DETECTING ANOMALOUS EVENTS FROM UNLABELED VIDEOS VIA TEMPORAL MASKED AUTO-ENCODING

Jingtao Hu[1,†], *Guang Yu*[1,†], *Siqi Wang*[1,*], *En Zhu*[1,*], *Zhiping Cai*[1] *and Xinzhong Zhu*[2]

[1]National University of Defense Technology, China; [2]Zhejiang Normal University, China
Email: {hujingtao17, guangyu, wangsiqi10c, enzhu, zpcai}@nudt.edu.cn, zxz@zjnu.edu.cn

2022 IEEE International Conference on Multimedia and Expo (ICME) | 978-1-6654-8563-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICME52920.2022.9859873

ABSTRACT

Unsupervised video anomaly detection (UVAD) intends to discern anomalous events from fully unlabeled videos. However, existing UVAD methods suffer from poor performance. Inspired by recent masked autoencoder (MAE) [1], we propose Temporal Masked Auto-Encoding (TMAE) as an effective end-to-end UVAD method. Specifically, we first denote video events by spatial-temporal cubes (STCs), which are built by temporally consecutive foreground patches from unlabeled videos. Then, half of patches in an STC are masked along the temporal dimension, while a vision transformer (ViT) is trained to exploit unmasked patches to predict masked patches. The rare and unusual nature of anomaly will result in a poorer prediction for anomalous events, which enables us to discriminate anomalies from unlabeled videos and compute the anomaly scores. Furthermore, to utilize motion clues in videos, we also propose to apply TMAE on optical flow, which can further boost performance. Experiments show that TMAE significantly outperforms existing UVAD methods by a notable margin (3.9%–6.6% AUC).

Index Terms— Unsupervised video anomaly detection, masked autoencoder

1. INTRODUCTION

Video analysis is an important subarea in multimedia community. Among the topics of video analysis, video anomaly detection (VAD) [2] intends to detect anomalous events in videos that diverge from the frequently-occurred routine, which can bring tremendous benefits for various scenarios such as social emergency management. However, VAD has been a challenging problem. This is due to the difficulty of modeling high-dimensional and complex video data, and the rare, unusual and ambiguous nature of anomaly, i.e., anomalies are highly unpredictable events that have vague semantics and a low probability of occurrence. Therefore, it is unpractical to collect anomalous data for modeling, which makes the supervised classification paradigm inapplicable for VAD.

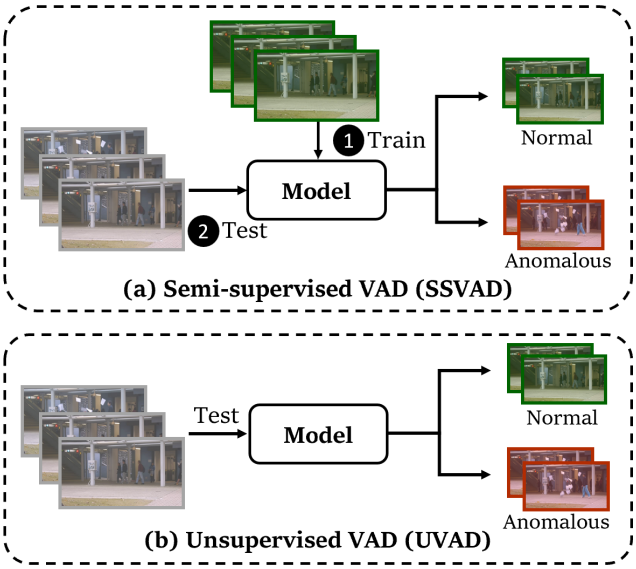

Fig. 1: Contrast of SSVAD and UVAD.

Consequently, most existing VAD methods follow the ***semi-supervised*** setup (see Fig. 1(a)), which requires pure normal videos to build a normalcy model. Then VAD is realized by judging whether the testing video data conform to this model during inference. Although the semi-supervised VAD (SSVAD) is a feasible paradigm that avoids collecting anomalous data, it can be still labor-intensive and time-consuming to label the collected videos for building a training set that contains only normal videos. Therefore, a promising alternative to SSVAD is ***unsupervised*** VAD (UVAD) (see Fig. 1(b)), which intends to discern anomalous events from fully unlabeled videos, so as to avoid the heavy labeling burden.

