Supplement to "*Networking Systems for Video Anomaly Detection: A Tutorial and Survey*"

JING LIU, YANG LIU, JIEYU LIN, JIELIN LI, PENG SUN, BO HU, LIANG SONG, AZZEDINE BOUKERCHE, and VICTOR C.M. LEUNG

This is supplementary material to our tutorial paper submitted to the ACM CSUR entitled *"Networking Systems for Video Anomaly Detection: A Tutorial and Survey"* for readers with specific needs and orientations. Due to the limited pages of the main text and some of the contents have been partially covered by existing survey or research papers, we have moved the detailed explanations of the 1) latest advances, 2) classical methods, and 3) research cases to this supplemental material, in order to further enhance the readability of the main text and its value as an guidance for beginners. Specifically, Section 1 states the recent advances in UVAD, WsVAD, and FuVAD, summarizing their main contributions and implementations. Section 2 provides a summarized analysis of the classical methods, including the fundamental principles and architecture overview. Section 3 provides additional explanations of our research cases in smart cities and modern industries to promote the reader to conduct NSVAD research for real-world applications.

1 DETAILED PRESENTATION OF RECENT ADVANCES

1.1 UVAD Methods

1.1.1 GNL with Single Proxy Task. Table 1 summarizes the basic structures and main contributions of existing GNL models with single proxy tasks. Specifically, the Predictive Convolutional LSTM network (PC-LSTM) [50] stands out for its innovative use of a composite conv-LSTM network in predicting video sequence evolutions and using prediction errors to gauge normality scores, demonstrating the utility of LSTM in understanding temporal dynamics. Similarly, the Fully Convolutional Feedforward Autoencoder (FF-AE) proposed by Hasan *et al.* [21] highlights the shift towards end-to-end feature representation learning, using traditional handcrafted features to train anomaly classifiers. The Deep Incremental Slow Feature Analysis Network (D-IncSFA) [23] further advances this by learning from raw data, showcasing the en-to-end feature extraction capability. Smeureanu *et al.* [65] leverage a pre-trained VGG network for extracting deep spatial-temporal features, marking a pivotal move towards integrating deep learning with traditional AD techniques.

The introduction of video prediction tasks proposed proposed by Liu *et al.* [32], using GAN networks, marks a significant evolution, emphasizing the prediction of normalcy in events through generative models. This concept is expanded in FFPN [44], which explores predictive networks' design principles and meta-learning for scene adaptability, reinforcing video prediction's superiority in understanding video contexts. Li *et al.* [31] address the issue of detail loss in autoencoders with their Spatial-Temporal U-net (STU-net), blending U-net's spatial representation strength with ConvLSTM's temporal modeling prowess. AnomalyNet [94] and the Incremental Spatial-Temporal Learner (ISTL) [51] introduce novel approaches to background noise reduction and temporal anomaly evolution.

Recent advancements include the Adversarial Event Prediction (AEP) network proposed by Yu *et al.* [83], using adversarial learning to discern abnormal events, and the ROADMAP model proposed by Wang *et al.* [71], which employs

Authors' Contact Information: Jing Liu, jingliu19@fudan.edu.cn; Yang Liu, yangliu@cs.toronto.edu; Jieyu Lin; Jielin Li; Peng Sun, peng.sun568@duke.edu; Bo Hu, bohu@fudan.edu.cn; Liang Song, songl@fudan.edu.cn; Azzedine Boukerche, aboukerc@uOttawa.ca; Victor C.M. Leung, vleung@ece.ubc.ca.

Table 1. Summary of Single Proxy Task-based GNL methods.

Ref.	Backbone	Contributions
2016 [23]	CNN	Building a deep incremental slow feature analysis network for learning abstraction and global high-level representation.
2016 [21]	AE	Designing a fully convolutional feed-forward AE to learn both local features and classifiers as an end-to-end framework.
2017 [10]	AE, GAN	Combining variational auto-encoder and GAN to model the spatial-temporal features of normal events efficiently.
2017 [42]	ConvLSTM, AE	Integrating CNN and ConvLSTM with auto-encoder to model the appearance and motion patterns of normal events.
2017 [65]	CNN	Using pre-trained CNN to extract appearance features and use OC-SVM to discriminate normal and abnormal events.
2018 [63]	CnovLSTM	Proposing a composite auto-encoder based on convolutional LSTM, 2D and 3D convolution to predict normal events.
2018 [32]	U-net	Detecting anomalies by measuring the difference between predicted and real frames with a prediction framework.
2019 [31]	ConvLSTM	Using U-net and convolutional LSTM based prediction framework to model spatial and temporal information efficiently.
2019 [94]	Sparse LSTM	Integrating feature learning, sparse representation, and dictionary learning in a unified framework to detect anomalies.
2019 [51]	ConvLSTM	Proposing incremental spatial-temporal learners to address the challenges of AD in real-time videos.
2019 [81]	U-net, ConvRNN	Unify reconstruction andprediction methods in an end-to-end deep predictive coding network framework.
2019 [19]	AE	Using memory to limit the generalization power of the AE and proposes a sparse addressing mechanism.
2019 [45]	ResNet, AE	Proposing a sparse coding-inspired neural network model for efficient video anomaly detection.
2020 [9]	3D ConvSLTM	Using 3D Conv-LSTM to model spatial-temporal features and use residual blocks to eliminate the gradient disappearance.
2020 [11]	U-net, GAN	Using two discriminators to improve the generator's ability to characterize the spatial-temporal patterns of normal events.
2020 [4]	U-net	Modeling the normality by performing forward and reverse frame prediction with two independent U-net.
2020 [56]	AE	Using an external memory network to weaken the generalization ability of the deep auto-encoder to anomalous events.
2021 [71]	ConvGRU	Using multi-path prediction framework to predict frames to better handle different scales of objects and regions.
2021 [44]	GAN, U-net	Identifying design principles for prediction-based VAD networks and designing a future frame prediction networks.
2022 [83]	GAN	Training predictive model to discriminate anomalies by the adversarial learning of past and future frames.
2022 [91]	PredRNN	Using ST-LSTM and adversarial learning to learn the evolution of motion and appearance in the short and long term.

a multi-path convGRU for enhanced frame prediction. STC-Net [91] represents a leap forward with its focus on learning long- and short-term patterns, using adversarial learning to refine its capabilities. U-shaped Swin Transformer Network with Dual Skip Connections (USTN-DSC) [80] proposes keyframe-based video event recovery agent task to mine high-level visual features and temporal contextual relationships in videos. Cheng *et al.* [79] introduced a diffusion model to VAD in order to learn the distribution of normal samples without involving any additional advanced semantic feature extraction models. STR-VAD in [72] performs VAD by probing spatio-temporal relationships among objects.

An emerging concern is the potential for models to overgeneralize, making them less effective in distinguishing between normal and abnormal events. This is where memory networks, such as the Memory-enhanced Autoencoder (memAE) [19] and the attention-based memory network proposed by Park *et al.* [56], play a crucial role. These networks introduce mechanisms to limit model generalization, ensuring the memorization of normal event features and improving anomaly detection accuracy. The inclusion of memory networks has been a game-changer, offering a novel approach to balancing complex learning structures and the specificity required for effective anomaly detection. This evolution underscores the field's progression towards more nuanced and adaptable NSVAD models.

1.1.2 GNL with Multiple Proxy Tasks. Videos, as complex spatial-temporal series, offer a unique challenge in detecting anomalies, which may present as deviations in spatial appearance or temporal motion. Recognizing the need to address both dimensions, researchers have explored the use of multiple proxy tasks. These approaches not only model appearance and motion separately but also aim to understand the intricate relationships between these dimensions, such as consistency and correlation.

