71.81**

It can be seen in this table that the AUC score of PAMAE-KF is frequently better than that of MNAD. The AUC values of PAMAE-KF are higher than those of MNAD in most experiment settings. Particularly, PAMAE-KF-NP consistently achieves the best result on all three datasets in both settings, i.e., with and without the memory module. For example, the AUC of PAMAE-KF-NP without the memory module on Ped2, Avenue, and ShanghaiTech are 97.35%, 85.58%, and 71.20%, respectively, while these values of MNAD are only 94.30%, 84.50%, and 66.80%, respectively. Similarly, the AUC scores of PAMAE-KF-NP with the memory module are also higher than those of MNAD by margins of 1.21%, 0.20%, and 1.31%. These results provide evidence for the usefulness of the proposed techniques for generating pseudo anomaly data.

Comparing the three proposed key frame selection techniques, table 1 shows that the KF-NP technique often helps the PAMAE-KF model achieve the best performance, while

the performance of PAMAE-KF using KF-TH and KF-MD is roughly equal. Specifically, the AUC scores of PAMAE-KF-MD without the memory module are slightly higher than those of PAMAE-KF-TH on Ped2 and Avenue datasets. Conversely, when using the memory module, the AUC score of PAMAE-KF-TH is slightly better than that of PAMAE-KF-MD.

Overall, the results in this subsection show that using the proposed key frame selection techniques to generate pseudo anomaly data helps the memory-augmented autoencoder, i.e., MNAD [18], achieve better performance in detecting anomaly events in videos.

6.2. Comparing with the other methods for video anomaly detection

This subsection compares PAMAE-KF with some recent methods for video anomaly detection. The best version of PAMAE-KF, i.e., PAMAE-KF-NP, is compared with two classes of video anomaly detection methods. The first class are memory-augmented methods, including MemAE [13], MNAD [18], and AMMC-Net [19], while the second class involves data-augmented methods, including OG [14], LNTRA-Patch [15], FastAno [16], PseudoBound-Repeat, PseudoBound-Patch, PseudoBound-Fusion and PseudoBound-Noise [4]. Table 2 shows the AUC (%) scores of PAMAE-KF and the other methods for detecting abnormal events in video.

Table 2. AUC scores of PAMAE-KF and the other methods

Year	Methods	Ped2	Avenue	ShanghaiTech
A. **Memory-augmented methods**				
2019	MemAE [13]	94.10	83.30	71.20
2020	MNAD [18]	97.0	88.50	70.50
2021	AMMC-Net [19]	96.60	86.60	**73.70**
B. **Data augmented methods**				
2020	OG [14]	98.10	-	-
2021	LNTRA-Patch [15]	94.77	84.91	72.46
2022	FastAno [16]	96.30	85.30	72.20
2023	PseudoBound-Repeat [4]	93.69	81.87	72.58
2023	PseudoBound-Patch [4]	95.33	85.36	**72.77**
2023	PseudoBound-Fusion [4]	94.16	82.79	71.52
2023	PseudoBound-Noise [4]	97.78	82.11	72.02
-	PAMAE-KF	**98.21**	**88.70**	71.81

It can be observed from table 2 that PAMAE-KF often achieves the highest AUC score in both classes on two datasets: Ped2 and Avenue. For example, on Ped2, the AUC score of PAMAE-KF is 98.21%, while the values of AMMC-Net and PseudoBound-Patch are 96.60% and 95.53%, respectively. Similarly, on the Avenue dataset, the AUC score of PAMAE-KF is also higher than that of MNAD and PseudoBound-Patch by margins of 0.20% and 3.34%, respectively. In the ShanghaiTech dataset, the result of PAMAE-KF is not as good as in Ped2 and Avenue. However, the AUC score of PAMAE-KF is still

better than that of the two methods in the first class, i.e., MemAE and MNAD, and that of PseudoBound-Fusion in the second class. Moreover, no data augmented methods can achieve better results compared to the best memory-based method, AMMC-Net. The reason could be that the scenarios in the videos of the ShanghaiTech dataset are highly complicated. Thus, the augmented anomaly data could be mis-predicted as the normal data. Our future research will further investigate this issue.