Although UVAD has emerged in the literature, existing UVAD methods still suffer from sub-optimal video representations and limited modeling power. To represent videos, most UVAD methods [3–5] still involve hand-crafted feature descriptors like 3D gradients to extract video event features. However, hand-crafted features often suffer from sub-optimal discriminative ability, which can be the bottleneck of UVAD performance. Meanwhile, hand-crafted feature engineering can be exhausting and inflexible when faced with different video scenes. As for modeling, most UVAD methods [3–5]

† Equal contributions
* Corresponding authors

rely on a classic model like logistic regression to perform learning and detection. Nevertheless, in many cases classic models often underperform deep neural networks (DNN) in modeling capability. The most recent work [6] for the first time utilizes a DNN to realize a self-trained deep regression for better modeling and anomaly scoring, but it still requires initial detection results from a classic model like isolation forest [7] to guide the learning of the DNN at the early stage. The initial results obtained by classic models may mislead the subsequent learning. Due to the above two limitations, existing UVAD methods suffer from poor VAD performance. For an intuitive comparison, existing UVAD methods are still significantly inferior to recent state-of-the-art SSVAD methods by about 5%–10% AUC on commonly-used VAD datasets.

As both hand-crafted feature descriptors and classic models can be less expressive in UVAD, high-quality video representations and strong modeling capability will be the key to effective UVAD. To this end, we notice that a newly proposed self-supervised learner, i.e., masked autoencoder (MAE) [1], can serve as a powerful and promising solution to unsupervised representation learning in UVAD. The core idea of MAE is to mask a part of patches on images and then train a DNN to recover the missing patches from unmasked patches. With the efficient learning paradigm of MAE, high-quality representations can be obtained for downstream recognition tasks. Meanwhile, as a DNN based model, MAE itself possesses strong modeling capabilities. Thus, a natural idea for UVAD is to utilize MAE to realize simultaneous representation learning and modeling of video events in an end-to-end manner. However, since MAE is designed to learn image representations by applying masking on the spatial dimension of 2D images, it does not consider the temporal information, which actually plays a pivotal role in videos. Besides, anomalous events often appear as temporal context anomalies.

Inspired by MAE, we propose a novel method named Temporal Masked Auto-Encoding (TMAE) for effective and end-to-end UVAD. To be more specific, we first localize video foreground and extract temporally consecutive foreground patches to build spatial-temporal cubes (STC), which represent events in videos and act as the basic processing units. Then, unlike MAE that applies masking on the spatial dimension, we mask half of the foreground patches along the temporal dimension in an STC. Finally, we train a vision transformer (ViT) to predict masked patches with unmasked patches. Due to the rare and unusual nature of anomaly, anomalous events tend to generate larger prediction errors, which enables us to directly utilize the prediction errors as anomaly scores to discriminate anomalies from unlabeled videos. Since motion provides valuable clues for VAD, we additionally propose to apply TMAE on the corresponding optical flow of STCs for more effective UVAD. Extensive experimental results on commonly-used VAD datasets show that TMAE significantly outperforms state-of-the-art UVAD methods by an evident margin (3.9%–6.6% AUC).

2. RELATED WORK

Semi-supervised VAD (SSVAD). The majority of efforts formulate VAD as a semi-supervised learning task, where anomalous video events are not available in the training process. Traditional SSVAD methods usually use hand-crafted descriptors (e.g. trajectory [8], histogram of gradient (HoG) [9], histogram of optical flow (HoF) [10], 3D gradients [11], etc.) to extract normal patterns at first, which is a necessary step before model training. With the rapid development of deep learning, plentiful works have made fruitful progress by integrating DNN into SSVAD to realize simultaneous representation learning and video event modeling. This promising end-to-end strategy stimulates researchers to explore various DNN architectures for SSVAD, such as LSTM [12], U-Net [13], GAN [14], predictive autoencoder [15] and so on. Readers can refer to [2] for a survey of recent SSVAD methods.