3D convolutional networks has inspired methods like the Spatial-Temporal Autoencoder (STAE) proposed by Zhao *et al.* [92], which models normality through dual decoders for reconstruction and prediction. A design mirrored in [52] also seeks to understand the correspondence between appearance and motion. Xu *et al.* [78] propose the Appearance and Motion DeepNet (AMDN), which utilizes separate autoencoders for learning from frames and optical flows, demonstrating the power of specialized descriptors in anomaly detection.

The use of GANs in this domain, as explored by Ravanbakhsh *et al.* [58] and further developed in DD-GAN [11] and OGNet [85], highlights the adaptability of GANs to multiple proxy tasks. These models leverage GANs' ability to generate and discriminate, improving VAD by focusing on motion continuity and reconstruction fidelity.

Further contributions include Chang *et al.*'s [2] dual autoencoder approach for capturing distinct spatial and temporal information and the Dual-Stream Deep Spatial-Temporal Autoencoder (DSTAE) proposed by Li *et al.* [29], which emphasizes the role of convLSTM in temporal reasoning. The Appearance-Motion Joint Autoencoder framework proposed by Liu *et al.* [34] and STM-AE [37] showcase innovative strategies for fusing spatial and temporal features, with the latter introducing an external memory network to better capture normality patterns. The AnoPCN [81] model advances concept by integrating error refinement with deep predictive encoding.

Recent works proposed by Cai *et al.* [1], Cheng *et al.* [7], and Ning *et al.* [53] underscore the growing interest in understanding the coherence between appearance and motion. These studies propose novel frameworks for leveraging the intrinsic consistency between these dimensions, aiming for a deeper comprehension of regular events. Hierarchical Semantic Contrast (HSC) in [67] introduces scene-level and object-level contrast learning to deal with the diversity of normal patterns to improve the model's ability to discriminate complex events.

1.1.3 LPM with Spatial-Temporal Patch (STP).. STP methods are grounded in the assumption that anomalies manifest as deviations in local information. They address this by partitioning videos into information cubes, applying methods like Three-dimensional Equi-scale Segmentation (3D-ESE), Equi-spatial Information Intensity Segmentation (ESIIS),

Table 2. Summary of Spatial-Temporal Patch-based Methods.

Ref.	Patch Formulation	Detection Logic
2010 [48]	Slicing the video into equal-sized S&T patches	Deviation to learned dynamic textures
2013 [60]	Dense sampling at different spatial and temporal scales	Modeling S&T arrangements
2016 [95]	Equating video into $3 \times 3 \times 7$ pixel patches	Binary classification with FCN
2016 [8]	Slicing the video into equal-sized S&T patches	Learning normality prototypes
2017 [61]	Equating the video sequence and resizing the objects	Mahalanobis distance to Gaussian models
2017 [43]	Multiple patches sampled at multiple scales	Reconstruction error
2019 [70]	Foreground extraction and keeping only part regions	Reconstruction error
2020 [75]	Equating RGB video and corresponding optical flow frames	One-class classification
2021 [26]	Equatingvideo sequences into spatial-temporal patches	Reconstruction error
2022 [33]	Equating frames into 8×8 pathches along spatial dimension	Prediction error for regions of interest

and High Information Density Filtering (HIDF) to model features independently. This approach enables nuanced understanding of spatial-temporal dynamics, catering to different densities of spatial information and filtering out irrelevant background for focused anomaly detection.

Innovations in this field include Deep-Cascade [61] and Spatial-Temporal Cascade Autoencoder (ST-CaAE) [26], both of which employ cascaded structures for local normality modeling, significantly reducing computational load by prioritizing regions of interest. S^2-VAE [70] highlights the use of foreground object detection in preprocessing to streamline input to variational autoencoders, optimizing the modeling of normality. DeepOC proposed by Wu *et al.* [75] and AST-AE proposed by Liu *et al.* [33] further refine this approach by integrating deep one-class classifiers and attention mechanisms to enhance anomaly detection efficiency.

1.1.4 *LPM with Foreground Objects Detection (FOD).*.

FOD methods, leveraging high-performance object detection models [59], offer an interpretable approach by analyzing specific attributes of foreground objects. This aligns closely with human intuition and has proven effective in complex scenarios like crowd anomaly detection. Techniques range from the LDGK model proposed by Ryota *et al.* [22], which utilizes multi-task learning for semantic information acquisition, to the DCF model [25], focusing on pose classification and motion modeling. The innovative use of visual cloze tests proposed by Yu *et al.* [82] and the Online Video Anomaly Detection (OAD) scheme proposed proposed by Doshi *et al.* underscore the significance of contextual information in enhancing anomaly detection.

Recent advancements include the Background-Agnostic framework [18], emphasizing instance segmentation to minimize background noise, and the OSIN network, which incorporates object detection to enrich scene understanding. Notably, Georgescu *et al.* [17] and the HF^2-VAD framework [40] introduce self-supervised learning tasks and hybrid modeling strategies, respectively, to achieve a comprehensive anomaly detection mechanism. The bidirectional prediction architecture (BiP) proposed by Chen *et al.* [3] and the Hierarchical Scene Normality-Binding Modeling (HSNBM) framework further exemplify the field's move towards integrating advanced prediction models and memory-augmented networks for a more detailed analysis of anomalies. Doshi *et al.* [12] propose a two-stream network to separately learn interactions between different targets and individual skeleton pose changes to perform interpretable VAD.

1.2 WsVAD Methods

Table 3. Summary of Unimodal WsVAD Models.

Ref	Fearure	Decision Logic	Contributions
2018 [66]	C3D	MIL ranking	Proposing to use video-level labels to supervise FCN to compute frame-level anomaly scores.
2019 [49]	ResNet-50, VGG-16	Binary classification	Using dual-stream CNNs to extract spatial and temporal features from video frames and optical flow.
2019 [93]	C3D, TSN^{RGB}, TSN$^{Optical\ flow}$	Action classification	Treating WAED as a supervised task under noise labels and using GCN to correct noise labels.
2019 [96]	VGG16, C3D, Inception, I3D	MIL ranking	Proposing a temporal-enhanced network to learn motion-aware features for MIL ranking model.
2020 [69]	I3D^{RGB}, I3D$^{Optical\ flow}$, I3D^{conc}	MIL ranking	Designing dynamic multi-instance learning loss and center loss for expanding the inter-class distance and reducing the intra-class distance of normal instances.
2020 [88]	C3D	MIL ranking	Using a clustering algorithm to generate pseudo-labels to assist the training of regression model.
2020 [86]	C3D	MIL ranking	Proposing a random batch-based training strategy to reduce the correlation between batches.
2021 [68]	C3D^{RGB}, I3D^{RGB}	MIL ranking	Proposing RTFM to explore the important temporal correlations for identification of positive instances.
2021 [47]	I3D	MIL ranking	Fusing S&T contexts to perform weakly-supervised video anomaly localization while proposing an enhancement strategy to eliminate noise interference.
2021 [15]	C3D^{RGB}, I3D^{RGB}	MIL ranking	Using a generator to generate reliable pseudo labels and extract task-specific deep representations.
2022 [24]	C3D	MIL ranking	Proposing a deep temporal coding scheme to capture the temporal evolution of video examples over time, reducing the false alarm rate of anomaly detection.
2022 [90]	I3D^{RGB}	MIL ranking	Exploring the temporal relationships between video clips, and capturing the task-specific features.
2022 [27]	C3D, TSN^{RGB}, TSN$^{Optical\ flow}$	MIL ranking	Presenting a graph convolutional network for cleaning label noise with integrated feature similarity and temporal consistency of anomaly analysis.
2022 [28]	CNN	Binary classification	Using 2D convolutional networks and echo state networks to obtain local ratio representations, and then using 3D convolutional networks to extract spatial-temporal features.
2022 [35]	I3D	MIL ranking	Using RNNs to capture temporal correlations and using a clustering algorithm to generate pseudo-labels for the training of MIL regression model.
2022 [38]	C3D^{RGB}, I3D^{RGB}	MIL ranking	Proposing recurrent criss-cross attention to explore the connection between local S&T representations.