6.3. Comparing with the other key frame selection techniques

This subsection compares three proposed key frame selection techniques with two other popular methods including DE-Entropy and K-means. DE-Entropy leverages the advantage of evolution through three key operations - mutation, crossover, and selection - to intelligently identify the most important and representative frames in a video sequence. Meanwhile, K-means selects key frames as cluster centers of the clustering process of extracted features of the frames. The results of this experiment are presented in table 3. In this table, MNAD refers to the base line model, i.e., the memory-augmented autoencoder [18].

Table 3. AUC scores of PAMAE-KF model with different extracted key frames: boldface indicates better performance than MNAD, and underlining indicates the best result

Methods	Ped2	Avenue	ShanghaiTech
MNAD w/ Mem [18]	97.00	88.50	70.50
PAMAE-KF w/ DE-Entropy	96.67	85.30	68.67
PAMAE-KF w/ K-means	**97.26**	87.27	**71.74**
PAMAE-KF w/ KF-TH	**97.91**	88.04	**71.06**
PAMAE-KF w/ KF-MD	**97.65**	87.76	**70.95**
PAMAE-KF w/ KF-NP	**98.21**	**88.70**	**71.81**

Table 3 shows that the AUC scores of PAMAE-KF with key frames extracted by our three techniques (PAMAE-KF-TH, PAMAE-KF-MD, PAMAE-KF-NP) are higher than those of PAMAE-KF with DE-Entropy and K-means. For instance, the AUC scores of PAMAE-KF-NP are higher than those of PAMAE-KF with K-means by margins of 0.95%, 1.43%, and 0.07% on Ped2, Avenue and ShanghaiTech, respectively, while the values of PAMAE-KF with DE-Entropy are the lowest among all models. Overall, the results above prove the superiority of our techniques compared to DE-Entropy or K-means techniques.

Table 3 also indicates that not all key frame selection techniques can improve the anomaly detection performance of the model. For example, the AUC scores of the model with the key frames extracted by the DE-Entropy technique were lower than those of MNAD across all three datasets. In detail, the AUC of PAMAE-KF with DE-Entropy on Ped2, Avenue, and ShanghaiTech are 96.67%, 85.30%, and 68.67%, respectively, while these values of MNAD are only 97.00%, 88.50%, and 70.50%, respectively. On the contrary, our best proposed technique, PAMAE-KF-NP, improves the accuracy on all three datasets. For example, the AUC scores of PAMAE-KF-NP are higher than

those of MNAD on the Ped2, Avenue and ShanghaiTech datasets by margins of 1.21%, 0.20% and 1.31%, respectively.

