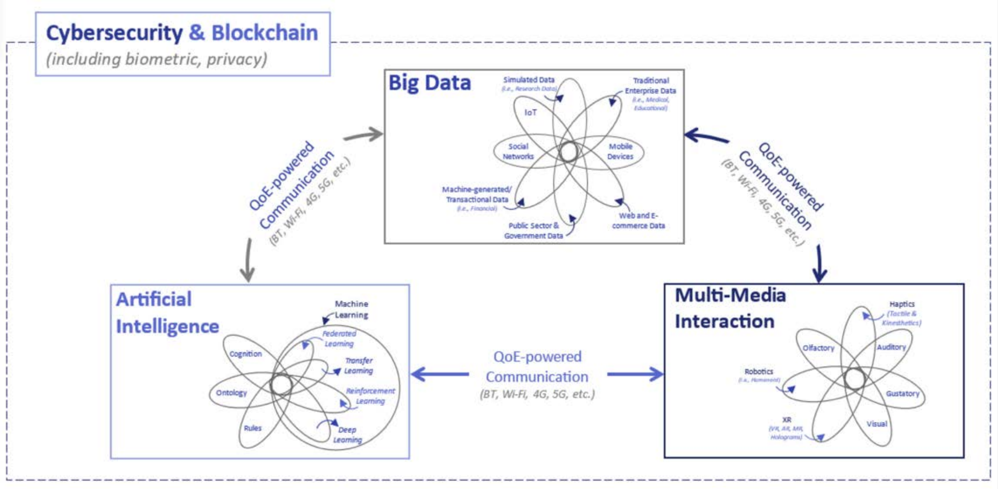
IEEE CONSUMER TECHNOLOGY SOCIETY (CTSOC)

ISSUE: FEBRUARY 2024

CTSOC-NCT NEWS ON CONSUMER TECHNOLOGY



©2023, Prof.A. El Saddik, elsaddik@uottawa.ca., reproduced with permission

THE INTEGRATION OF CHATGPT WITH THE METAVERSE FOR MEDICAL CONSULTATIONS

TABLE OF CONTENTS

EDITOR'S NOTE

As a representative of the Editorial Board of IEEE CTSoc News on Consumer Technology (NCT), it is my pleasure to present the February 2024 edition.

Our cover story highlights the pioneering integration of ChatGPT with the Metaverse for medical consultations, a concept initially featured in IEEE Consumer Electronics Magazine. This groundbreaking application promises to transform healthcare delivery, expanding access and engaging patients in unprecedented ways by uniting advanced language models with the Metaverse for healthcare services.

Following, we offer an engaging interview with the research team of Multimensional Insight Lab, under the esteemed leadership of Prof. Sanghoon Lee at Yonsei University in South Korea. The team introduces their cutting-edge Human Multi Dimensional Insight, a concept that broadens our understanding of interaction within digital domains.

We conclude this issue with an insightful article from Prof. Xinbo Gao and his team at Chongqing University of Posts and Telecommunications. They discuss an emerging paradigm shift in research, moving from Video Anomaly Detection to the predictive realms.

Enjoy the insights and inspiration within these pages!

Wen-Huang Cheng
Editor-in-Chief

EDITOR: RONG CHAO

ARTICLE TITLE

The Integration of ChatGPT With the Metaverse for Medical Consultations

AUTHOR(S)

A El Saddik, S Ghaboura

JOURNAL TITLE

IEEE Consumer Electronics Magazine

JOURNAL VOLUME AND ISSUE

Volume: 13, Issue: 3

DATE OF THE ARTICLE

October 2023

PAGE NUMBERS FOR THE ARTICLE

06 – 15

Recent years have seen a promising synergy between healthcare and the Metaverse, leading to the development of virtual healthcare environments. This convergence offers accessible and immersive healthcare experiences and has the potential to transform the delivery of medical services and enhance patient outcomes. However, the reliance on specialist presence in the Metaverse for medical support remains a challenge. Meanwhile, the newly launched large language model chatbot, ChatGPT by OpenAI, has emerged as a game-changer, providing human-like responses and facilitating interactive conversations. By integrating this cutting-edge language model with the Metaverse for medical purposes, they aim to revolutionize healthcare delivery, enhance access to care, and increase patient engagement. The study proposes a new medical Metaverse model utilizing GPT-4 as a content creator, highlighting its potential, addressing challenges and limitations, and exploring various application fields. They conclude by outlining their ongoing efforts to transform this concept into a practical reality.