Unsupervised VAD (UVAD). Compared with SSVAD, fewer works are devoted to UVAD. The groundbreaking work [3] operates on shuffled hand-crafted video event data and considers dramatic changes as anomalous activities. [4, 5] improves the change detection in [3] by using the unmasking technique [16] to enhance the performance. Different from aforementioned change detection paradigm in essence, Pang et al. [6] address UVAD problem by firstly obtaining the initial detection results from a pre-trained DNN and an isolation forest [7]. Then they promote the performance based on self-trained ordinal regression.

3. METHODOLOGY

As introduced in Sec. 1, MAE [1] has been proposed as a astonishingly successful self-supervised learner, which can conduct effective unsupervised representation learning. Thus, we are motivated by MAE, and propose TMAE to learn high-quality video representations and accomplish end-to-end UVAD. To our best knowledge, this is the first work to tailor MAE for VAD. The overall process of TMAE is shown in Fig. 2, which mainly consists of the following three steps:

3.1. Video Event Extraction

The localization and extraction of video events prove to be pretty important for VAD [17, 18], which can make the subsequent modeling focus on the meaningful foreground objects rather than the irrelevant background of video frames. Following the video event extraction scheme in [18, 19], we first perform localization on each video frame to obtain foreground objects. Then, for each object, D foreground patches are extracted according to its location from the current and $D-1$ temporally consecutive frames. Finally, we resize the D patches into $x_i \in \mathbb{R}^{H\times W}(i = 1, 2, \ldots, D)$, and stack them into a spatial-temporal cube (STC) $\mathbf{C} \in \mathbb{R}^{H\times W\times D} =$

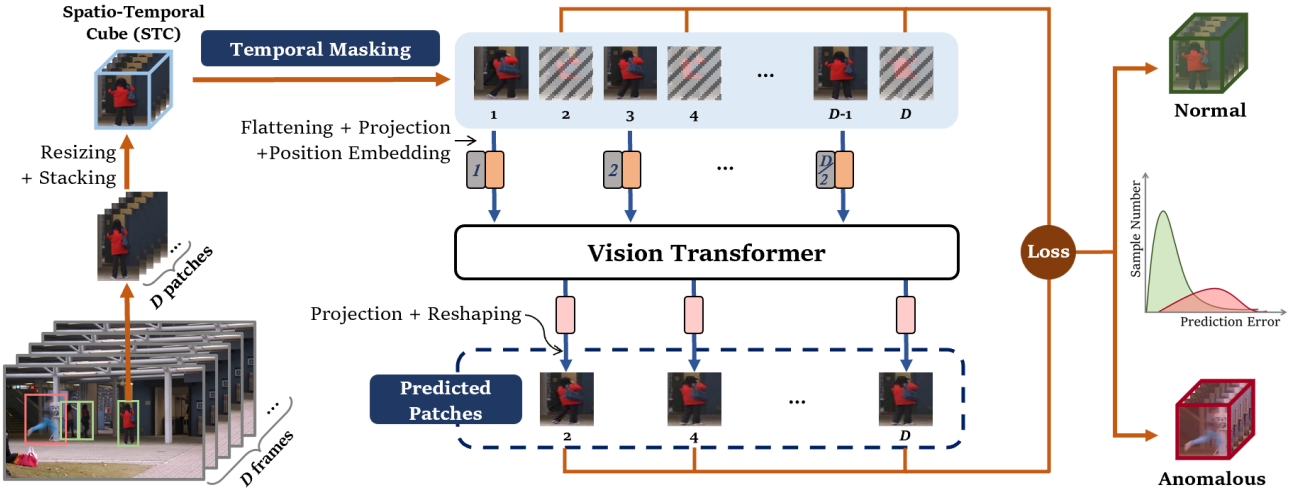

Fig. 2: The proposed UVAD method framework. Given a set of unlabeled videos, we first **(1)** localize foreground and extract sequential foreground patches at the same location to build precise video events (i.e. STCs), and then **(2)** mask the patches along the temporal dimension in an STC and feed the rest of them into ViT to predict the invisible patches. We finally **(3)** utilize prediction error as anomaly score to discriminate anomalous video events in a strictly unsupervised manner.