6.4. Qualitative results

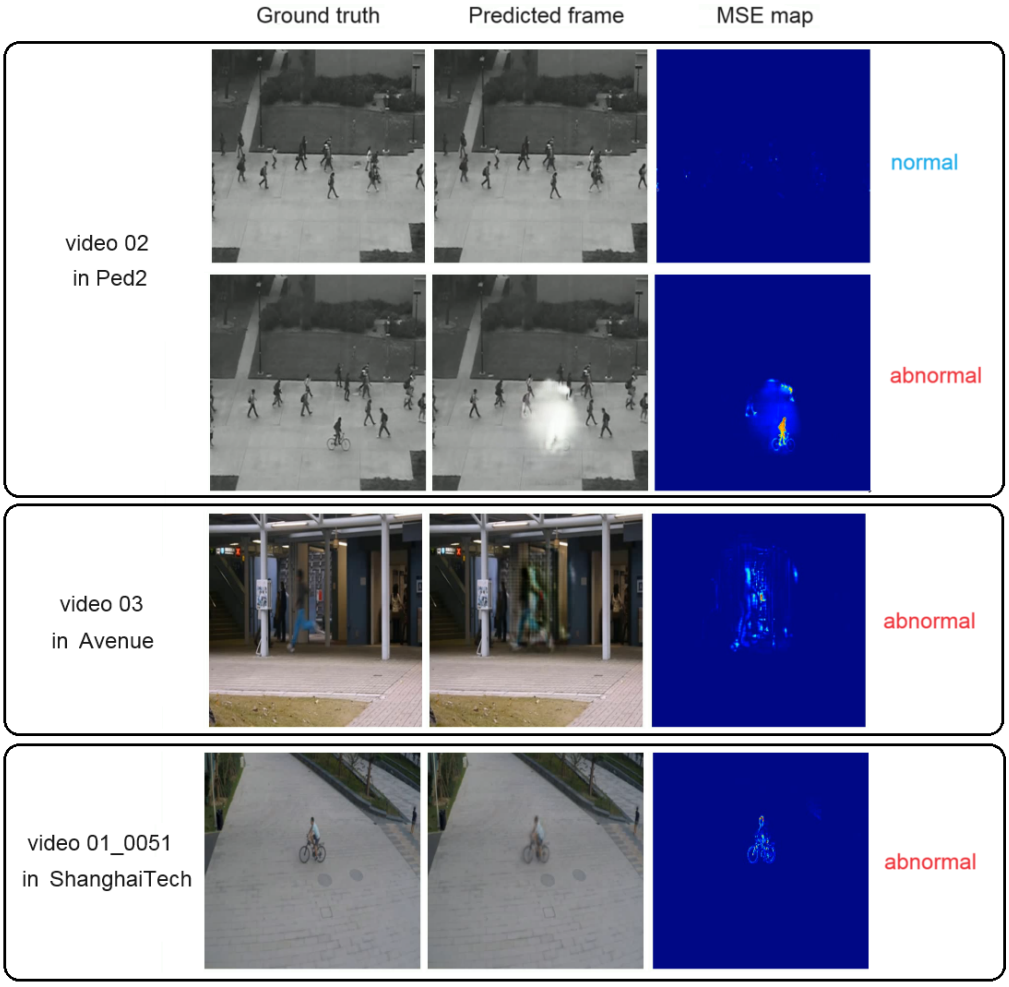

Fig. 3. A visual representation of anomaly detection using the PAMAE-KF-NP method is provided. Each row represents a video from the Ped2, Avenue, and ShanghaiTech datasets, respectively. From left to right, the columns display the ground truth (left), predicted frame (middle), and MSE (Mean Square Error) map (right)

This section analyzes the behavior of PAMAE-KF by visualizing its mean square error maps and its detection scores when predicting a sequence of frames.

First, the MSE map is calculated by the MSE between an input frame and its corresponding predicted frame. An example of the MSE map in the three datasets is presented in figure 3. In this figure, the ground truth is presented on the left, the predicted frames are presented in the middle, and the MSE maps are on the right. It

can be seen from this figure that when the abnormal object appears in the frame, its predicted error is greater than the background. Thus, these abnormal objects can be detected by the PAMAE-KF models.

