FROM VIDEO ANOMALY DETECTION TO PREDICTION: MAKING ABNORMAL JUDGMENTS IN ADVANCE

XINBO GAO.

Chongqing University of Posts and Telecommunications, Chongqing, China
IEEE, IET, CIE, CCF, CAAI Fellow

gaoxb@cqupt.edu.cn

JIAXU LENG

Chongqing University of Posts and Telecommunications, Chongqing, China

lengjx@cqupt.edu.cn

MINGPI TAN

Chongqing University of Posts and Telecommunications, Chongqing, China

tanmingp11@163.com

Abstract

Video Anomaly Detection (VAD), detecting abnormal events in videos that deviate from expected normal patterns, has become a research hotspot due to its potential applications. However, detecting abnormal events that have occurred is relatively meaningless in some situations (e.g., traffic accidents), where an advanced judgment for anomalies is much more significant. To this end, we introduce a new, challenging yet valuable task, named Video Anomaly Prediction (VAP). In this article, we take a systematic look at the VAP task, including its definition, challenges, corresponding baseline method and so on. Moreover, we point out some future opportunities that we will focus on to accelerate the development of this task.

What are Anomalies?

Anomaly analyses are essential with critical applications in video surveillance, automatic driving, Consumer electronics and so on. According to Karl Raimund Popper's famous theory that scientific theories must be falsifiable, we can appreciate the definition of anomalies and the great significance of anomaly analyses. Anomalies are usually defined as deviations from a common rule or what is regarded as standard, normal, or expected and distinguished with noise that has no value. In videos, anomalies are specifically defined as irregular behaviours or objects that do not conform to the normality of the current scene, following the definition 1 provided in [1]. Fig.1 shows some examples from public VAD datasets, UCSD Ped2[2] and CUHK Avenue[3].

Definition 1 *Video anomalies can be thought of as the occurrence of unusual appearance or motion attributes or the occurrence of usual appearance or motion attributes in unusual locations or times.*

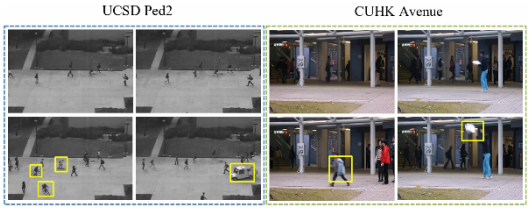


Fig1. Examples of anomalies in videos. The first row shows normal events while the second row shows abnormal events like driving vehicles on the sidewalk and throwing papers.

How does VAD Work?

One characteristic of anomalies is the low probability of occurrence, which requires much effort in collecting anomalies for constructing a supervised dataset. Therefore, the VAD task is usually regarded as an out-of-distribution (OOD) detection, training a model to fit the distribution of normal patterns only on normal data. At the test time, the abnormal events will deviate from the learned distribution. From how to fit the distribution of regular patterns, the existing VAD methods can be roughly divided into three categories: reconstruction-based, prediction-based and classification-based methods. Reconstruction-based[4][5][6] and prediction-based[7][8][9] methods measure the model's ability to fit the normal distribution by the frame reconstruction and prediction quality. With this principle, those methods regard the reconstruction/prediction errors between the reconstructed/predicted frames and their ground truths as the anomaly scores to quantify the extent of abnormalities. Considering that classification is the nature of abnormal judgment, normal or abnormal, some classification-based VAD methods have been proposed[10][11][12]. As it is expensive to collect data with abnormal annotations, classification-based methods have two paradigms. One is to train a one-class classifier using only normal data, and the other is to build anomaly hypotheses and generate pseudo-anomalous data for training binary classification models.

Why We Need VAP?