$[x_1, x_2, \ldots, x_D]$ (the number of channels is ignored). In this way, the STC $\mathbf{C}$ can represent a video event more accurately and serves as the basic processing unit in UVAD. The whole process is shown in the left part of Fig. 2.

3.2. Temporal Masked Auto-Encoding

The core component of our approach is the proposed temporal masked auto-encoding (TMAE) scheme. It is noted that the major hindrance in UVAD is the full absence of supervision information. In light of this, we naturally attempt to introduce MAE into UVAD, so as to learn effective video representations by DNNs under the fully unsupervised scenario. However, the original framework of MAE is designed for learning 2D image representations along the spatial dimension, which cannot exploit the vital temporal information in videos. To enable effective representation learning from videos, we propose the novel TMAE scheme, which encourages DNN to learn informative temporal patterns in video events by masking on the temporal dimension of STCs.

Specifically, given an STC $\mathbf{C} = [x_1, x_2, \ldots, x_D]$, we first mask half of foreground patches along the temporal dimension in the STC. It should be noted that we apply masking at the patch-level, i.e., a patch is either fully masked or fully visible. To mask half of the patches, we explore three different masking strategies: 1) Interval masking: The patches at the even temporal positions in the STC are masked, e.g., $\{x_i | i = 2, 4, \ldots, D-2, D\}$ are masked and the remaining patches $\{x_i | i = 1, 3, \ldots, D-3, D-1\}$ serve as the input of DNN. 2) Block-wise masking: The former or latter half of the consecutive patches are masked. 3) Random masking: Masking is applied at random temporal positions in the STC. We adopt interval masking as the default masking strategy.

After we apply masking in the STC, we then train a DNN that takes the unmasked patches as input to predict the masked patches for high-quality video representation learning, which is similar to the learning paradigm of MAE. Following MAE, we also choose vision transformer (ViT) [20] as our DNN backbone, which is due to the two considerations: 1) As a newly-emerging model in computer vision, ViT has shown powerful modeling capabilities. 2) ViT is naturally eligible for long-range temporal data like STCs. Concretely, for the input of ViT, i.e., the unmasked patches $\{x_i \in \mathbb{R}^{H \times W} | i = 1, 3, \ldots, D-3, D-1\}$, we first flatten them to 1D tokens $\{t_j \in \mathbb{R}^{(H \cdot W)} | j = 1, 2, ..., \frac{D}{2}\}$, and map each token into a low-dimension embedding $t_j^{'} \in \mathbb{R}^{dim}$ by a trainable linear projection. To retain temporal position information in the patch sequence, we add learnable positional embeddings $E_{pos} \in \mathbb{R}^{\frac{D}{2} \times dim}$ to the token embeddings and obtain the following token sequence $Y^{(0)}$:

$$Y^{(0)} = [t_1^{'}; t_2^{'}; ...; t_{\frac{D}{2}-1}^{'}; t_{\frac{D}{2}}^{'}] + E_{pos} \tag{1}$$

Then we feed $Y^{(0)}$ into L consecutive identical blocks of ViT. Each block performs the same computing process and consists of three modules, i.e., multi-head self-attention (MSA), layer normalization (LN) and multi-layer perception (MLP). Take the l^{th} block as an example, it takes the output $Y^{(l-1)}$ of the previous $(l-1)^{th}$ block as input to compute its output $Y^{(l)}$:

$$\begin{aligned} \tilde{Y}^{(l)} &= MSA(LN(Y^{(l-1)})) + Y^{(l-1)}, \\ Y^{(l)} &= MLP(LN(\tilde{Y}^{(l)})) + \tilde{Y}^{(l)}, \end{aligned} \tag{2}$$

Finally, we project each item $y_j \in \mathbb{R}^{dim}$ in the output tokens $Y^{(L)}$ to the original dimension of the patches $y_j^{'} \in$

$\mathbb{R}^{(H \cdot W)}$ and reshape it into the initial size $x_j^{'} \in \mathbb{R}^{H \times W}$ as the prediction of masked patches. The generated STC $\mathbf{C}_{pred} = [x_1^{'}, x_2^{'}, ..., x_{D/2}^{'}]$ is regarded as the prediction of $\mathbf{C}_{mask} = [x_2, x_4, ..., x_{D-2}, x_D]$, which is the complement set of input $\mathbf{C}_{unmask} = [x_1, x_3, ..., x_{D-3}, x_{D-1}]$. Finally, the ViT is trained by minimizing prediction error. We adopt the mean square error (MSE) loss as the prediction loss.

