FUVAS: Few-shot Unsupervised Video Anomaly Segmentation via Low-Rank Factorization of Spatio-Temporal Features

Jiaxiang Jiang
jiaxiang.jiang@intel.com
Intel Labs

Ibrahima J. Ndiour
ibrahima.j.ndiour@intel.com
Intel Labs

Mahesh Subedar
mahesh.subedar@intel.com
Intel Labs

Omesh Tickoo
omesh.tickoo@intel.com
Intel Labs

***Abstract*—Video anomaly detection (VAD) methods analyze untrimmed videos to make temporal decisions at the frame level to identify abnormal events. An important challenge of VAD approaches is the accurate spatial segmentation of the anomalous regions within frames to provide interpretability of anomalies. In this paper, we introduce FUVAS, a fast few-shot unsupervised VAD method ideal for low-data scenarios. FUVAS efficiently identifies temporal anomalies and spatially segments them within each frame of the input video. Our approach harnesses rich video features extracted from pre-trained 3D deep neural networks (DNNs) and performs out-of-distribution detection in the spatio-temporal deep feature space induced by short temporal segments of the video input using low-rank factorization techniques. The proposed approach is agnostic to the choice of 3D DNN backbone architecture and supports both convolutional and transformer models. We present comprehensive results and ablation studies across popular datasets, demonstrating the quality, computational efficiency, and wide applicability of our method.**
Our code is available at: https://github.com/openvinotoolkit/anomalib/tree/main/src/anomalib/models/video/fuvas

I. Introduction

Video anomaly detection (VAD) serves as a critical component in computer vision, tasked with identifying rare or irregular incidents in video streams. Although extensive literature exists on image anomaly detection [1]–[7], these methods often underperform and drastically increase computational and memory requirements when applied to videos. VAD approaches reliant on weakly or semi-supervised frameworks require high-level video labels or supplementary contextual cues [8], [9], which can become unfeasible owing to the scarcity of abnormal video data during the training phase. Also, these methods have difficulty in extrapolating to novel anomaly types unseen during their training. Traditional methods rely on handcrafted features such as flows (trajectories) between frames or histogram of flows [10], [11], and foreground segmentation masks [12]. These handcrafted features are highly subjective and dataset-dependent. Recently, most VAD techniques have relied on deep learning [4], [13]–[23].

Unsupervised methodologies train on anomaly-free training video sets, with the aim of detecting anomalies of an arbitrary nature at test time. These methods strive to identify irregularities without prior guidance. However, most approaches in this category [4], [14]–[21] fail to provide meaningful explanations for their decisions as they only provide binary labels for each frame. A few methods can provide further explainability by segmenting the video anomalies [22], [24]–[28]. While the approach in [27] uses a Siamese network to compare patches of normal sequences, generative approaches for video frames were explored in [22], [24]–[26], [28]. A significant drawback of these approaches is their high demand for memory and computational resources, as well as the necessity for a large number of training video examples.

Regarding training data requirements, anomaly detection methods already face challenges, especially in environments with limited data, such as industrial settings where data collection and normal data segregation are complex and resource-intensive. Recent literature [29], [30] explores model performance in few-shot anomaly detection with low-data scenarios where training data is severely limited.

To tackle these challenges, we propose the Few-shot Unsupervised Video Anomaly Segmentation (FUVAS) method. FUVAS operates effectively in low-data regimes, offering not only anomaly identification but also pixel-level anomaly segmentation for each frame. Our contributions are as follows. We introduce an extremely efficient approach that leverages low-rank factorization techniques for unsupervised VAD. This approach is agnostic to the features obtained from specific DNN architecture (supports both convolutional and transformer-based models) and eliminates the need for DNN retraining. It also significantly reduces the training data requirements. We present extensive experiments to demonstrate improved segmentation and reduced computational requirements.