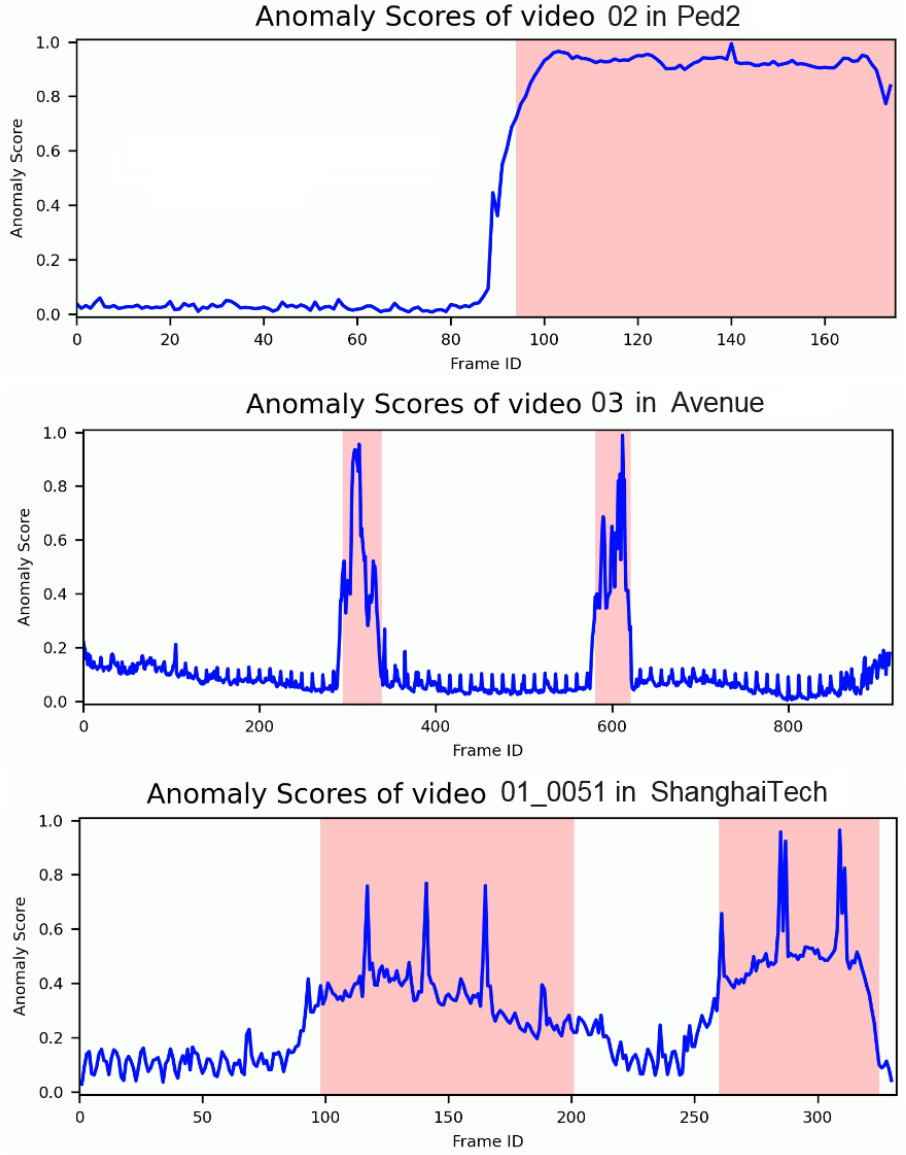

Fig. 4. The anomaly score plots of three selected videos in three datasets: the pink shaded areas indicate frames labeled as anomalous, while the blue line represents the anomaly scores of frames calculated by PAMAE-KF-NP

Secondly, figure 4 displays the anomaly scores of PAMAE-KF for three videos from the Ped2, Avenue, and ShanghaiTech datasets, respectively. In this figure, the blue line represents the anomaly scores plotted against the sequence of frames, while the shaded areas indicate the intervals when abnormal objects appear. It can be observed that when

an abnormal event occurs in the frame sequence, the anomaly score sharply increases. Conversely, these scores remain relatively low for normal events. Thus, this score serves as a good indicator for detecting abnormal events in videos.

7. Conclusion and future work

In this article, we proposed the novel model for anomaly detection based on the key frame selection techniques and a memory-augmented autoencoder. The proposed model, shortened as PAMAE-KF, includes two components: a data generator and a memory-augmented autoencoder. The data generator generates pseudo anomaly data from the original videos by using the key frame selection techniques. The memory-augmented autoencoder is trained in a supervised mode using both the original normal frame and the pseudo anomaly frames. We evaluated the effectiveness of PAMAE-KF on three benchmarking datasets: Ped2, Avenue and ShanghaiTech. The experiments show that PAMAE-KF outperform the base model, i.e., MNAD [18] and some recent methods for video anomaly detection. Further analysis also helps to understand the performance of PAMAE-KF.

There are some future research directions arisen from this article. First, the pseudo anomaly data generated in this work is based on discontinuous motion. In the future, we aim to generate pseudo anomaly data using other techniques such as patching, masking, etc. Secondly, it is also possible to use other key frame selection techniques such as shot-based methods, clustering-based methods and optical flow methods. Last but not least, we also plan to apply the pseudo anomaly data to other models and test them on other video anomaly detection datasets to better understand its weaknesses and strengths.

Acknowledgment

The work in this article was funded by VINGROUP INNOVATION FOUNDATION (VINIF), under grant number VINIF.2023.DA059.

References

[1] S. Anoopa and A. Salim, "Survey on anomaly detection in surveillance videos," *Materials Today: Proceedings*, vol. 58, 2022, pp. 162–167, Doi: 10.1016/j.matpr.2022.01.171.

[2] M. Baradaran and R. Bergevin, "A critical study on the recent deep learning based semi-supervised video anomaly detection methods," *Multimedia Tools and Applications*, 2023, pp. 1–47, Doi: 10.1007/s11042-023-16445-z.