Despite great success, the VAD task is not enough for some situations with high-impact events, such as traffic accidents or terrorist attacks. To this end, we introduce the VAP task. Instead of detecting anomalies that have occurred as VAD does, VAP aims to make abnormal judgments in advance for events that have not happened at the current time. If we can make an early warning before the anomalous event occurs based on the trend of the event, it is of great significance to prevent dangerous accidents and avoid loss of life and property. However, since there are no ground truths in the VAP task, reconstruction-based and prediction-based VAD methods that rely on ground truths to calculate anomaly scores cannot solve VAP. Moreover, classification-based VAD methods tend to classify the current input rather than encourage learning feature representations of the future which is the key to VAP. In addition, there remain two main challenges to handling VAP: i) Anomalies are difficult to conform to the expectation directly because of their unbounded and rare characteristics. ii) Due to the spatial-temporal consistency, it is tough to obtain reliable corresponding semantic representations for multi-frame VAP through multi-frame prediction. Inspired by human cognition, humans have corresponding memories to judge whether future behaviours conform to the normality of the current scene. Besides, the work of

Song et al in Science found a 93% potential predictability in human behaviour [13]. To this end, we proposed a semantic pool-based VAP framework, a new baseline, which learns a semantic pool to memorize the normal semantic patterns at training time. At test time, the future frame is abnormal when its semantic feature does not belong to the trained semantic pool[14].

How to Solve VAP?

A. Problem Statement

To have a clearer picture of the VAP task, we emphasize the difference from the prediction-based VAD here. Instead of VAD leveraging the previous frames to predict the current frame to calculate the anomaly score on the frame level with its ground truth, VAP aims to obtain the semantic feature representation of the future frame to calculate the anomaly score on the feature level.

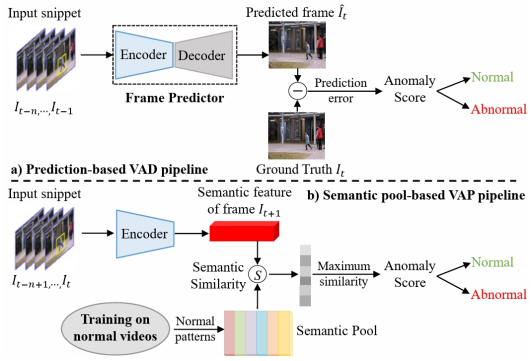


Fig.2 The pipeline of classical VAD and our introduced VAP. Best viewed zoom in.

As shown in Fig.2, we assume that the current moment is at time *t*. The superiority of VAP is that it can make an abnormal judgment on future time *t*+1, even though there are no ground truths. Fig.2 a) shows the pipeline of VAD, given the input snippet with consecutive frames ($I_{t-n,\cdots,}I_{t-1}$), we stack all these frames across the channel and send them into the frame predictor to predict the current frame I_t. Then, the prediction error between I_t and predicted $\hat{I}_t$ is calculated and used to make an abnormal judgment of time *t*. Differently, given the input snippet ($I_{t-n+1,\cdots,}I_t$) for the VAP task, as shown in Fig.2 b), we encourage the encoder to learn semantic feature of the future frame and obtain a semantic pool that stores normal semantic patterns during training. At the test, the semantic pool in the VAP task plays the role of ground truth in VAD, and we calculate the similarities through the semantic feature of the target future frame I_{t+1} and the memorized patterns in the obtained semantic pool. Then, the maximum similarity score is selected to make an abnormal judgment of future time.

B. Semantic pool-based VAP: A New Baseline

According to the pipeline of VAP, there are two key factors: future semantic learning and semantic pool building. As shown in Fig.3, our proposed baseline model mainly consists of two Channel-selected Shift Encoder (CSE), two Multiple Frames Prediction modules (MFP), a Semantic Pool Building Module (SPBM), and two kinds of constraints, prediction loss and Semantic Similarity Loss (SSLoss). Note that the CSE and two kinds of constraints are designed for future semantic learning. The SPBM is applied to memorize the normal semantic patterns.