3.3. Anomaly Scoring and Performance Boosting

After we perform TMAE with extracted video events, we need to discriminate anomaly by calculating the anomaly score for each event. A straightforward solution is to feed learned representations into a classic anomaly detection model like isolation forest. However, as we discussed in Sec. 1, such classic models can be sub-optimal, while the end-to-end anomaly scoring is more favorable. To discriminate anomalous events, we notice two important observations in practice: 1) Compared with normal video events, the abnormal ones occupy a much smaller proportion in videos. 2) The patterns of normal events are usually simpler and more predictable, while the abnormal ones tend to be more complex and irregular. Therefore, such rare and unusual nature of anomaly will make it more difficult for ViT to conduct masking and prediction on abnormal events in TMAE, which can be reflected by the prediction errors in training. As a result, the prediction errors of TMAE will be discriminative and enable us to distinguish anomalies in a convenient end-to-end manner.

For each STC, the anomaly score is measured by pixel-wise MSE loss, which can be calculated as follows:

$$MSE(\mathbf{C}) = \|\mathbf{C}_{mask} - \mathbf{C}_{pred}\|_2^2 \tag{3}$$

Since the temporal patterns are essentially embedded in an object's motion, we further utilize dense optical flow [21] as *auxiliary motion clues* to boost TMAE. Optical flow describes the pixel-level motion between two adjacent frames in videos, which can be calculated through a pre-trained DNN model like [22]. Therefore, given the optical flow maps of video frames, we can similarly extract motion clue cubes (MCCs) $\hat{\mathbf{C}} \in \mathbb{R}^{H \times W \times D}$ and feed the unmasked part of MCCs to the same pipeline above as the original video data.

Considering both original video frames and optical flow maps, the final anomaly score of a given $\mathbf{C}$ is defined as:

$$Score(\mathbf{C}) = \alpha \cdot \frac{MSE(\mathbf{C}) - \mu}{\sigma} + \beta \cdot \frac{MSE(\hat{\mathbf{C}}) - \hat{\mu}}{\hat{\sigma}} \tag{4}$$

where μ and σ denote the mean and standard deviation of all STCs' MSE losses; $\hat{\mu}$ and $\hat{\sigma}$ denote the mean and standard deviation of all MCCs' MSE losses; α and β are hyper-parameters to measure the importance of the two parts. Besides, we take the maximum score of all STCs on a frame as the anomaly score of this frame for evaluations. It is worth noting that the TMAE is an end-to-end unsupervised method that directly detects anomalies from the testing set, i.e. an offline transductive solution. Therefore, computing $\mu, \sigma, \hat{\mu}, \hat{\sigma}$ is plausible for TMAE. Besides, we do not report running time here because the training and inference are actually one learning process for the transductive TMAE solution.

3.4. Comparison with State-of-the-art Methods

In this section, we compare TMAE method with existing UVAD solutions. Meanwhile, we also include representative SSVAD methods as a reference. The results are shown in Table 1.

Table 1: AUC comparison with UVAD and SSVAD methods. MC is the abbreviation of motion clues.

	Method	ped1	ped2	CUHK	SHTech
SSVAD	CAE [23]	81.0%	90.0%	70.2%	-
	ConvLSTM-AE [12]	75.5%	88.1%	77.0%	-
	SRNN [24]	-	92.2%	81.7%	68.0%
	Recounting [25]	-	92.2%	-	-
	Frame-Prediction [26]	83.1%	95.4%	85.1%	72.8%
	Att-prediction [27]	83.9%	96.0%	86.0%	-
	Mem-AE [28]	-	94.1%	83.3%	71.2%
	Mem-Guided [29]	-	97.0%	88.5%	70.5%
	SRNN-AE [30]	-	92.2%	83.5%	69.6%
UVAD	Discriminative [3]	59.6%	63.0%	78.3%	-
	Unmasking [4]	68.4%	82.2%	80.6%	-
	MC2ST [5]	71.8%	87.5%	84.4%	-
	STDOR [6]	71.7%	83.2%	-	-
	TMAE w/o MC	74.7%	93.1%	88.1%	70.8%
	TMAE w MC	**75.7%**	**94.1%**	**89.8%**	**71.4%**