[3] Y. Liu, D. Yang, Y. Wang, ..., and L. Song, "Generalized video anomaly event detection: Systematic taxonomy and comparison of deep models," *ACM Computing Surveys*, 2023, Doi: 10.1145/3645101.

[4] M. Astrid, M. Z. Zaheer, and S.-I. Lee, "Pseudobound: Limiting the anomaly reconstruction capability of one-class classifiers using pseudo anomalies," *Neurocomputing*, vol. 534, 2023, pp. 147–160, Doi: 10.1016/j.neucom.2023.03.008.

[5] L. Wang, J. Tian, S. Zhou, H. Shi, and G. Hua, "Memory-augmented appearance-motion network for video anomaly detection," *Pattern Recognition*, vol. 138, 2023, p. 109335, Doi: 10.1016/j.patcog.2023.109335.

[6] Z. Yang, J. Liu, Z. Wu, P. Wu, and X. Liu, "Video event restoration based on keyframes for video anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 592–14 601, Doi: 10.1109/CVPR52 729.2023.01 402.

[7] V.-T. Le and Y.-G. Kim, "Attention-based residual autoencoder for video anomaly detection," *Applied Intelligence*, vol. 53, no. 3, 2023, pp. 3240–3254, Doi: 10.1007/s10489-022-03613-1.

[8] T.-N. Nguyen and J. Meunier, "Anomaly detection in video sequence with appearance-motion correspondence," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1273–1283, Doi: 10.1109/ICCV.2019.00 136.

[9] R. T. Ionescu, F. S. Khan, M.-I. Georgescu, and L. Shao, "Object-centric auto-encoders and dummy anomalies for abnormal event detection in video," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7842–7851, Doi: 10.1109/CVPR.2019.00 803.

[10] Z. Liu, Y. Nie, C. Long, Q. Zhang, and G. Li, "A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 588–13 597, Doi: 10.1109/ICCV48 922.2021.01 333.

[11] T. Li, X. Chen, F. Zhu, Z. Zhang, and H. Yan, "Two-stream deep spatial-temporal auto-encoder for surveillance video abnormal event detection," *Neurocomputing*, vol. 439, 2021, pp. 256–270, Doi: 10.1016/j.neucom.2021.01.097.

[12] Y. Chang, Z. Tu, W. Xie, ..., and J. Yuan, "Video anomaly detection with spatio-temporal dissociation," *Pattern Recognition*, vol. 122, 2022, p. 108213, Doi: 10.1016/j.patcog.2021.108213.

[13] D. Gong, L. Liu, V. Le, ..., and A. v. d. Hengel, "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1705–1714, Doi: 10.1109/ICCV.2019.00 179.

[14] M. Z. Zaheer, J.-h. Lee, M. Astrid, and S.-I. Lee, "Old is gold: Redefining the adversarially learned one-class classifier training paradigm," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14 183–14 193, Doi: 10.1109/CVPR42 600.2020.01 419.

[15] M. Astrid, M. Z. Zaheer, J.-Y. Lee, and S.-I. Lee, "Learning not to reconstruct anomalies," *British Machine Vision Conference (BMVC)*, 2021, Doi: 10.48550/arXiv.2110.09742.

[16] C. Park, M. Cho, M. Lee, and S. Lee, "Fastano: Fast anomaly detection via spatio-temporal patch transformation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 2249–2259, Doi: 10.1109/WACV51 458.2022.00 197.

[17] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection–a new baseline," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6536–6545, Doi: 10.1109/CVPR.2018.00 684.

[18] H. Park, J. Noh, and B. Ham, "Learning memory-guided normality for anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14 372–14 381, Doi: 10.1109/CVPR42 600.2020.01 438.

[19] R. Cai, H. Zhang, W. Liu, S. Gao, and Z. Hao, "Appearance-motion memory consistency network for video anomaly detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 2, 2021, pp. 938–946, Doi: 10.1609/aaai.v35i2.16 177.

[20] S. Sun and X. Gong, "Hierarchical semantic contrast for scene-aware video anomaly detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 22 846–22 856, Doi: 10.1109/CVPR52 729.2023.02 188.

[21] M. Rochan, L. Ye, and Y. Wang, "Video summarization using fully convolutional sequence networks," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 347–363, Doi: 10.1007/978–3–030–01 258–8_22.