INTERVIEW WITH RESEARCH TEAM OF MULTIDIMENSIONAL INSIGHT LAB IN YONSEI UNIVERSITY, SOUTH, KOREA

Lab website : http://insight.yonsei.ac.kr/gnuboard/
Youtube link : https://www.youtube.com/channel/UC68Rvsp15g1xReaKt86QnHA/

Professor Sang Hoon Lee

School of Electrical and Electronic Engineering, College of Medicine, Department of Radiology, Yonsei University, Korea

Sang Hoon Lee is a professor at School of Electrical and Electronic Engineering, Yonsei University, South Korea. Prof. Lee established Multi Dimensional Insight Laboratory in 2003. Recent research interests are Image/Video Quality Assessment, Artificial Intelligence Multi-Modal Signal processing, 3D Multi-Camera Systems, 3D reconstruction, Human Avatar Modeling and Synthesis Assessment, 3D Face/Pose/Motion/Cloth Synthesis Assessment

Human Multi Dimensional Insight

Our lab not only engaged in regular weekly meetings and seminars, but we also pertake in a variety of leisure activities beyond research. We organize a membership training (MT) in winter and summer. While skiing and having tea time at the top of the mountain, we helped and taught each other, which served as a great opportunity for the researchers to become closer. Playing basketball allowed us to foster a sence of teamwork. Eventually, aiming for a relaxed atmosphere where conversations are not strictly confined to work.

Research Team : (Left to right) Prof. Lee, Moonkyeong Choi, Sangwoo Seo, Hyunse Yoon, Jaekyung Kim, Jungwoo Huh, Mingyu Jang, Seokkeun Choi, Hoseok Tong, Jungsu Kim, Seongjean Kim, Yeseung Park, Hyucksang Lee, Seonghwa Choi, Jeonghaeng Lee, Kyungjune Lee, Seongmin Lee, Suwoong Heo

Your lab's name is quite unique. Can you tell us the origin of its name?

Our lab was first established in 2003. Initially, the name of the lab was the Wireless Network Laboratory. In the early 2000s, OFDM/MIMO-based 3rd/4th generation communications were leading the industry, and Yonsei University, in collaboration with Samsung Electronics on a long-term project, focused extensively on developing core technologies, which our lab also participated in. Therefore, our main research areas were OFDM/MIMO based radio resource management, network protocol design, and cross-layer optimization between multimedia and wireless networks. Starting from 2010, communication technology showed excellent performance, approaching the theoretical link capacity of Shannon&#39;s theory, which led us to shift our research focus to the increasingly important area of multimedia data. In fact, with my background as a Ph.D in image and video processing, it was relatively easier for us to approach the field of image analysis based on visual perception. One of the research areas was Quality-of-Experience (QoE) studies focusing on the visual fatigue of 3D stereoscopic images and videos. Although we faced challenges in migration, the advent of deep learning technology later made it a great junction for our current field of processing 3D data by combining computer graphics and computer vision. The gradual migration to AI technology occurred in 2015, and it became unnecessary to change the lab's name. We decided on our lab's name to maintain flexibility due to the ever-changing flow of research, and looking back, it seems to have been a wise decision. They say change is the only constant in research... except for the coffee machine, which has been constant in its refusal to work properly (laugh).

What insights guide the research topics in your lab, and what technologies are you focusing on from a long-term perspective?

Currently, AI technology is evolving at a remarkable pace. When AI technology first emerged, we believed that once the development of algorithms reached a saturation point, the competition for data acquisition would accelerate. Now, it seems we are at that period. In response to the question of what our next target technology will be, it is creating AI that can share emotions with humans. This often feels like a technology introduced in movies, seemingly far into the future. However, with sufficient data and supportive algorithms, implementing technology necessary for life is becoming a common reality. Yet, it's when AI technology can touch human emotions that AI will truly integrate deeply into our lives.

Therefore, our lab is working to create a virtual human akin to a real person and breathe life into AI technology by synthesizing it in human form, enabling it to experience emotions that a person can feel. Acquiring data is considered one of the most crucial components for this. We are developing a camera system to acquire 4D human data and contemplating from various angles how to infuse individual emotions into it.

It seems that the endpoint of such technology will lead to a flow towards a human-centric metaverse. While the continuous development of generative AI has made it easy to create 2D contents such as images, the quality for synthesizing in the 3D domain is still insufficient for service. This indicates that the evolution of generative AI in the 3D domain and a human-centric metaverse will likely be the ongoing trend.

What was the most crucial element in pursuing these goals?

The importance of data is increasing more than ever. To address this, we have developed a camera system. Constructing the current system required a significant amount of time and effort. Among the challenges were communication between me and my students, among students themselves, and even between the school and the lab. Understanding why we need to build the system, how to construct it, and what needs to be built first required a lot of effort to narrow down thoughts and reach a consensus. Additionally, the technical approaches and research methods vary among lab members, as do their interests. Overcoming these differences and deciding what to prioritize and what to place as a secondary concern required much deliberation.