1.2.1 Unimodal Methods. We have summerized the decision logic and main contributions of existing unimodal WsVAD models in Table 3. Specifically, Snehashis *et al.* [49] utilized a dual-stream CNN for spatial and temporal feature extraction, comparing architectures like ResNet 50 and Inception V3. Zhu and Newsam [96] emphasized motion information's importance, proposing a Temporal Augmented Network for motion-aware feature learning. Zhong *et al.* [93] approached WAED as a supervised task with noisy labels, suggesting graph convolutional networks for label rectification. The Anomaly Regression Net (ARNet) [69] and a two-stage framework proposed by Waseem *et al.* [28] focused on discriminative feature learning and spatial-temporal feature fusion, respectively.

Tian *et al.* [68] introduced Robust Temporal Feature Metric Learning (RTFM) employing dilated convolutions and self-attention for accurate feature learning. Muhammad *et al.* [88] used clustering for pseudo-label generation, introducing a clustering loss for enhanced anomaly detection. The CLAWS model [86] proposed a random batch training strategy and a clustering distance-based loss to handle label noise.

Ammar *et al.* [24] developed a Deep Temporal Encoding-Decoding (DTED) solution for capturing spatial-temporal evolution patterns, employing joint loss optimization. The Weakly Supervised Temporal Relation Learning (WSTR) [90] utilized transformer technology for mining semantic correlations, while the Within-Video Abnormality Spotting and Localization (WASL) [47] focused on fusing temporal and spatial contexts with high-order context encoding.

Lastly, Feng *et al.* [15] presented a Multi-Instance Self-Training (MIST) framework for generating reliable pseudo-labels and extracting task-specific representations. Liu *et al.* [35] introduced a Self-guiding Multi-instance Ranking (SMR) framework using clustering for pseudo-label generation and exploring task-relevant features. The Spatial-Temporal Attention (STA) framework [38] aimed at understanding the relationship between local and global features, employing recurrent cross-attention operations for feature enhancement. In response to the vulnerability of MIL to anomalous fragments with simple contexts that lead to high false alarms, Lv *et al.* [46] proposed Unbiased MIL (UMIL) to eliminate contextual bias. Chen *et al.* [6] propose a feature amplification mechanism and amplitude contrast loss to enhance the discrimination of feature amplitude detection anomalies

1.2.2 Multi-modal Methods. Wu *et al.* [76, 77] introduced HL-Net, a three-branch network for violence behavior detection that employs similarity and proximity priors for capturing long-distance correlations and local positional relations, alongside a score branch for dynamic proximity capture. Pang *et al.* [55] proposed a feature fusion network that utilizes a bilinear pooling mechanism for visual and audio information fusion, facilitating mutual learning for enhanced feature representations in multi-modal violence detection tasks. The Modality-aware Contrastive Instance Learning with Self-Distillation strategy [84] addresses modality heterogeneity through lightweight dual-stream networks and a self-distillation module that bridges the semantic gap between multi-modal features.

Wei *et al.* [73] proposed the Multi-modal Supervised Attention Fusion (MSAF) framework for implicit multi-modal data alignment, refining video-level ground truth into pseudo-clip-level labels and employing attention fusion guided by these labels for feature fusion and anomaly score prediction. Shang *et al.* [62] tackled the challenge of dataset limitations by proposing mutual distillation to transfer knowledge from larger to smaller datasets, along with a multi-modal attention fusion strategy for more discriminative feature representations. The Audio-Guided Attention Network (AGAN) [57] extracts video and audio features, enhances them temporally through the cross-modal aware local awakening network, and computes anomaly scores with temporal convolution, showcasing the potential of multi-modal approaches in overcoming the constraints of current VAD research.

Text Empowered Video Anomaly Detection (TEVAD) [5] computes text features using subtitles of videos to capture the semantics of anomalous events. Zhang *et al.* [89] propose an enhanced two-stage self-training framework that utilizes completeness and uncertainty properties. Feng *et al.* [14] use a feature set that captures the temporal synchronization between video frames and sound to train an autoregressive model to generate audiovisual feature sequences for video forensics. Multimodal Motion Conditioned Diffusion [16] considers skeletal representation and utilizes SOTA diffusion model to generate multimodal future human poses.

2 DETAILED EXPLANATION OF CLASSICAL METHODS

2.1 UVAD

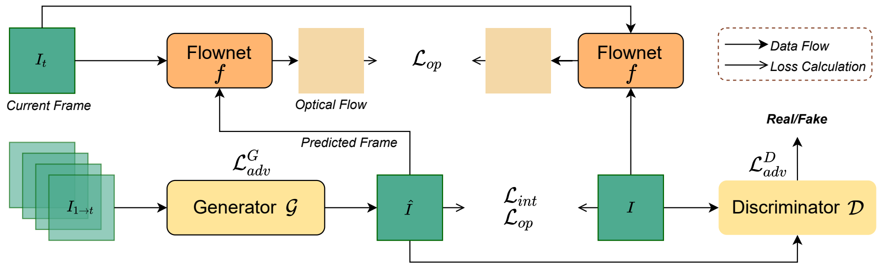


Fig. 1. Overview of the Future Frame Prediction framework. FFP consists of a pre-trained optical flow network [13] f, a generator $\mathcal{G}$, and a discriminator $\mathcal{D}$. During training, $\mathcal{G}$ accepts a continuous sequence $\boldsymbol{I}_{1:t}$ of t frames as input to predict the next frame $\hat{\boldsymbol{I}_{t+1}}$, while $\mathcal{D}$ attempts to differentiate between the predicted frame $\hat{\boldsymbol{I}_{t+1}}$ and the real future frame $\boldsymbol{I}_{t+1}$. The learnable parameters of $\mathcal{G}$ and $\mathcal{D}$ are optimized with optical flow loss $\mathcal{L}_{op}$, intensity loss $\mathcal{L}_{int}$, gradient losses $\mathcal{L}_{gd}$, and adversary loss $\{\mathcal{L}^{D}_{adv}, \mathcal{L}^{G}_{adv}\}$.

2.1.1 Future Frame Prediction (FFP) Framework. FFP [32] introduces the prediction proxy task into UVAD for the first time, laying the research foundation for prediction methods. The authors believe that traditional reconstruction proxy tasks based on autoencoders, while easy to train, may be ineffective in detecting anomalies due to the strong learning ability of deep neural networks on simple tasks, leading to missed detections. In contrast, Video Prediction (VP) requires models to explicitly demonstrate the spatial-temporal pattern evolution of inferred video data, empowering models to understand the internal logic of regular events. Therefore, they propose the video prediction framework shown in Fig. 1, which is based on the assumption that predictors (i.e., generators $\mathcal{G}$) learned on massive regular events cannot infer the appearance and motion information of unseen anomalous samples. Inspired by the successful application of GANs in VP tasks, the core components of the FFP framework are a U-Net-based generator $\mathcal{G}$ for predicting future frames and a patch discriminator $\mathcal{D}$ introduced from the Least Square GAN.