[22] P. Saini, K. Kumar, S. Kashid, A. Saini, and A. Negi, "Video summarization using deep learning techniques: a detailed analysis and investigation," *Artificial Intelligence Review*, vol. 56, no. 11, 2023, pp. 12347–12385, Doi: 10.1007/s10462-023-10444-0.

[23] Y. Jiang, K. Cui, B. Peng, and C. Xu, "Comprehensive video understanding: Video summarization with content-based video recommender design," in *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 2019, Doi: 10.1109/ICCVW.2019.00195.

[24] S. Mei, G. Guan, Z. Wang, S. Wan, M. He, and D. D. Feng, "Video summarization via minimum sparse reconstruction," *Pattern Recognition*, vol. 48, no. 2, 2015, pp. 522–533, Doi: 10.1016/j.patcog.2014.08.002.

[25] J. Wei, X. Yang, and Y. Dong, "User-generated video emotion recognition based on key frames," *Multimedia Tools and Applications*, vol. 80, 2021, pp. 14343–14361, Doi: 10.1007/s11042-020-10203-1.

[26] K. T. Abraham, M. Ashwin, D. Sundar, T. Ashoor, and G. Jeyakumar, "An evolutionary computing approach for

solving key frame extraction problem in video analytics," in *2017 International conference on communication and signal processing (ICCSP)*. IEEE, 2017, pp. 1615–1619, Doi: 10.1109/ICCSP.2017.8 286 663.

[27] P. Mangai, M. K. Geetha, and G. Kumaravelan, "Temporal features-based anomaly detection from surveillance videos using deep learning techniques," in *2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS)*. IEEE, 2022, pp. 490–497, Doi: 10.1109/ICAIS53 314.2022.9 742 960.

[28] G. M. E. Elahi and Y.-H. Yang, "Online learnable keyframe extraction in videos and its application with semantic word vector in action recognition," *Pattern Recognition*, vol. 122, 2022, p. 108273, Doi: 10.1016/j.patcog.2021.108273.

[29] M. K. Asha Paul, J. Kavitha, and P. A. Jansi Rani, "Key-frame extraction techniques: A review," *Recent Patents on Computer Science*, vol. 11, no. 1, 2018, pp. 3–16, Doi: 10.2174/2213275911666180719111118.

[30] X. Huang, C. Zhao, C. Gao, L. Chen, and Z. Wu, "Synthetic pseudo anomalies for unsupervised video anomaly detection: A simple yet efficient framework based on masked autoencoder," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5, Doi: 10.1109/ICASSP49 357.2023.10 094 296.

[31] A. Acsintoae, A. Florescu, M.-I. Georgescu, ..., and M. Shah, "Ubnormal: New benchmark for supervised open-set video anomaly detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 20 143–20 153, Doi: 10.1109/CVPR52 688.2022.01 951.

[32] N.-C. Ristea, F.-A. Croitoru, R. T. Ionescu, ..., and M. Shah, "Self-distilled masked auto-encoders are efficient video anomaly detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 984–15 995, Doi: 10.48 550/arXiv.2306.12 041.

[33] M. Astrid, M. Z. Zaheer, and S.-I. Lee, "Synthetic temporal anomaly guided end-to-end video anomaly detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 207–214, Doi: 10.1109/ICCVW54 120.2021.00 028.

[34] K. T. Abraham, M. Ashwin, D. Sundar, T. Ashoor, and G. Jeyakumar, "Empirical comparison of different key frame extraction approaches with differential evolution based algorithms," in *Intelligent Systems Technologies and Applications*. Springer, 2018, pp. 317–326, Doi: 10.1007/978–3–319–68 385–0_27.

[35] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14. Oakland, CA, USA, 1967, pp. 281–297.

[36] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 1, 2013, pp. 18–32, Doi: 10.1109/TPAMI.2013.111.

[37] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 FPS in MATLAB," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 2720–2727, Doi: 10.1109/ICCV.2013.338.

[38] W. Luo, W. Liu, and S. Gao, "A revisit of sparse coding based anomaly detection in stacked RNN framework," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 341–349, Doi: 10.1109/ICCV.2017.45.

Manuscript received 14-04-2024; Accepted 25-06-2024.