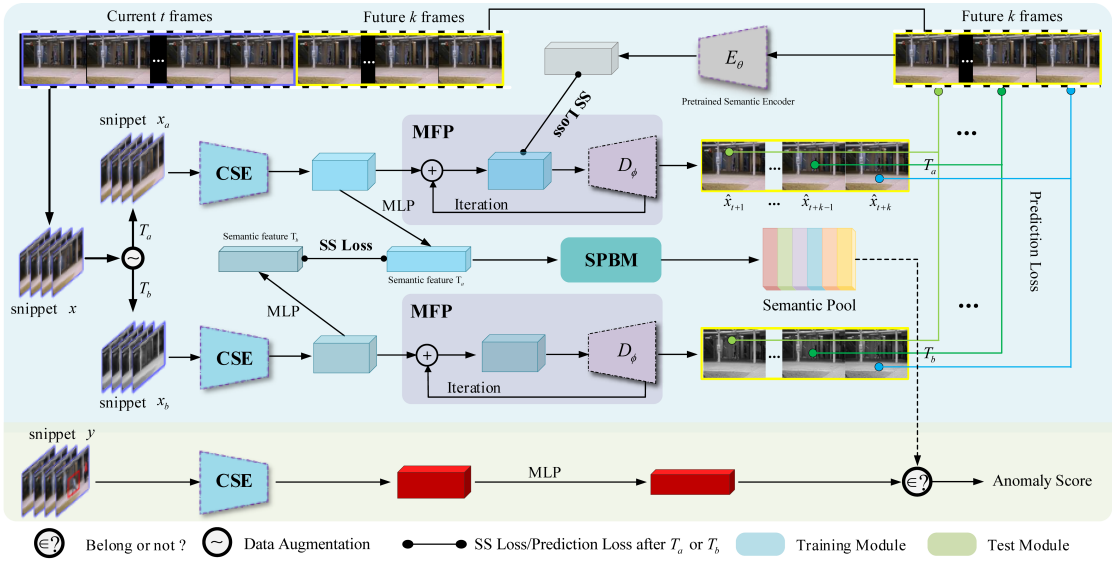
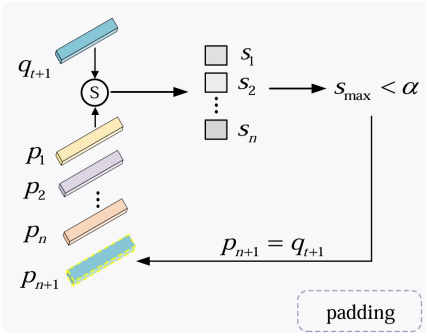


Fig. 3 Overview of our proposed baseline for the VAP task. At the training stage, we utilize CSEs, MFPs, an SPBM, a pre-trained semantic encoder, and two kinds of constraints, prediction loss and SSLoss, to learn the semantic features of future frames and build a robust dynamic semantic pool that memorizes the semantic patterns of future frames. At the test stage, the future frame is abnormal when its semantic feature does not belong to the semantic pool obtained in the training stage.

Future Semantic Learning. To extract temporal information in videos, we put forward a novel encoder based on TSM[15], called CSE. According to TSM, temporal information can be modelled by shifting the channels along the temporal dimension. Differently, considering the characteristics of the video anomalies, we shift channels with large feature changes along the temporal dimension to reduce the influence of constant background information and focus on the areas with large changes in motion, which have a high correlation to anomalies. Besides, we introduce the SSLoss, maximizing the semantic agreement of the two semantic features, to guarantee that the output of CSE represents the semantic representations of the future frame.

Semantic Pool Building. In our work, we aim to establish a semantic pool from normal videos. In our semantic pool, each item represents a semantic pattern of normality. As shown in Fig.4, our SPBM performs padding and updating the items.

The padding strategy aims to select semantic patterns, which are not similar to the memorized items. Based on this padding strategy, we store different semantic patterns of normal data, which considers the diversity. The updating strategy wants to find a common semantic representation between different normal semantic patterns so that we can further save the capacity and complexity of the semantic pool. Based on this, we consider the consistency between different normal data through feature fusion.



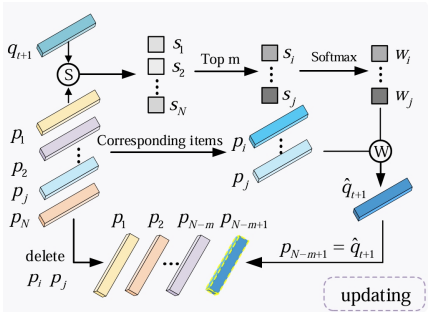


Fig.4 Details of padding and updating strategies.