During the training phase, the $\mathcal{G}$ based on U-net accepts t consecutive frames $\boldsymbol{I}_{1:t}$ as input and attempts to predict the next frame, while the $\mathcal{D}$ attempts to differentiate between input images as real or predicted. The authors introduce intensity, gradient, and optical flow constraints to measure the differences between predicted $\hat{\boldsymbol{I}}$ and ground truth future frames $\boldsymbol{I}$ from spatial appearance, temporal motion, and gradient perspectives, denoted as $\mathcal{L}_{int}$, $\mathcal{L}_{gd}$, and $\mathcal{L}_{op}$, respectively. Additionally, the generator and the discriminator enhance each other through adversarial learning. The objective function of $\mathcal{D}$ is to discriminate $\boldsymbol{I}$ as real and output 1 for $\hat{\boldsymbol{I}}$. While $\mathcal{G}$ aims to generate as realistic predicted frames as possible that can be recognized as 1 by $\mathcal{D}$.

During the testing phase, FFP only uses the well-trained $\mathcal{G}$ to predict the next frame and quantitatively calculates the degree of anomaly by measuring the difference between the output and ground truth. Specifically, the authors use Peak Signal-to-Noise Ratio (PSNR) to calculate the quantitative difference between $\boldsymbol{I}$ and $\hat{\boldsymbol{I}}$. To better illustrate the relationship between this value and the degree of anomaly, PSNR of all frames of the same test video is normalized to the anomaly score s in the range [0,1] using max-min normalization. Since the larger the difference between images, the smaller their PSNR, a smaller s indicates a higher degree of anomaly.

FFP utilizes the unique temporal structure information of video as self-supervised signals to learn video normality, opening up a new track of video self-supervised learning for UVAD—future frame prediction. Subsequent research has shown that even with the similar network, prediction proxy tasks can achieve additional performance gains compared to reconstruction, completely distinguishing VAD research from conventional AD tasks on non-temporal data. Most of the prediction-based UVAD models also use GAN as backbone and introduce optical flow to learning temporal

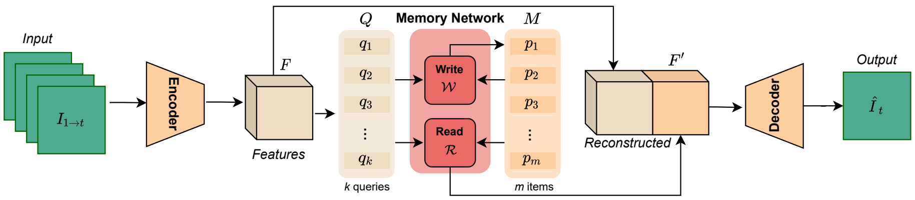
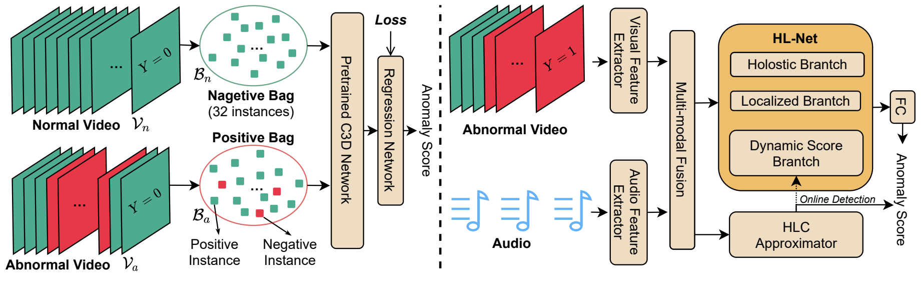


Fig. 2. Pipeline for Memory-Guided Video Normality Learning (MGVNL). MGVNL utilizes a memory network $\mathcal{M}$ to store prototype patterns of normal events to limit the overgeneralization of encoder and decoder to anomalies.

dynamics with inspiration from FFP. Moreover, FFP framework is easy to implement and shows significant performance improvements through new proxy tasks, inspiring following researchers to focus on efficient proxy task design. They have proposed a series of video-specific tasks for normality learning, such as bidirectional prediction, spatial-temporal puzzles, and stochastic motion representation, whihc significantly advance the development of UVAD.

2.1.2 Memory-Guided Normality Learning. AD community always emphasizes the diversity of anomalies, thus unsupervised methods only use easily identifiable regular events to train models, thereby avoiding defining and collecting rare diverse anomalies. However, a long-standing problem is that the spatial-temporal features of regular events are also diverse and exhibit a wide range of pattern overlap with anomalous events, which may result in models trained on normal samples effectively reconstructing anomaly samples through learned diverse normal patterns, even without having seen them. Numerous studies have shown that deep neural networks may characterize abnormal events with small reconstruction/prediction errors due to excessive generalization. To address this, Park *et al.* [56] proposed Memory-Guided Video Normality Learning, which embeds an external memory network $\mathcal{M}$ between the encoder and decoder to weaken the generalization ability of deep models, as shown in Fig. 2. Experiments show that the memory network brings considerable performance improvement to UVAD with low training cost and parameter count, eliminating obstacles for large-scale video representation models in the VAD application. Many subsequent UVAD studies have referenced this work using memory networks to constrain the overgeneralization of deep neural networks and proposed new addressing mechanisms and optimization strategies.

The memory network is essentially a 2D matrix, denoted by $\boldsymbol{M} = \{\boldsymbol{p}_1, \boldsymbol{p}_1, \cdot, \boldsymbol{q}_m\} \in \mathbb{R}^{m\times C}$, which contains m items of dimension C and is embedded between the encoder and decoder. During training, $\mathcal{M}$ writes the prototype patterns of spatial-temporal features $\boldsymbol{F} \in \mathbb{R}^{H\times W\times C}$ extracted by the encoder into $\boldsymbol{M}$ through write operation, and then uses historical memory items through read operations to construct $\boldsymbol{F'}$. The decoder uses the concatenated $\boldsymbol{F'}$ and $\boldsymbol{F}$ as input to reconstruct input sequences or predict future frames. During testing, the concatenated features of regular events are similar to $\boldsymbol{F}$ and can be effectively understanded by the decoder, while the that of unseen abnormal events will deviate significantly, leading to large decoding errors. First, $\boldsymbol{F}$ is unfolded along the spatial dimension into k query vectors, denoted as $\boldsymbol{Q} = \{\boldsymbol{q}_1, \boldsymbol{q}_1, \cdot, \boldsymbol{q}_k\} \in \mathbb{R}^{k\times C}$, where $k = H \times W$. The read operation aims to reconstruct $\boldsymbol{q}_i$ using memory items from $\boldsymbol{M}$: $\boldsymbol{F} \leftarrow \hat{\boldsymbol{q}} = \boldsymbol{w}_i \boldsymbol{M}$, where $\boldsymbol{w}_i \in \mathbb{R}^{1\times m}$ denotes the weighting coefficient. In [56], $\boldsymbol{w}$ is defined as the cosine similarity between $\boldsymbol{q}_i$ and al items, followed by softmax normalization. The write operation uses $\boldsymbol{Q}$ to update $\boldsymbol{M}$ to record prototype patterns of regular events. Similar to the read operation, the weighting coefficient is defined as the normalized cosine similarity between $\boldsymbol{p}_i$ and all query vectors. To ensure the numerical scale of memory items remains comparable during the update process, the authors perform $\mathbf{L}_2$ normalization on the weighted memory items: $\boldsymbol{M} \leftarrow \hat{\boldsymbol{p}}_i = \mathbf{L}_2(\boldsymbol{p}_i + \boldsymbol{w}_i Q)$.