4. EMPIRICAL EVALUATIONS

4.1. Experimental Setup

To evaluate TMAE, we conduct experiments on four public datasets: UCSD ped1 and ped2 [31], CUHK Avenue [11], and ShanghaiTech [26]. To perform UVAD, we follow the previous works [3–5] and only use the testing set of a dataset, while labels are only used for evaluations. For quantitative evaluations, the widely-used frame-level AUC is adopted as the evaluation metric. The implementation parameters are as follows: For both STCs and MCCs, they are built in the same size with $H = W = 32$ and $D = 8$ respectively. As for ViT, we follow the original architecture [20] and set the number of blocks $L = 4$, the embedding dimension $dim = 512$. Meanwhile, the number of heads in MSA is set to 12 and the hidden dimension in MLP is 1024. ViT is optimized by an Adam optimizer in PyTorch with a learning rate of 0.0001. The batchsize is 256 and the training epoch is 100. Particularly, since ped1 dataset suffers from evident foreground

depth variation, we uniformly divide its frames into 4×1 regions and use a separated ViT to process each region. As to scoring, the coefficients α and β in Eq. (4) are set to 1 and 0.5 respectively. The final frame anomaly scores are smoothed by a sliding window with the window size of 15.

From Table 1, we can draw three conclusions as follows: **(1)** Our method attains a remarkable AUC gain of **3.9%**, **6.6%** and **5.4%** respectively on ped1, ped2 and CUHK Avenue datasets when compared with the best performer of existing UVAD methods. It is noticeable that we can also yield satisfactory performance on the challenging ShanghaiTech dataset, which has not been explored before under the fully unsupervised setting. When compared with the pioneering baseline in [3], we even obtain about 10%-30% improvement among different datasets. **(2)** Our unsupervised solution achieves highly competitive performance when compared with classic SSVAD methods. For example, on CUHK Avenue, our method accomplishes the best AUC among both UVAD and SSVAD methods in Table 1. Meanwhile, the performance of our UVAD method is comparable to the representative SSVAD method [26] on the latest ShanghaiTech dataset. **(3)** Taking *motion clues* into consideration enhances the model and brings consistent improvement of 1%, 1%, 1.7%, 0.6% on the four datasets respectively. In particular, on CUHK Avenue dataset, the involvement of motion clues enables our method to outperform recent SSVAD methods. **In conclusion, the results demonstrate the effectiveness of our novel end-to-end UVAD solution.**

4.2. Visualization

To offer an intuitive illustration, we visualize the predictions of masked patches and corresponding prediction errors. As shown in Fig. 3, for each dataset, the odd columns are representative normal video events (e.g. walking), and the even columns are anomalous events (e.g. car, bicycle, running). TMAE produces evidently better prediction for normality than anomaly, which is consistent with the expectation and facilitates anomaly detection.

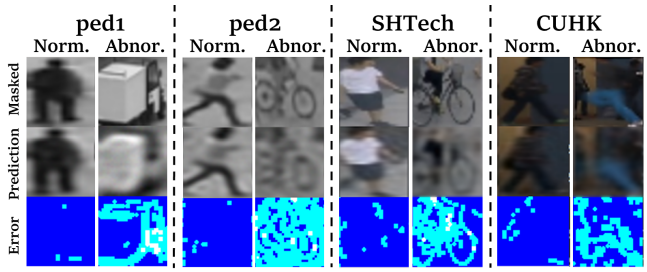

Fig. 3: Visualization the predicted masked patches.