II. Approach

Problem Definition: Consider a set of videos $\mathbf{V} = \{V_1, V_2, ..., V_n\}$, where n is the number of videos. Video anomaly segmentation is formulated as: given an input video $V_{input} \in R^{T\times 3\times H\times W}$, $(Det, Seg) = A(V_{input})$. Here, A is an anomaly segmentation model, T is the length of the input video sequence (expressed in number of frames) and (H, W) is the video frame's spatial dimensions. $Det \in \mathbb{R}^T$ contain the frame-

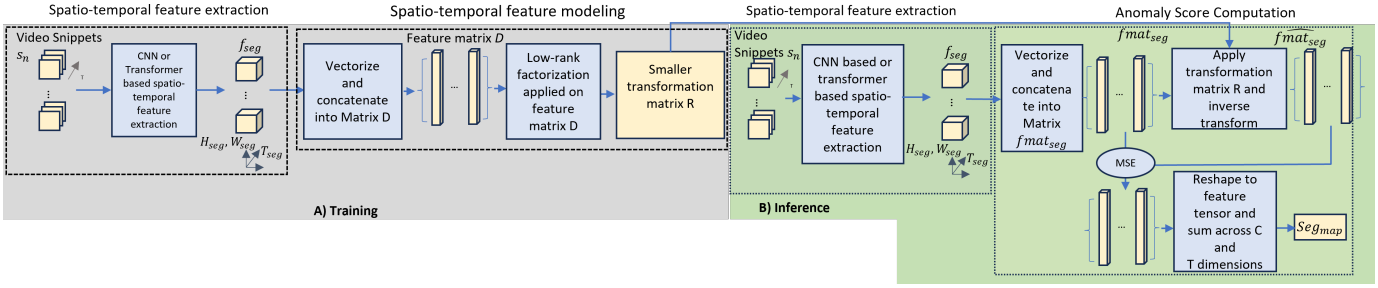

Fig. 1. FUVAS involves extracting and modeling features. During training, the vectorized spatio-temporal features $fvec_{seg}$ (extracted using a frozen 3D model) are aggregated into a feature matrix D. Low-rank factorization is then applied to D to learn a transformation matrix that compresses features into low-dimensional embeddings, retaining only the relevant normal patterns from the training set. At inference, the learned compression model reconstructs normal patterns from the training distribution, yielding $\hat{fmat}_{seg}$. The element-wise mean squared error (MSE) matrix of $fmat_{seg}$ and $\hat{fmat}_{seg}$ is used to generate anomaly segmentation maps.

level anomaly detection results and $Seg \in \mathbb{R}^{T\times H\times W}$ contain the anomaly segmentation maps.

A. FUVAS Overview

Figure 1 provides an overview of our approach, comprising three key components: spatio-temporal feature extraction, spatio-temporal feature modeling, and anomaly score computation. The spatio-temporal feature extraction utilizes pre-trained video feature extraction backbones to generate latent feature representations from video frame snippets. In the spatio-temporal feature modeling part, low-rank factorization is utilized to learn the normal pattern information from the spatio-temporal features. The anomaly score computation module measures the compression error as a proxy for uncertainty that will be used for out-of-distribution detection (OOD), with inliers exhibiting low uncertainty and outliers high uncertainty. In other words, FUVAS learns an invertible transformation matrix through low-rank factorization of the training features' data matrix. At test time, we first apply the learned transformation to the test features' data matrix, followed by its inverse. We then measure the resulting compression errors and use these errors as uncertainty estimates for detecting OOD instances in the feature space.

B. Spatio-temporal deep feature extraction

The input video sequence V_{input} is divided into fixed-length snippets Sn_i ($i \in \{1, 2, \ldots, m\}$). At this stage, the goal is to use a 3D deep neural network to extract the spatio-temporal features $f^i_{\text{seg}} \in \mathbb{R}^{C_{\text{seg}}\times T_{\text{seg}}\times H_{\text{seg}}\times W_{\text{seg}}}$ and $f^i_{\text{det}} \in \mathbb{R}^{C_{\text{det}}\times T_{\text{det}}\times H_{\text{det}}\times W_{\text{det}}}$. For generality, our notation distinguishes between detection and segmentation. In practice, we tap a given layer of the backbone to extract a spatio-temporal feature that is used for both detection and segmentation (i.e. $f^i_{\text{seg}} = f^i_{\text{det}}$). Our approach supports both convolutional neural networks (CNNs) and transformer-based vision models. Next, we show how these spatio-temporal features can be effectively used for video anomaly detection and segmentation.