(a) Multiple Instance Ranking Framework (b) Holistic and Localized Network

Fig. 3. Architecture overview of **(a)** Multi-Instance Ranking (MIR) framework and **(b)** Holistic and Localized Network. The MIR employs a regression network to calculate anomaly scores with the supervision from video-level labels. The HN-Net aims to mine anomaly cues from both video and audio to enhance the performance of the MIL framework in violence detection.

In recent years, researchers have improved the addressing mechanisms and optimization strategies of memory networks. For example, Lv *et al.* proposed using attention networks to calculate weight coefficients during the update process and first embedded the memory network into the decoder to further enhance the model's learning ability for event prototypes. MSN-net uses loss functions inspired by information theory to constrain memory entry updates and proposes a Top-k attention-based addressing mechanism to alleviate the negative impact of event diversity on prototype feature learning.

2.2 WsVAD

2.2.1 Multiple Instance Ranking (MIR) framework. MIR [66] introduced video-level labels into VAD task and proposed the MIL-based weakly supervised route. Its basic structure is illustrated in Fig. 3(a), comprising a pre-trained C3D network for spatial-temporal feature extraction and a regression network for calculating frame-level anomaly scores. In contrast to UVAD models in Fig. 1 and 2, WsVAD have access to pre-defined anomaly samples and video-level labels Y during training. MIR first divides the video into 32 non-overlapping fixed-length segments, each containing 16 consecutive frames. The authors treat each segment as an instance, and all instances from the same video constitute a bag. Clearly, all instances in the negative bag $\mathcal{B}_n$ formed by normal videos $\mathcal{V}_n$ are negative, while only a few instances in the positive bag $\mathcal{B}_a$ are positive.

MIR frames the VAD under the supervision of video-level labels as a regression problem, aiming to use weak-semantic annotations to supervise a fully connected network (FCN) to output instance-level anomaly scores ranging from [0, 1], where higher scores indicate a higher likelihood of anomalies. An intuitive optimization strategy is to use the ranking loss, denoted as $f(\mathcal{V}_a) > f(\mathcal{V}_n)$, which encourages the regression network to output higher scores for $\mathcal{V}_a$. However, during the training phase, only video-level labels are available, meaning that it is known whether the input video contains anomalies, but the specific temporal positions of anomalies are unknown. Inspired by MIL, the authors propose a multi-instance ranking loss to encourage the highest score of instances in $\mathcal{B}_a$ to exceed that in $\mathcal{B}_n$: $\max_{i\in\mathcal{B}_a} f\left(\mathcal{V}_a^i\right) > \max_{i\in\mathcal{B}_n} f\left(\mathcal{V}_n^i\right)$. Furthermore, drawing from the Hinge loss in SVM for classification, this objective equation is optimized as $\max\left(0, 1 - \max_{i\in\mathcal{B}_a} f\left(\mathcal{V}_a^i\right) + \max_{i\in\mathcal{B}_n} f\left(\mathcal{V}_n^i\right)\right)$. Additionally, considering the sparsity of positive instances in $\mathcal{B}_a$ and the continuity of events, MIR introduce sparsity constraint $C_{sparsity}$ and smoothness constraint $C_{smoothness}$. Regularization constraints are added to the model weights $\boldsymbol{W}$ to prevent overfitting. Therefore, the final objective function was balanced with hyper-parameters $\{\lambda_1, \lambda_2\}$, as follows:

$$l\left(\mathcal{B}_a, \mathcal{B}_n\right) = \max\left(0, 1 - \max_{i \in \mathcal{B}_a} f\left(\mathcal{V}_a^i\right) + \max_{i \in \mathcal{B}_n} f\left(\mathcal{V}_n^i\right)\right) + \lambda_1 \overbrace{\sum_{i}^{n-1)} \left(f\left(\mathcal{V}_a^i\right) - f\left(\mathcal{V}_a^{i+1}\right)\right)^2}^{C_{smoothness}} + \lambda_2 \overbrace{\sum_{i}^{n} f\left(\mathcal{V}_a^i\right)}^{C_{sparsity}}, \tag{1}$$

Researchers follow the MIL ranking of MIR and have proposed many innovative WsVAD works, driving model performance improvement and field development. Mainstream UVAD solutions, such as reconstruction/prediction-based models, require obtaining the entire video first and then performing offline detection. In contrast, WsVAD methods like MIR can online output anomaly scores for continuous video streams, promoting the application of VAD in IoVT.

2.2.2 Holistic and Localized Network. To fully unleash the potential of deep learning in multimodal violence detection, Wu *et al.* [76] collected a large-scale violence behavior dataset with weak semantic annotations, comprising video and audio files, pioneering research in multimodal VAD. Details of the dataset will be presented in Sec. **??**, while this section focuses on the Holistic and Localized Network (HLN) proposed by them as a case study for video-audio anomaly detection under weak supervision settings. The core structure of the HLN network is illustrated in Fig. 3(b), consisting of three parts: data preprocessing, offline detection, and online detection units. The authors first extract deep features for video and audio separately using the I3D and VGG networks, which are then concatenated along the channel dimension and fed into a fully connected network for fusion. The HL-Net comprises holistic and localized branches, designed to capture long-range and short-range relationships in the fused video-audio features, respectively. Inspired by graph neural networks, the holistic branch employs feature similarity to define the overall relationship matrix. Additionally, the localized branch considers position distances in non-local operations using a proximity-based local relationship matrix. Since the aforementioned HL-Net requires access to the entire video due to its long-range dependencies, it cannot perform online detection, which is crucial for intelligent surveillance systems. To address this, the authors introduce a Holistic and Localized Cue (HLC) approximator, used under the guidance of HL-Net to generate accurate predictions using past video segments. Furthermore, the dynamic score branch computes the response at a particular position as the weighted sum of features from all positions, further enhancing online detection performance. Due to the availability of only video-level labels, the training of the above modules still follows multiple instance learning.

2.3 FuVAD

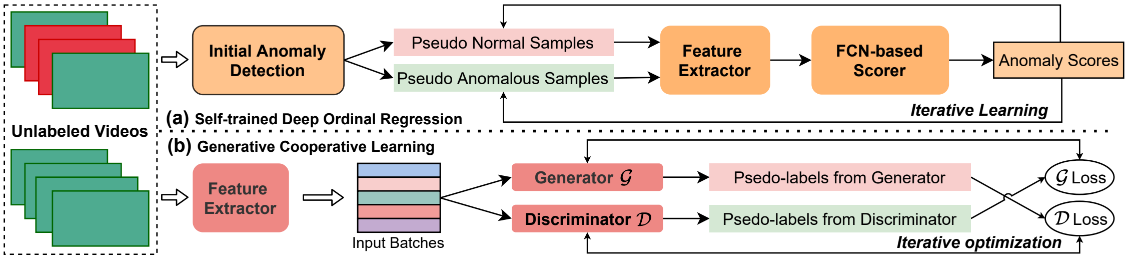


Fig. 4. Architecture overview of (a) SDOR [54] and (b) GCL [87].