4.3. Discussion

We first explore how different temporal masking schemes affect the performance. The temporal block-wise masking strategy is implemented by masking first half patches in STCs and MCCs, and predicting the latter half. As shown in Table 2, the results are constantly worse than our default temporal masking strategy, and the AUC drops sharply by about 5% on the ped2 dataset. We also test random temporal masking strategy on different benchmarks: Despite its superiority to block-wise masking, it is still inferior to the interval masking strategy. Such results reveal that masking with equal intervals in STCs is the most suitable way for conducting UVAD.

Table 2: AUC w.r.t. different temporal masking strategies.

Masking	ped1	ped2	CUHK	SHTech
Block-wise	74.5%	89.5%	87.2%	68.7%
Random	75.6%	**94.1%**	89.6%	68.7%
Interval	**75.7%**	**94.1%**	**89.8%**	**71.4%**

Then, we also analyze the parameter sensitivity of several key parameters: **(a)** The hyper-parameter relative proportion between α and β, and **(b)** The number of patches in an STC and that in a MCC. Specifically, to unveil the influence of auxiliary motion clues on the final anomaly score, we change the hyper-parameter α and β in Eq. (4) from 0.5 to 1.0. As shown in Table 3, different combinations of α and β only cause up to 1.5% fluctuations on AUC. As for the number of patches in cubes (i.e. STCs and MCCs), we can constantly obtain satisfactory performance under different cube sizes as displayed in Fig. 4, while $D = 8$ is the optimal setting for TMAE.

Table 3: Sensitivity analysis on the weight of MC.

$\alpha : \beta$	ped1	ped2	CUHK	SHTech
0.5:1.0	75.5%	93.5%	88.3%	71.4%
1.0:0.5	75.7%	94.1%	89.8%	71.4%
1.0:1.0	76.0%	94.1%	89.5%	71.5%

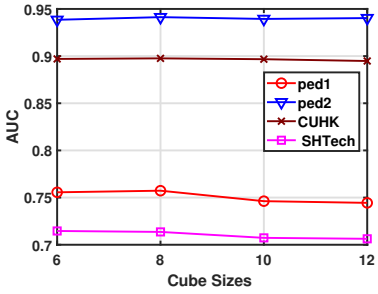

Fig. 4: Parameter sensitivity analysis on cube sizes.

5. CONCLUSION

In this paper, we make the first attempt to introduce the powerful self-supervised leaner MAE into VAD. Specifically, we

propose a novel TMAE framework that can capture temporal information by stimulating DNN to learn the temporal patterns from unlabeled video events. TMAE can learn high-quality representations and detect anomalous video events in an end-to-end manner. Compared with the state-of-the-art UVAD and SSVAD approaches, extensive experiments validate the effectiveness of TMAE. In the future, we will explore the idea of combining other effective techniques (e.g., spatial masking) with our temporal masking to further strengthen the model.

Acknowledgements. The work is supported by National Natural Science Foundation of China (62006236), NUDT Research Project (ZK20-10) and HPCL Autonomous Project (202101-15).

6. REFERENCES

[1] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Doll'ar, and Ross B. Girshick, "Masked autoencoders are scalable vision learners," *ArXiv*, vol. abs/2111.06377, 2021.

[2] Bharathkumar Ramachandra, Michael Jones, and Ranga Raju Vatsavai, "A survey of single-scene video anomaly detection," *IEEE TPAMI*, 2020.

[3] Allison Del Giorno, J Andrew Bagnell, and Martial Hebert, "A discriminative framework for anomaly detection in large videos," in *ECCV*, 2016, pp. 334–349.

[4] Radu Tudor Ionescu, Sorina Smeureanu, Bogdan Alexe, and Marius Popescu, "Unmasking the abnormal events in video," in *ICCV*, 2017, pp. 2895–2903.

[5] Yusha Liu, Chun-Liang Li, and Barnabás Póczos, "Classifier two sample test for video anomaly detections," in *BMVC*, 2018, p. 71.

[6] Guansong Pang, Cheng Yan, Chunhua Shen, Anton van den Hengel, and Xiao Bai, "Self-trained deep ordinal regression for end-to-end video anomaly detection," *CVPR*, pp. 12170–12179, 2020.

[7] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou, "Isolation forest," in *ICDM*, 2008, pp. 413–422.