C. Feature modeling - Anomaly score computation

Consider the features $f^i_{seg} \in \mathbb{R}^{C_{seg}\times T_{seg}\times H_{seg}\times W_{seg}}$, we first vectorize (flatten) them to $fvec^i_{seg}$. For a training set with m video snippets, this results in a spatio-temporal feature matrix $D_{n\times m} = [fvec^1_{seg}, \ldots, fvec^m_{seg}]$, where $n = C_{seg} \times T_{seg} \times H_{seg} \times W_{seg}$. Because of the high dimensionality of the features, $D_{n\times m}$ is rank-deficient. Therefore, we use low-rank factorization to decompose D into smaller matrices $R_{n\times r}$ and $S_{r\times m}$, where r is the rank of D. The matrix $R_{n\times r}$ captures normal patterns from the training data and serves as a compression transformation for feature matrices. Note that there exists several implementation methods for low-rank factorization, including iterative techniques and linear algebra-based approaches [31]. In our implementation, we use Singular Value Decomposition (SVD) as our low rank factorization method for training efficiency.

During the anomaly score computation, $R_{n\times r}$ is applied to a batch of $fvec_{seg}$ denoted as $fmat_{seg}$ to obtain a batch of low-dimensional vector embeddings $femb_{seg} \in \mathbb{R}^{r\times k}$ where k is the number of video snippets in a batch. Subsequently, $R^*_{n\times r}$ reconstructs $\hat{fmat}_{seg}$ from $femb_{seg}$. $R^*_{n\times r}$ can be obtained by using the pseudo-inverse of $R_{n\times r}$. Since $R_{n\times r}$ contains only normal patterns from the training data, the decompressed $\hat{fmat}_{seg}$ should similarly reflect only such normal patterns. Thus, we define the compression error matrix as the element-wise mean squared difference between $fmat_{seg}$ and $\hat{fmat}_{seg}$. This error is an effective uncertainty estimate for out-of-distribution detection [6], [32]. Subsequently, the compression error vectors (i.e. columns of the compression error matrix) are reshaped to match the original tensor dimensions $C_{\text{seg}} \times T_{\text{seg}} \times H_{\text{seg}} \times W_{\text{seg}}$, akin to f_{seg}. Finally, we aggregate along C_{seg} and T_{seg} dimensions to form a map, which is then resized to match the spatial dimensions of the input snippet Sn_i, producing an anomaly segmentation map Seg_{map} for the center frame. Detection proceeds similarly, with the additional step of averaging across the map to derive a scalar value d_i.

III. Experiments and Results

We thoroughly evaluate our approach, comparing it to state-of-the-art unsupervised anomaly detection and segmentation techniques. Testing in the few-shot setting with both convolutional and transformer models shows our method's versatility and effectiveness in low-data regimes. We also analyze com-

TABLE I

BENCHMARKING RESULTS FOR VIDEO ANOMALY DETECTION (VIA AUC METRIC) AND SEGMENTATION (VIA PAUC METRIC). THE FIRST HALF OF THE TABLE CONTAINS METHODS THAT HAVE BOTH DETECTION AND SEGMENTATION CAPABILITIES. ALL METHODS BELOW THE CENTER LINE ARE VIDEO ANOMALY DETECTION-ONLY APPROACHES. *NO SEGMENTATION GROUND TRUTH FOR CUHK AVENUE DATASET.

Methods	UCSD PED1		UCSD PED2		ShanghaiTech Campus		CUHK Avenue	
	AUC	PAUC	AUC	PAUC	AUC	PAUC	AUC	*PAUC
LDSN [27]	86.0	80.4	94.0	93.0	-	-	87.2	NA
DeepOC [26]	83.5	63.1	96.9	95.0	-	-	86.6	NA
BMAN [25]	-	-	96.6	86.7	76.2	-	90.0	NA
IPRAD [28]	82.6	78.4	96.2	93.1	71.5	-	83.7	NA
Tam-net [24]	83.5	69.9	98.1	95.7	-	-	78.3	NA
GMFC-VAE [22]	**94.9**	71.4	92.2	78.2	-	-	83.4	NA
Ours	82.9	**98.1**	95.5	**97.6**	78.5	**97.5**	86.7	NA
EVAL [14]	-		-		76.6		86.0	
AMSRC [33]	-		99.3		76.3		**93.8**	
VERBK [15]	-		98.1		73.8		89.9	
DMAD [17]	-		**99.7**		78.8		92.8	
HSCSA [34]	-		98.1		**83.4**		93.7	
FPDM [20]	-		-		78.6		90.1	

putational complexity during training and inference on various platforms.