2.3.1 Self-trained Deep Ordinal Regression. The basic structure of SDOR is shown in Fig. 4(a), consisting of an initial anomaly detection module, feature extractor, and anomaly scorer. Unlike unsupervised methods that rely on carefully curated training data, SDOR utilizes imbalanced normal and anomaly samples without any labels available. The authors argue that such a setup can avoid the costs of data curation and labeling, enabling the model to directly learn from massive amounts of raw real-world videos, extending NSVAD applications to online video content detection. Furthermore, in SDOR, feature learning and score training are jointly optimized, avoiding the suboptimal model performance due to excessive reliance on pre-trained feature extractor in WsVAD. Specifically, the authors first use Isolation Forests on non-specifically oriented video data to separate potentially anomalous instances with significant differences from the raw data, then employ such pseudo-samples to supervise both the feature extraction module and anomaly scorer. Given that label noise from initial pseudo-samples may lead to inaccuracies in the output anomaly scores, SDOR introduces iterative learning to update pseudo-labels.

2.3.2 Generative Cooperative Learning. In contrast to the adversarial learning in GANs [64], GCL employs a cooperative learning between generator $\mathcal{G}$ and discriminator $\mathcal{D}$ to optimize each other, as shown in Fig. 4(b). Specifically, $\mathcal{G}$ is implemented as an autoencoder tasked with reconstructing the features of input videos. Based on reconstruction errors and a threshold, $\mathcal{G}$ labels the input as pseudo-normal or pseudo-anomalous, where the pseudo-labels are used to train $\mathcal{D}$. For pseudo-anomalies with high confidence (samples with significantly large feature reconstruction errors), the authors introduce negative learning to encourage $\mathcal{G}$ to further exaggerate the errors. The $\mathcal{D}$ is a simple fully connected network, used to compute the anomaly probability of input samples for supervising $\mathcal{G}$.

In implementation, the authors employ ResNext3d [20] as the feature extractor to characterize unlabelled video sequences. To eliminate the covariate shift between intra-batch and inter-batch features, they perform random sampling on feature vectors to constitute input batch features for $\mathcal{G}$ and $\mathcal{D}$. $\mathcal{G}$ is trained in a self-supervised learning manner, where the loss is defined as the MSE between input and reconstructed features. According on the low frequency of anomalies, the authors regard samples whose reconstruction errors rank within the top fixed percentage as pseudo-anomalies. In contrast, $\mathcal{D}$ is trained using pseudo-labels generated by $\mathcal{G}$, aiming to minimize the binary cross-entropy loss. The process of generating pseudo-labels for $\mathcal{D}$ mirrors that of $\mathcal{G}$, where samples with high probability values exceeding a predefined threshold are considered anomalies. For the pseudo abnormal samples identified by $\mathcal{G}$, the authors adjust the output targets during $\mathcal{G}$'s training process to be a vector of all ones rather than the input features, thereby further exaggerating the reconstruction errors between anomalies and normals.

3 FURTHER PRESENTATION OF RESEARCH CASES

Due to the presence of multiple entities (e.g., humans, machines, and products) with different spatial-temporal interactions in modern factories, NSVAD for smart industries faces greater challenges compared to simple scenarios such as public spaces where anomalies are primarily human-centric. Inspired by memory networks, we explicitly identified the key to open-set NSVAD in dynamic industrial environments as efficiently representing diverse normal patterns while maintaining limited generalization to possible anomaly events. To this end, We designed a spatial-temporal memory augmented three-branch network [36] to explore prototype patterns of regular events in appearance, motion, and spatial-temporal interaction dimensions. Experimental results on real industrial datasets validated the effectiveness and superiority of our method, demonstrating that NSVAD can detect various manufacturing anomalies and enhance the safety of industrial processes.

In smart cities, regular videos from the real world and mobile internet often contain label-independent data shifts due to equipment specifications, acquisition angles, and external weather conditions. Moreover, some potentially severe anomalies exhibit minimal differences from regular events in spatial-temporal patterns. For example, vehicles traveling against traffic flow, while understandable to humans, may lead to false negatives for NSVAD models due to the lack of explicit trajectory analysis and preconceived notions that roads only allow travel in specified directions. Existing deep learning models often struggle to generalize effectively to regular events with data shifts while remaining sensitive to minor anomalies. To address this, we proposed the Causal Video Normality Learning (CVNL) in [39] to uncover the regularity with potential causal relationships to deal with diverse events in complex environments. The CNNL achieved significant performance gains in cross-scenario videos. Its robustness further expanded the application prospects of NSVAD in complex environments.

Inspired by edge artificial intelligence, we designed the first NSVAD system for large-scale IoT applications. Specifically, we propose an End-Cloud Cooperation (ECC) framework to comprehensively consider the detection performance of NSVAD with thousands of terminals and the collaborative allocation of communication and computational resources of networked devices. The ECE-based NSVAD system is still in its nascent stages, which aims to inspire researchers from the IoT and communication communities to collectively drive the real-world applications.

3.1 Appearance-Motion Prototype Network

Through observations of manufacturing processes, we categorize diverse industrial anomalies into three types: appearance-only (e.g., humans entering machine workspaces and foreign objects intruding on assembly lines), motion-only (e.g., sudden acceleration or abrupt stops of robotic arms), and appearance-motion united anomalies (e.g., product drops due to failed grabs). Such observations encourage us to design different structures and proxy tasks based on unique characteristics of different anomalies to address the diverse spatial patterns, high temporal dynamics, and complex spatial-temporal interactions in industrial environments. Inspired by existing multi-proxy task methods, we propose the AMP-Net as shown in Fig. **??**, which consists of appearance encoding $\mathcal{E}_a$, motion encoding $\mathcal{E}_m$, and spatial-temporal fusion decoding to respectively model spatial, temporal, and spatial-temporal regularity. Unlike existing multi-stream approaches [37] that do not distinguish between static appearance and dynamic motion and use the same encoding network to obtain spatial and temporal features, we think that $\mathcal{E}_a$ aims to understand local spatial contexts of regular events while $\mathcal{E}_m$ focuses on modeling global temporal changes. Therefore, we propose a fusion module based on channel attention to enhance $\mathcal{E}_a$'s understanding of multi-level appearance semantics and introduce temporal attention in $\mathcal{E}_m$ to empower the model to actively capture important dynamics.

Previous methods that separately model appearance and motion usually use simple error measurements from two branches to measure the spatial and temporal information deviation of test samples, while we propose a fusion module and decoding task to explore the normal spatial-temporal interactions of regular events. We assume that anomalies deviating from both appearance and motion information lack such inherent interactions. It is worth mentioning that to prevent AMP-Net from overly strong learning abilities to infer unseen anomaly instances using patterns learned from regular events, we introduce memory networks in both $\mathcal{E}_b$ and $\mathcal{E}_m$ to weaken the model's erroneous generalization ability to anomalies. During the training phase, spatial memory $\mathcal{M}_S$ and temporal memory $\mathcal{M}_T$ alternately perform read-write operations to store appearance and motion prototype features of regular industrial videos, while during the testing phase, only read operations are performed to amplify the feature deviation of anomaly events. We introduce a novel filtering mechanism in read operations to avoid the negative impact of personalized features of regular events. AMP-Net achieves frame-level AUCs of 98.7%, 92.4%, and 78.8% on UCSD Ped2 [30], CUHK Avenue [41], and ShanghaiTech [32] datasets, respectively. Moreover, case studies on real-world industrial datasets demonstrate AMP-Net's significantly higher detection capabilities for complex anomalies compared to contemporaneous UVAD schemes, indicating that constructing AIADS for industrial safety with NSVAD is worth exploring.