C. Superiority of VAP

As shown in Fig.5, there are two test clips with a total of 5 frames, and we assume that the 6-th frame has not happened yet. For each frame, the number on the top and under the bottom denote its label and frame index respectively. Note that 0 and 1 denote abnormal and normal frames, respectively. Existing VAD methods like MNAD-P[8] can only detect anomalies in frame 174 or 561, but our VAP method can make a judgment on the future frame 175 or 562.

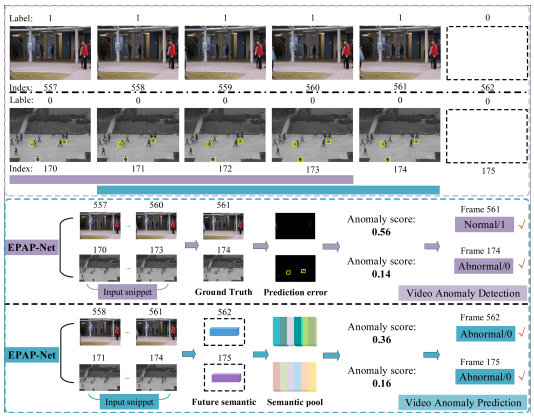


Fig.5 Difference in algorithm mechanisms between VAD and VAP. Best viewed zoom-in.

D. Multi-frame VAP

Moreover, to make the VAP task more meaningful, we design an MFP module to obtain semantic representations of future multi-frame for multi-frame VAP. We iteratively fuse the semantic features and the features from the higher layers of the decoder as the new inputs to predict the multiple future frames. Hence, we regard the features after feature fusion as the corresponding semantic representations of multiple future frames to make abnormal judgments.

What is the Next for VAP?

The significance of VAP is that we can receive an anomaly warning in advance when the abnormal event has not occurred. Compared with frame-level VAP which makes advanced judgments on single or multiple future frames, Time-level VAP finds future potential anomalies earlier, which is more valuable. Besides, future events are characterized by uncertainty. Therefore, we will explore uncertain learning to handle VAP.

Reference

[1] Venkatesh Saligrama, Janusz Konrad, Pierre-marc Jodoin. Video Anomaly Identification[J]. IEEE Signal Processing Magazine, 2010, 27(5): 18–33.

[2] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vas concelos. Anomaly detection in crowded scenes[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010: 1975–1981.

[3] Cewu Lu, Jianping Shi, Jiaya Jia. Abnormal event detection at 150 fps in matlab[C] //Proceedings of the IEEE International Conference on Computer Vision, 2013:2720–2727.

[4] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, et al. Learning temporal regularity in video sequences[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 733-742.

[5] Weixin Luo, Wen Liu, Dongze Lin, et al. Video anomaly detection with sparse coding inspired deep neural networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 43(3): 1070-1084.

[6] Wenrui Liu, Hong Chang, Xilin Chen, et al. Diversity-measurable anomaly detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2023: 12147-12156.

[7] Wen Liu, Weixin Luo, Dongze Lian, et al. Future frame prediction for anomaly detection–a new baseline[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 6536-6545.

[8] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020: 14372-14381.

[9] Cheng Yan, Shiyu Zhang, Yang Liu, et al. Feature Prediction Diffusion Model for Video Anomaly Detection[C]//Proceedings of the IEEE International Conference on Computer Vision, 2023: 5504-5514.

[10] Mohammad Sabokrou, Mohammad Khalooei, Mahmood Fathy, et al. Adversarially learned one-class classifier for novelty detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 3379-3388.

[11] Muhammad Zaigham Zaheer, Jinha Lee, Marcella Astrid, et al. Old is gold: Redefining the adversarially learned one-class classifier training paradigm[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020: 14183-14193.

[12] Zuhao Liu, Xiaoming Wu, Dian Zheng, et al. Generating anomalies for video anomaly detection with prompt-based feature mapping[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2023: 24500-24510.

[13] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility[J]. Science, 2010, 327(5968): 1018–1021.

[14] Jiaxu Leng, Mingpi Tan, Xinbo Gao, et al. Anomaly warning: Learning and memorizing future semantic patterns for unsupervised ex-ante potential anomaly prediction[C]//Proceedings of the ACM International Conference on Multimedia, 2022: 6746-6754.

[15] Ji Lin, Chuang Gan, Song Han. Tsm: Temporal shift module for efficient video understanding[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 7083–7093.