[8] Claudio Piciarelli, Christian Micheloni, and Gian Luca Foresti, "Trajectory-based anomalous event detection," *IEEE TCSVT*, vol. 18, no. 11, pp. 1544–1554, 2008.

[9] Bin Zhao, Li Fei-Fei, and Eric P Xing, "Online detection of unusual events in videos via dynamic sparse coding," in *CVPR*, 2011, pp. 3313–3320.

[10] Yang Cong, Junsong Yuan, and Ji Liu, "Sparse reconstruction cost for abnormal event detection," in *CVPR*, 2011, pp. 3449–3456.

[11] Cewu Lu, Jianping Shi, and Jiaya Jia, "Abnormal event detection at 150 fps in matlab," in *ICCV*, 2013, pp. 2720–2727.

[12] Weixin Luo, Wen Liu, and Shenghua Gao, "Remembering history with convolutional lstm for anomaly detection," in *ICME*, 2017, pp. 439–444.

[13] Yutao Hu, Xin Huang, and Xiaoyan Luo, "Adaptive anomaly detection network for unseen scene without fine-tuning," in *PRCV*, 2021, pp. 311–323.

[14] Zhi Zhang, Sheng-hua Zhong, and Yan Liu, "Video abnormal event detection via context cueing generative adversarial network," in *ICME*, 2021, pp. 1–6.

[15] Yuandu Lai, R. Liu, and Yahong Han, "Video anomaly detection via predictive autoencoder with gradient-based attention," *ICME*, pp. 1–6, 2020.

[16] Moshe Koppel, Jonathan Schler, and Elisheva Bonchek-Dokow, "Measuring differentiability: Unmasking pseudonymous authors," *JMLR*, vol. 8, no. 6, 2007.

[17] Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling Shao, "Object-centric auto-encoders and dummy anomalies for abnormal event detection in video," in *CVPR*, 2019, pp. 7842–7851.

[18] Guang Yu, Siqi Wang, Zhiping Cai, En Zhu, Chuanfu Xu, Jianping Yin, and Marius Kloft, "Cloze test helps: Effective video anomaly detection via learning to complete video events," in *ACM MM*, 2020, pp. 583–591.

[19] Siqi Wang, Guang Yu, Zhiping Cai, Xinwang Liu, En Zhu, Jianping Yin, and Qing Liao, "Video abnormal event detection by learning to complete visual cloze tests," *ArXiv*, vol. abs/2108.02356, 2021.

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[21] Denis Fortun, Patrick Bouthemy, and Charles Kervrann, "Optical flow modeling and computation: A survey," *CVIU*, vol. 134, pp. 1–21, 2015.

[22] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *CVPR*, 2017, pp. 2462–2470.

[23] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K. Roy-Chowdhury, and Larry S. Davis, "Learning temporal regularity in video sequences," in *CVPR*, 2016, pp. 733–742.

[24] Weixin Luo, Wen Liu, and Shenghua Gao, "A revisit of sparse coding based anomaly detection in stacked rnn framework," in *ICCV*, 2017, pp. 341–349.

[25] Ryota Hinami, Tao Mei, and Shin'ichi Satoh, "Joint detection and recounting of abnormal events by learning deep generic knowledge," in *ICCV*, 2017, pp. 3619–3627.

[26] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao, "Future frame prediction for anomaly detection–a new baseline," in *CVPR*, 2018, pp. 6536–6545.

[27] Joey Tianyi Zhou, Le Zhang, et al., "Attention-driven loss for anomaly detection in video surveillance," *IEEE TCSVT*, 2019.

[28] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel, "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *ICCV*, 2019.

[29] Hyunjong Park, Jongyoun Noh, and Bumsub Ham, "Learning memory-guided normality for anomaly detection," *CVPR*, pp. 14360–14369, 2020.

[30] Weixin Luo, Wen Liu, Dongze Lian, et al., "Video anomaly detection with sparse coding inspired deep neural networks," *IEEE TPAMI*, vol. 43, pp. 1070–1084, 2021.

[31] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos, "Anomaly detection in crowded scenes," in *CVPR*, 2010, pp. 1975–1981.