■

Anh Le received his B. Eng. degree and MSc Degree in Computer Science from Le Quy Don Technical University in 2010 and 2015, respectively. He worked at Institute of Simulation Technology for over 10 years in the role of a research scientist in the field of computer graphics and simulation. He is now a PhD student at Le Quy Don Technical University. His research interests are video anomaly detection, computer vision, computer graphics and simulation. E-mail: anhle@lqdtu.edu.vn.

Quang Uy Nguyen has much experience in Deep learning and anomaly detection. He has been the reviewer for a number of high-ranking journals including IEEE Transaction on Cybernetics, IEEE transaction on Evolutionary Computation, IEEE Internet of Things Journal, Information Sciences, Journal of Big Data. He is also the Program Committee member of various leading conferences such as GECCO, CEC, EuroGP, GLOBECOM, WCNC. Nguyen is also the director of the Intelligent Computing Research Group at Le Quy Don Technical University. He has successfully supervised four PhD students and currently supervises other five PhD students. His recent research interests are deep learning, representation learning, anomaly detection, image forensic. He has an H-Index of 17 and his publications have received more than 1350 citations. E-mail: quanguyhn@lqdtu.edu.vn.

Thi Huong Chu received her Bachelor of. Engineering degree in Applied Mathematics and Informatics from Hanoi University of Science and Technology and MSc Degree in Computer Science from Le Quy Don Technical University. She received the PhD degree in Mathematical Foundations for Informatics from Le Quy Don Technical University, in 2020. Her research interests are in the domain of Evolutionary Algorithms, Genetic Programming and Machine Learning. E-mail: huongktqs@gmail.com.

Hai-Hong Phan received a Ph.D. degree in computer science from the University of CY Cergy Paris, France in 2019. She has over ten years of experience in research, teaching, consulting, and implementing information technology projects. She has reviewed articles for some high-ranking international journals. She has published scientific papers in high-quality journals and conferences such as Multimedia Tools Applications, IET Image Processing, and International Conference on Pattern Recognition. Her researches focus on computer vision, image processing, action recognition, and face recognition. E-mail: hongpth@lqdtu.edu.vn.

PHÁT HIỆN BẤT THƯỜNG TRONG VIDEO ÁP DỤNG BỘ MÃ HÓA TỰ ĐỘNG ĐƯỢC TĂNG CƯỜNG BỘ NHỚ BẰNG LỰA CHỌN KHUNG HÌNH CHÍNH

Lê Anh, Nguyễn Quang Uy, Chu Thị Hường, Phan Hải Hồng

Tóm tắt

Trong bài báo này, chúng tôi đề xuất một phương pháp mới để huấn luyện một bộ mã hóa tự động được tăng cường bộ nhớ theo chế độ được giám sát bằng cách sinh các video bất thường giả dựa trên các kỹ thuật lựa chọn khung hình chính. Hầu hết các phương pháp phát hiện bất thường trong video sử dụng mô hình học máy để học các mẫu của các video bình thường. Bất kỳ video nào có các mẫu lệch đáng kể so với các mẫu đã học được coi là bất thường. Tuy nhiên, việc phát triển một mô hình học máy hiệu quả cho phát hiện bất thường trong video là một nhiệm vụ thách thức do sự thiếu hụt các bất thường. Cụ thể, các mẫu bất thường thường hiếm hơn nhiều và khó thu thập hơn so với các mẫu bình thường. Để giải quyết vấn đề này, trong bài báo này, chúng tôi đề xuất một phương pháp mới sử dụng kỹ thuật lựa chọn khung hình chính để tạo ra các giả bất thường. Các giả bất thường được tạo ra sau đó được kết hợp với dữ liệu bình thường để tạo ra tập dữ liệu được tăng cường. Sau đó, bộ mã hóa tự động được tăng cường bộ nhớ được huấn luyện trên các tập dữ liệu được tăng cường. Kết quả thử nghiệm cho thấy rằng các điểm số AUC của giải pháp đề xuất cao hơn các điểm số AUC của kiến trúc mạng cơ sở từ 0.20% đến 1.31% trên cả ba tập cơ sở dữ liệu được biết đến rộng rãi cho phát hiện bất thường trong video.

Từ khóa

Phát hiện bất thường video, bộ mã hóa tự động, bộ sinh giả bất thường, lựa chọn khung hình chính.