A. Experimental setup

During training, we calculate the features from the training samples using a 3D backbone. Each training video is divided into short snippets, i.e. chunks of consecutive frames. The 3D backbone takes a video snippet as input (e.g. a window of 13 consecutive frames) and generates a single spatio-temporal feature tensor. This tensor is flattened to obtain a feature vector. The feature vectors across all the windowed training samples are juxtaposed to form a $n \times m$ data matrix, where m is the number of training samples (snippets) and n is the vectorized feature dimension as defined in §II-C. Low-rank factorization is then applied to the training data matrix in order to learn the transformation matrix to apply to the features. At inference, the feature vectors are calculated in a similar fashion using the test video snippets. Anomalies are then detected in the spatio-temporal feature space by using the low-rank factorization compression error as an uncertainty score.

For our experiments, we utilize multiple 3D backbones all pre-trained on the Kinetics dataset [35]: I3D, X3D and Swin3D [36]. Similar to previous works [14], [17], [37], we use the CUHK Avenue [38], ShanghaiTech Campus [39], UCSD PED1 and PED2 datasets [40]. These datasets contain frame-level and pixel-level ground truth annotations, except for CUHK Avenue which has only frame-level annotations. To quantitatively evaluate the quality of detection (frame-level binary decisions) and segmentation (pixel-level masks), we employ traditional metrics [37] such as the area under the ROC curve (AUC) and per-pixel AUC (PAUC), as well as the more precise Intersection over Union (IoU) [41] as an additional metric for segmentation.

B. Main Results

Table I presents a benchmark of the state of the art. While we also present methods that can only perform anomaly detection in the bottom-half of Table I, we are mostly interested in benchmarking against methods that have both detection and segmentation capabilities. Findings show that our approach surpasses anomaly segmentation benchmarks by a significant margin in terms of PAUC evaluation metric, and performs comparably in terms of anomaly detection. Additionally, we evaluate our method using Intersection over Union (IoU), the standard metric for segmentation in computer vision. On the Shanghai Tech Campus, UCSD Ped1 and Ped2 datasets, FUVAS achieves IoU scores of 0.48, 0.43, and 0.43 respectively, indicating a strong correlation with the ground truth. For a sample frame in each dataset, Figure 2 shows the ground-truth segmentation map and the corresponding FUVAS-generated heatmap. The binarized masks were obtained by thresholding the segmentation map outputs obtained from our algorithm.

C. Few-shot Anomaly Detection and Segmentation

We demonstrate our method's superior performance using fractions of normal training video data (see Figure 3). Given video data's spatial and temporal redundancy, we randomly subsample training data at different rates, train our method, and measure the test set accuracy. This few-shot anomaly detection setup [29], relevant in data-scarce industrial settings, is repeated five times to reduce variability in this experiment. Our method excels in the few-shot setting, achieving optimal quality with as little as 5% or even 1% of the total training dataset, outperforming many methods that require more data.

D. Computational Study

Table III shows our method's efficiency for training and inference, with evaluations on Intel Xeon CPU, NVIDIA RTX3080, and A100 GPUs demonstrating its effectiveness and surpassing real-time requirements. We far exceed the performance of recent methods reported in [14] supplemental.

E. Ablation Studies

1) FUVAS Performance across backbones: Table IV shows the performance obtained while changing the 3D backbone model, and confirms that our approach is independent of the backbone architecture used for feature extraction.