Another crucial element is vision for the future. With the camera system we have constructed so far, the results have visibly improved, we have started to acquire visually pleasing high-quality data. At this point, we are contemplating how to organize the results and how to persuade others with our findings in a research paper. I believe that organizing these results and allowing individuals to develop their vision through this process could be a driving force for further research progress.

Besides research, what kind of activities are conducted in your lab?

Our lab not only engages in regular meetings and seminars related to research and development but also participates in various leisure activities beyond research. First and foremost, at the end of each semester, we go on a 2 nights and 3 days membership training in winter and summer. During this period, lab members are encouraged to share not only about their research but also personal updates and goals, or to brainstorm creative research topics in a relaxed atmosphere. Moreover, a unique and advantageous aspect of our lab is that, in addition to me, many students have a deep appreciation for music. We have been active like a small club, utilizing our strengths in composition, piano, guitar, vocals, etc., to create music together. Recently, we went beyond just creating songs to producing final tracks and releasing them on music sites. If you're interested, please visit our website. We warmly welcome support through the purchase of our music. It won't be that expensive (laugh). Lastly, when the lab schedule is relatively free or there's something to celebrate, we also have group meals at restaurants near the campus.

The reason we engage in such a variety of activities beyond research is based on the premise that, in addition to professional teamwork, the interpersonal relationships formed in lab life are also very precious. Aiming to brighten the lab atmosphere and achieve better results, we plan to continue these activities in the future.

What message would you like to impart to the younger generation?

Sometimes, when I talk to students working with AI, I liken it to the flow of water in a stream. The flow is fastest at the surface, between the water and the air, and slows down as you go deeper, due to the friction with the bottom. This surface speed can be compared to the trend-following implementation techniques of AI algorithms. Without analyzing how things work on the inside, one risks being swept away aimlessly by the strong current. To counteract this, one must row deeper or stir the bottom to propel the boat forward, requiring much hidden effort. Additionally, collaborating with fellow students to establish a technological foundation is essential. Otherwise, time may be wasted aimlessly. It's crucial to assess whether one can withstand such rapid currents. If you prefer to proceed calmly, laying a solid foundation with a long-term perspective, it's important to delve into slower currents at the bottom, where one isn't swept away and can research at their own pace. For instance, if you can acquire data, you can gain insights from it and, from there, find clues to human emotions.

I believe that rather than being dazzled by the high-performing algorithms created by major AI corporations, maintaining your pace and building your research from the ground up will lead to positive outcomes.

What message do young generations wish to convey to professors and senior researchers?

Young generations learning and engaging with AI feel that the field is changing very rapidly and sensitively. New technologies and papers are emerging quickly, and trends shift rapidly each year. They invest a lot of time in keeping up with and understanding these trends and request more time to concentrate on these trending technologies. While learning new software and trends is an exciting process, merely following them can sometimes cause us to lose sight of the core of our research. Placing importance on running and learning deep learning software may lead to forgetting what our actual research goals were. I believe that finding the core of one's research and setting the direction for it is a truly challenging process. It's like wandering around in a forest, thinking you are walking straight when, in fact, you might be lost. At this point, the guidance of senior researchers in establishing the core of research and providing direction can be immensely helpful. Tasks like code review and bug fixing can be sufficiently handled among peers, but the grand task of defining the core and direction of research benefits greatly from the experience of senior researchers. Even if the passion for learning leads young generations astray, it seems beneficial to engage in discussions with the senior generations more frequently to create positive synergy.

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.*

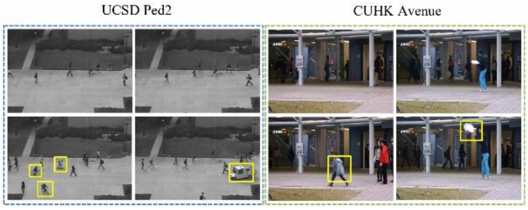


Fig1. Examples of anomalies in videos. The first row shows normal events while the second row shows abnormal events like driving vehicles on the sidewalk and throwing papers.

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.

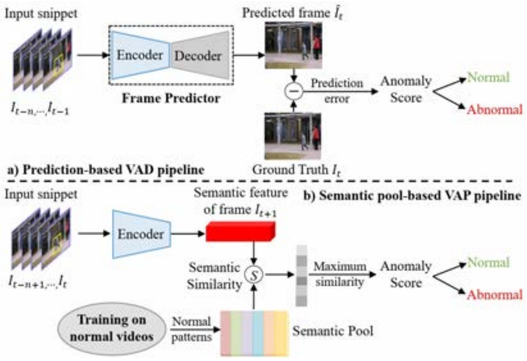


Fig.2 The pipeline of classical VAD and our introduced VAP. Best viewed zoom in.