3.2 Causal Video Normality Learning

CVNL consists of *a*) feature extractor, *b*) memory, *c*) prototype decomposer, and *d*) causally inspired characterizer. *a* and *b* are the memory-enhanced encoders used to record prototype patterns of regular events. Their principles and implementation have been presented in 2.1.2. The original deep features $\boldsymbol{F}$ of regular events contain shared semantics and personalized semantics. *c*, inspired by self-supervised sparse representation learning [74], uses a set of parallel FCNs to separate these two types of semantics into private features $\boldsymbol{F}_s$ and shared features $\boldsymbol{F}_p$. Most importantly, *d* is the core of CVNL, used to explore potential causal variables determining video regularity from $\boldsymbol{F}$. Specifically, inspired by the common cause principle and independent cause mechanism, we believe there is a set of jointly independent causal factors that can fully capture the statistical dependency from low-level observations to high-level descriptions of video normality. Additionally, the sparse shift-invariant assumption suggests that the influence of personalized features of normal events on learned stable causal factors and consistency is local and limited. Therefore, we construct *d* to learn unobservable causal factors and model the inherent consistency of regular events from the perspective of CRL. Specifically, the $\boldsymbol{F}$ the same batch of **b** video segments are fed into CIR in parallel, mapping $\boldsymbol{F}_p$ and $\boldsymbol{F}_s$ into the same causal representation space, denoted as $\boldsymbol{R}$ and $\boldsymbol{R}'$. The causal variables should maintain causal invariance for the so-called prototype decomposition intervention, i.e., the causal representation of $\boldsymbol{F}_p$ and $\boldsymbol{F}_s$ of the same**b** video segments should remain unchanged in terms of causal factors, i.e., consistency. To ensure that the causal factors are jointly independent, we construct three correlation matrices to constrain the cross-information of different dimensions of causal representation, denoted as $\boldsymbol{C}_1$, $\boldsymbol{C}_1$, and $\boldsymbol{C}_1$. The key optimization objective is to maximize the diagonal elements of the correlation matrix $\boldsymbol{C}_1$ (the diagonal elements of $\boldsymbol{C}_2$ and $\boldsymbol{C}_3$ are constant 1) and minimize their off-diagonal matrices.

REFERENCES

[1] Ruichu Cai, Hao Zhang, Wen Liu, Shenghua Gao, and Zhifeng Hao. 2021. Appearance-motion memory consistency network for video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 938–946.

[2] Yunpeng Chang, Zhigang Tu, Wei Xie, and Junsong Yuan. 2020. Clustering driven deep autoencoder for video anomaly detection. In *European Conference on Computer Vision*. Springer, 329–345.

[3] Chengwei Chen, Yuan Xie, Shaohui Lin, Angela Yao, Guannan Jiang, Wei Zhang, Yanyun Qu, Ruizhi Qiao, Bo Ren, and Lizhuang Ma. 2022. Comprehensive Regularization in a Bi-directional Predictive Network for Video Anomaly Detection. In *Proceedings of the American association for artificial intelligence*. 1–9.

[4] Dongyue Chen, Pengtao Wang, Lingyi Yue, Yuxin Zhang, and Tong Jia. 2020. Anomaly detection in surveillance video based on bidirectional prediction. *Image and Vision Computing* 98 (2020), 103915.

[5] Weiling Chen, Keng Teck Ma, Zi Jian Yew, Minhoe Hur, and David Aik-Aun Khoo. 2023. TEVAD: Improved video anomaly detection with captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5548–5558.

[6] Yingxian Chen, Zhengzhe Liu, Baoheng Zhang, Wilton Fok, Xiaojuan Qi, and Yik-Chung Wu. 2023. Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 387–395.

[7] Kai Cheng, Yang Liu, and Xinhua Zeng. 2023. Learning graph enhanced spatial-temporal coherence for video anomaly detection. *IEEE Signal Processing Letters* 30 (2023), 314–318.

[8] Peter Christiansen, Lars N Nielsen, Kim A Steen, Rasmus N Jørgensen, and Henrik Karstoft. 2016. DeepAnomaly: Combining background subtraction and deep learning for detecting obstacles and anomalies in an agricultural field. *Sensors* 16, 11 (2016), 1904.

[9] K Deepak, S Chandrakala, and C Krishna Mohan. 2021. Residual spatiotemporal autoencoder for unsupervised video anomaly detection. *Signal, Image and Video Processing* 15, 1 (2021), 215–222.

[10] Asimenia Dimokranitou. 2017. *Adversarial autoencoders for anomalous event detection in images*. Ph. D. Dissertation. Purdue University.

[11] Fei Dong, Yu Zhang, and Xiushan Nie. 2020. Dual discriminator generative adversarial network for video anomaly detection. *IEEE Access* 8 (2020), 88170–88176.

[12] Keval Doshi and Yasin Yilmaz. 2023. Towards interpretable video anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2655–2664.

[13] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. 2015. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 2758–2766.

[14] Chao Feng, Ziyang Chen, and Andrew Owens. 2023. Self-supervised video forensics by audio-visual anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10491–10503.

[15] Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. 2021. Mist: Multiple instance self-training framework for video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14009–14018.

[16] Alessandro Flaborea, Luca Collorone, Guido Maria D'Amely Di Melendugno, Stefano D'Arrigo, Bardh Prenkaj, and Fabio Galasso. 2023. Multimodal motion conditioned diffusion model for skeleton-based video anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10318–10329.

[17] Mariana-Iuliana Georgescu, Antonio Barbalau, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. 2021. Anomaly detection in video via self-supervised and multi-task learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12742–12752.

[18] Mariana Iuliana Georgescu, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. 2021. A background-agnostic framework with adversarial training for abnormal event detection in video. *IEEE transactions on pattern analysis and machine intelligence* 44, 9 (2021), 4505–4523.

[19] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. 2019. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1705–1714.

[20] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 6546–6555.

[21] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. 2016. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 733–742.

[22] Ryota Hinami, Tao Mei, and Shin'ichi Satoh. 2017. Joint detection and recounting of abnormal events by learning deep generic knowledge. In *Proceedings of the IEEE international conference on computer vision*. 3619–3627.

[23] Xing Hu, Shiqiang Hu, Yingping Huang, Huanlong Zhang, and Hanbing Wu. 2016. Video anomaly detection using deep incremental slow feature analysis network. *IET Computer Vision* 10, 4 (2016), 258–267.

[24] Ammar Mansoor Kamoona, Amirali Khodadadian Gostar, Alireza Bab-Hadiashar, and Reza Hoseinnezhad. 2023. Multiple instance-based video anomaly detection using deep temporal encoding–decoding. *Expert Systems with Applications* 214 (2023), 119079.

[25] Kwang-Eun Ko and Kwee-Bo Sim. 2018. Deep convolutional framework for abnormal behavior detection in a smart surveillance system. *Engineering Applications of Artificial Intelligence* 67 (2018), 226–234.

[26] Nanjun Li, Faliang Chang, and Chunsheng Liu. 2020. Spatial-temporal cascade autoencoder for video anomaly detection in crowded scenes. *IEEE Transactions on Multimedia* 23 (2020), 203–215.

[27] Nannan Li, Jia-Xing Zhong, Xiujun Shu, and Huiwen Guo. 2022. Weakly-supervised anomaly detection in video surveillance via graph convolutional label noise cleaning. *Neurocomputing* 481 (2022), 154–167.

[28] Nannan Li, Jia-Xing Zhong, Xiujun Shu, and Huiwen Guo. 2022. Weakly-supervised anomaly detection in video surveillance via graph convolutional label noise cleaning. *Neurocomputing* 481 (2022), 154–167.