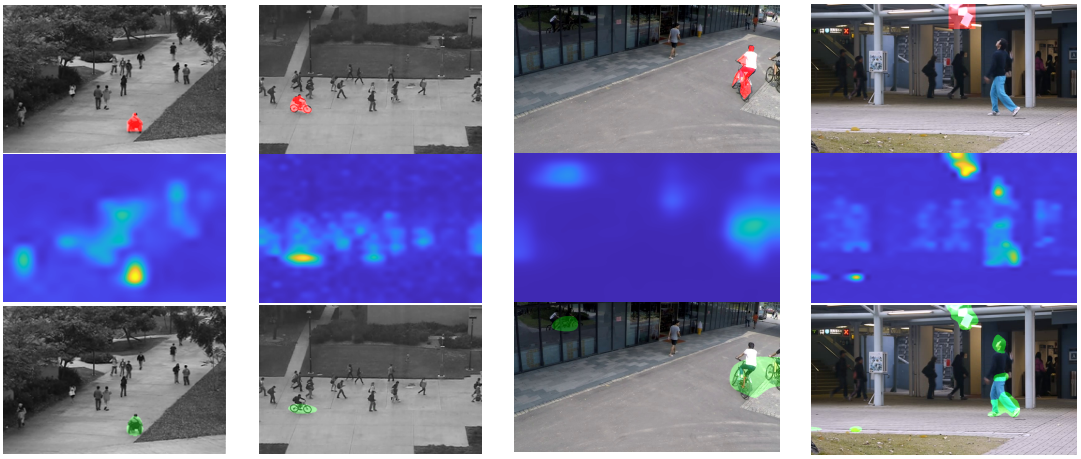

Fig. 2. **Visualization of segmentation results. Top Row:** Ground truth anomaly segmentation masks. **Center Row:** FUVAS-generated anomaly segmentation heatmaps. **Bottom Row:** FUVAS-generated anomaly segmentation binary masks (obtained by thresholding the heatmaps).

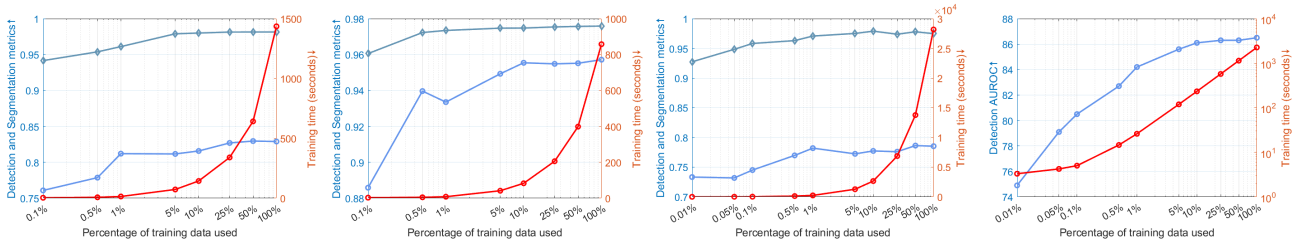

Fig. 3. Few-shot anomaly detection and segmentation on UCSD PED1, UCSD PED2, ShanghaiTech Campus and CUHK Avenue datasets, respectively. Red, blue and teal curves (when applicable) respectively depict the training time, detection AUC and segmentation PAUC. When trained with a fraction of the available training data (as low as 1%), FUVAS produces superb quality similar to full-data training, with extremely low computational and memory costs.

TABLE II
FUVAS SEGMENTATION ACROSS VARIOUS LAYERS OF SWIN3D ON SHANGHAI TECH CAMPUS (STC) AND UCSD PED2 DATASETS.

Dataset	features.1		features.2		features.3	
	PAUC	IoU	PAUC	IoU	PAUC	IoU
STC	93.0	0.47	95.7	0.47	96.3	0.48
PED2	97.3	0.45	97.5	0.45	97.6	0.46

2) FUVAS Performance across layers: Using Swin3D, this ablation study examines the performance of our approach across layers on the same backbone architecture. Findings show that our method is remarkably insensitive to layer selection (see Table II). Similar results are obtained on X3D and I3D backbones.

TABLE III
FUVAS COMPUTATIONAL STUDY: WE MEASURE THE TRAINING TIME FOR THE VIDEO DATA AND THE SUPPORTED INFERENCE FRAME RATE (AT TWO PREDICTIONS PER SECOND FOR 30 FPS INPUT) ON CPU (XEON) AND GPU (RTX 3080 AND A100).