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.

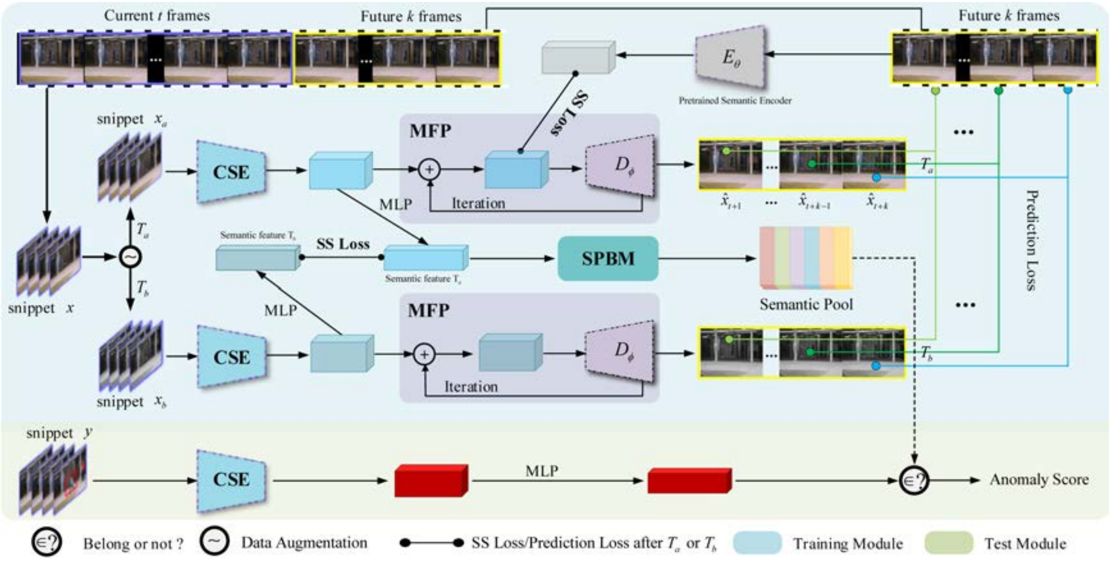
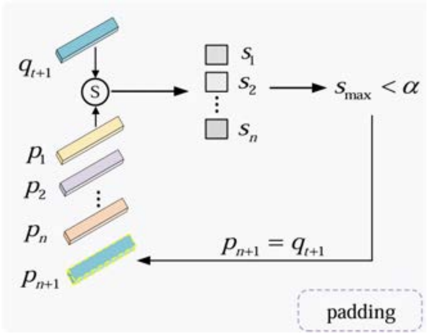


Fig. 3 Overview of our proposed baseline for the VAP task. At the training stage, we utilize CSEs, MFPs, an SPBM, a pre-trained semantic encoder, and two kinds of constraints, prediction loss and SSLoss, to learn the semantic features of future frames and build a robust dynamic semantic pool that memorizes 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.



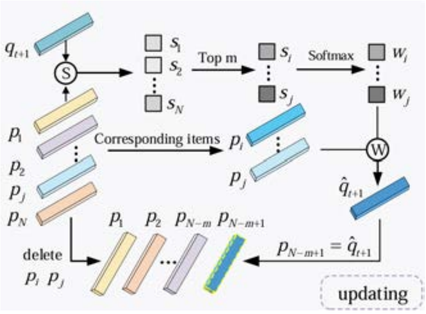


Fig.4 Details of padding and updating strategies.

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.

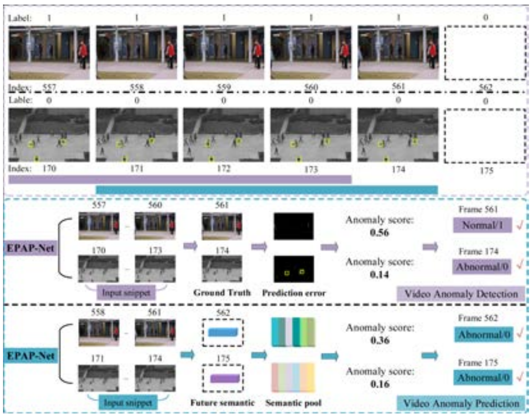


Fig.5 Difference in algorithm mechanisms between VAD and VAP. Best viewed zoom-in.

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.