	X3D		Swin3D	
	train (s) ↓	infer (fps) ↑	train (s) ↓	infer (fps) ↑
Xeon	15.3	18.1	2.9	102.3
3080	10.2	70.6	2.5	137.1
A100	3.0	107.2	1.2	319.5

TABLE IV
FUVAS SEGMENTATION PERFORMANCE ACROSS VARIOUS BACKBONES ON SHANGHAI TECH CAMPUS (STC) AND UCSD PED2 DATASETS.

Dataset	X3D		I3D		Swin3D	
	PAUC	IoU	PAUC	IoU	PAUC	IoU
STC	97.5	0.48	95.7	0.48	96.3	0.48
PED2	97.0	0.43	96.2	0.42	97.6	0.46

IV. CONCLUSION

This work presented a fast and efficient method for unsupervised video anomaly segmentation that excels in low-data regimes. Our approach employs low-rank factorization on spatio-temporal features extracted by a 3D DNN and calculates uncertainty estimates for short video segments to discriminate between inliers and outliers. Demonstrated advantages of FUVAS include: (i) state-of-the-art quality in video anomaly segmentation (ii) generalizability to new abnormal scenarios without requiring prior knowledge of anomalies; (iii) model-agnostic feature extraction, compatible with both convolutional and transformer-based models; (iv) fast, effective training with limited data; and (v) low computational complexity and memory requirements, enabling real-time deployment across various compute platforms.

REFERENCES

[1] T. Defard, A. Setkov, A. Loesch, and R. Audigier, "PaDiM: a patch distribution modeling framework for anomaly detection and localization," in *IEEE International Conference on Pattern Recognition (ICPR)*, pp. 475–489, 2021. 1

[2] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14298–14308, 2022. 1

[3] S. Zhang and J. Liu, "Feature-constrained and attention-conditioned distillation learning for visual anomaly detection," in *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2945–2949, 2024. 1

[4] D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, 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*, pp. 1705–1714, 2019. 1

[5] C. Zhou and R. C. Paffenroth, "Anomaly detection with robust deep autoencoders," in *23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 665–674, 2017. 1

[6] I. J. Ndiour, N. A. Ahuja, E. U. Genc, and O. Tickoo, "FRE: A fast method for anomaly detection and segmentation," in *34th British Machine Vision Conference (BMVC), Aberdeen, UK*, 2023. 1, 2

[7] L. Ruff, J. R. Kauffmann, R. A. Vandermeulen, G. Montavon, W. Samek, M. Kloft, T. G. Dietterich, and K.-R. Müller, "A unifying review of deep and shallow anomaly detection," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 756–795, 2021. 1

[8] H. Ye, K. Xu, X. Jiang, and T. Sun, "Learning spatio-temporal relations with multi-scale integrated perception for video anomaly detection," in *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4020–4024, 2024. 1

[9] M. Zhang, J. Wang, J. Wang, Q. Qi, Z. Zhuang, H. Sun, and N. Xiao, "Robust video anomaly detection framework via prior knowledge and multi-path frame prediction," in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. 1

[10] S. Wu, B. E. Moore, and M. Shah, "Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes," in *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 2054–2060, IEEE, 2010. 1

[11] Y. Cong, J. Yuan, and J. Liu, "Abnormal event detection in crowded scenes using sparse representation," *Pattern Recognition*, vol. 46, no. 7, pp. 1851–1864, 2013. 1

[12] B. Antić and B. Ommer, "Video parsing for abnormality detection," in *International conference on computer vision*, pp. 2415–2422, 2011. 1

[13] X. Han, X. Wang, K. Jiang, W. Liu, R. Hu, X. Pan, and X. Xu, "Mutuality attribute makes better video anomaly detection," in *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2670–2674, 2024. 1

[14] A. Singh, M. J. Jones, and E. G. Learned-Miller, "EVAL: Explainable video anomaly localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18717–18726, 2023. 1, 3

[15] 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*, pp. 14592–14601, 2023. 1, 3

[16] A. Al-lahham, N. Tastan, M. Z. Zaheer, and K. Nandakumar, "A coarse-to-fine pseudo-labeling (c2fpl) framework for unsupervised video anomaly detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6793–6802, 2024. 1

[17] W. Liu, H. Chang, B. Ma, S. Shan, and X. Chen, "Diversity-measurable anomaly detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12147–12156, 2023. 1, 3

[18] W. Luo, W. Liu, and S. Gao, "Remembering history with convolutional lstm for anomaly detection," in *2017 IEEE International conference on multimedia and expo (ICME)*, pp. 439–444, IEEE, 2017. 1

[19] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 733–742, 2016. 1

[20] C. Yan, S. Zhang, Y. Liu, G. Pang, and W. Wang, "Feature prediction diffusion model for video anomaly detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5527–5537, 2023. 1, 3

[21] Y. S. Chong and Y. H. Tay, "Abnormal event detection in videos using spatiotemporal autoencoder," in *Advances in Neural Networks-ISNN 2017: 14th International Symposium, ISNN 2017, Sapporo, Hakodate, and Muroran, Hokkaido, Japan, June 21–26, 2017, Proceedings, Part II 14*, pp. 189–196, Springer, 2017. 1

[22] Y. Fan, G. Wen, D. Li, S. Qiu, M. D. Levine, and F. Xiao, "Video anomaly detection and localization via gaussian mixture fully convolutional variational autoencoder," *Computer Vision and Image Understanding*, vol. 195, p. 102920, 2020. 1, 3

[23] G. Shen, Y. Ouyang, and V. Sanchez, "Video anomaly detection via prediction network with enhanced spatio-temporal memory exchange," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3728–3732, 2022. 1

[24] X. Ji, B. Li, and Y. Zhu, "Tam-Net: Temporal enhanced appearance-to-motion generative network for video anomaly detection," in *International Joint Conference on Neural Networks (IJCNN)*, 2020. 1, 3

[25] S. Lee, H. G. Kim, and Y. M. Ro, "Bman: Bidirectional multi-scale aggregation networks for abnormal event detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 2395–2408, 2019. 1, 3

[26] P. Wu, J. Liu, and F. Shen, "A deep one-class neural network for anomalous event detection in complex scenes," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 7, pp. 2609–2622, 2019. 1, 3

[27] B. Ramachandra, M. Jones, and R. Vatsavai, "Learning a distance function with a siamese network to localize anomalies in videos," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2598–2607, 2020. 1, 3

[28] Y. Tang, L. Zhao, S. Zhang, C. Gong, G. Li, and J. Yang, "Integrating prediction and reconstruction for anomaly detection," *Pattern Recognition Letters*, vol. 129, pp. 123–130, 2020. 1, 3

[29] N. Belton, M. T. Hagos, A. Lawlor, and K. M. Curran, "Fewsome: One-class few shot anomaly detection with siamese networks," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2978–2987, 2023. 1, 3

[30] S. Damm, M. Laszkiewicz, J. Lederer, and A. Fischer, "Anomalydino: Boosting patch-based few-shot anomaly detection with dinov2," *arXiv preprint arXiv:2405.14529*, 2024. 1

[31] G. H. Golub and C. F. Van Loan, *Matrix computations*. JHU press, 2013. 2

[32] I. J. Ndiour, N. A. Ahuja, and O. Tickoo, "Subspace modeling for fast out-of-distribution and anomaly detection," in *2022 IEEE International Conference on Image Processing (ICIP)*, pp. 3041–3045, 2022. 2

[33] X. Huang, C. Zhao, and Z. Wu, "A video anomaly detection framework based on appearance-motion semantics representation consistency," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2023. 3

[34] S. Sun and X. Gong, "Hierarchical semantic contrast for scene-aware video anomaly detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22846–22856, 2023. 3

[35] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017. 3

[36] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3202–3211, 2022. 3

[37] B. Ramachandra, M. J. Jones, and R. R. Vatsavai, "A survey of single-scene video anomaly detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2293–2312, 2022. 3

[38] 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*, pp. 2720–2727, 2013. 3

[39] W. Liu, D. L. W. Luo, and S. Gao, "Future frame prediction for anomaly detection – a new baseline," in *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[40] V. Mahadevan, W.-X. LI, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1975–1981, 2010. 3

[41] Z. Wang, E. Wang, and Y. Zhu, "Image segmentation evaluation: a survey of methods," *Artificial Intelligence Review*, vol. 53, pp. 5637–5674, 2020. 3