[29] Tong Li, Xinyue Chen, Fushun Zhu, Zhengyu Zhang, and Hua Yan. 2021. Two-stream deep spatial-temporal auto-encoder for surveillance video abnormal event detection. *Neurocomputing* 439 (2021), 256–270.

[30] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. 2013. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence* 36, 1 (2013), 18–32.

[31] Yuanyuan Li, Yiheng Cai, Jiaqi Liu, Shinan Lang, and Xinfeng Zhang. 2019. Spatio-Temporal Unity Networking for Video Anomaly Detection. *IEEE Access* 7 (2019), 172425–172432.

[32] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. 2018. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6536–6545.

[33] Yang Liu, Shuang Li, Jing Liu, Hao Yang, Mengyang Zhao, Xinhua Zeng, Wei Ni, and Liang Song. 2021. Learning Attention Augmented Spatial-temporal Normality for Video Anomaly Detection. In *2021 3rd International Symposium on Smart and Healthy Cities (ISHC)*. IEEE, 137–144.

[34] Yang Liu, Jing Liu, Jieyu Lin, Mengyang Zhao, and Liang Song. 2022. Appearance-Motion United Auto-Encoder Framework for Video Anomaly Detection. *IEEE Transactions on Circuits and Systems II: Express Briefs* 69, 5 (2022), 2498–2502.

[35] Yang Liu, Jing Liu, Wei Ni, and Liang Song. 2022. Abnormal Event Detection with Self-guiding Multi-instance Ranking Framework. In *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 01–07.

[36] Yang Liu, Jing Liu, Kun Yang, Bobo Ju, Siao Liu, Yuzheng Wang, Dingkang Yang, Peng Sun, and Liang Song. 2024. AMP-Net: Appearance-Motion Prototype Network Assisted Automatic Video Anomaly Detection System. *IEEE Transactions on Industrial Informatics* 20, 2 (2024), 2843–2855.

[37] Yang Liu, Jing Liu, Mengyang Zhao, Dingkang Yang, Xiaoguang Zhu, and Liang Song. 2022. Learning Appearance-Motion Normality for Video Anomaly Detection. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.

[38] Yang Liu, Jing Liu, Xiaoguang Zhu, Donglai Wei, Xiaohong Huang, and Liang Song. 2022. Learning Task-Specific Representation for Video Anomaly Detection with Spatial-Temporal Attention. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2190–2194.

[39] Yang Liu, Zhaoyang Xia, Mengyang Zhao, Donglai Wei, Yuzheng Wang, Liu Siao, Bobo Ju, Gaoyun Fang, Jing Liu, and Liang Song. 2023. Learning Causality-inspired Representation Consistency for Video Anomaly Detection. In *Proceedings of the 31st ACM International Conference on Multimedia*. 203–212.

[40] Zhian Liu, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. 2021. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13588–13597.

[41] Cewu Lu, Jianping Shi, and Jiaya Jia. 2013. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*. 2720–2727.

[42] Weixin Luo, Wen Liu, and Shenghua Gao. 2017. Remembering history with convolutional lstm for anomaly detection. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 439–444.

[43] Weixin Luo, Wen Liu, and Shenghua Gao. 2017. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE international conference on computer vision*. 341–349.

[44] Weixin Luo, Wen Liu, Dongze Lian, and Shenghua Gao. 2021. Future frame prediction network for video anomaly detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[45] Weixin Luo, Wen Liu, Dongze Lian, Jinhui Tang, Lixin Duan, Xi Peng, and Shenghua Gao. 2019. Video anomaly detection with sparse coding inspired deep neural networks. *IEEE transactions on pattern analysis and machine intelligence* 43, 3 (2019), 1070–1084.

[46] Hui Lv, Zhongqi Yue, Qianru Sun, Bin Luo, Zhen Cui, and Hanwang Zhang. 2023. Unbiased multiple instance learning for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8022–8031.

[47] Hui Lv, Chuanwei Zhou, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. 2021. Localizing anomalies from weakly-labeled videos. *IEEE transactions on image processing* 30 (2021), 4505–4515.

[48] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. 2010. Anomaly detection in crowded scenes. In *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 1975–1981.

[49] Snehashis Majhi, Ratnakar Dash, and Pankaj Kumar Sa. 2020. Two-Stream CNN architecture for anomalous event detection in real world scenarios. In *International Conference on Computer Vision and Image Processing*. Springer, 343–353.

[50] Jefferson Ryan Medel and Andreas Savakis. 2016. Anomaly detection in video using predictive convolutional long short-term memory networks. *arXiv preprint arXiv:1612.00390* (2016).

[51] Rashmika Nawaratne, Damminda Alahakoon, Daswin De Silva, and Xinghuo Yu. 2019. Spatiotemporal anomaly detection using deep learning for real-time video surveillance. *IEEE Transactions on Industrial Informatics* 16, 1 (2019), 393–402.

[52] Trong-Nguyen Nguyen and Jean Meunier. 2019. Anomaly detection in video sequence with appearance-motion correspondence. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1273–1283.

[53] Zhiyuan Ning, Zile Wang, Yang Liu, Jing Liu, and Liang Song. 2024. Memory-enhanced appearance-motion consistency framework for video anomaly detection. *Computer Communications* 216 (2024), 159–167.

[54] Guansong Pang, Cheng Yan, Chunhua Shen, Anton van den Hengel, and Xiao Bai. 2020. Self-trained deep ordinal regression for end-to-end video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12173–12182.

[55] Wen-Feng Pang, Qian-Hua He, Yong-jian Hu, and Yan-Xiong Li. 2021. Violence detection in videos based on fusing visual and audio information. In *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2260–2264.

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

[57] Yujiang Pu and Xiaoyu Wu. 2022. Audio-Guided Attention Network for Weakly Supervised Violence Detection. In *2022 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE)*. IEEE, 219–223.

[58] Mahdyar Ravanbakhsh, Enver Sangineto, Moin Nabi, and Nicu Sebe. 2019. Training adversarial discriminators for cross-channel abnormal event detection in crowds. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1896–1904.

[59] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 779–788.

[60] Mehrsan Javan Roshtkhari and Martin D Levine. 2013. An on-line, real-time learning method for detecting anomalies in videos using spatio-temporal compositions. *Computer vision and image understanding* 117, 10 (2013), 1436–1452.

[61] Mohammad Sabokrou, Mohsen Fayyaz, Mahmood Fathy, and Reinhard Klette. 2017. Deep-cascade: Cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes. *IEEE Transactions on Image Processing* 26, 4 (2017), 1992–2004.

[62] Yimeng Shang, Xiaoyu Wu, and Rui Liu. 2022. Multimodal Violent Video Recognition Based on Mutual Distillation. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, 623–637.

[63] Prakhar Singh and Vinod Pankajakshan. 2018. A Deep Learning Based Technique for Anomaly Detection in Surveillance Videos. In *2018 Twenty Fourth National Conference on Communications (NCC)*. 1–6.

[64] Rituraj Singh, Anikeit Sethi, Krishanu Saini, Sumeet Saurav, Aruna Tiwari, and Sanjay Singh. 2024. Attention-guided generator with dual discriminator GAN for real-time video anomaly detection. *Engineering Applications of Artificial Intelligence* 131 (2024), 107830.

[65] Sorina Smeureanu, Radu Tudor Ionescu, Marius Popescu, and Bogdan Alexe. 2017. Deep appearance features for abnormal behavior detection in video. In *International Conference on Image Analysis and Processing*. Springer, 779–789.

[66] Waqas Sultani, Chen Chen, and Mubarak Shah. 2018. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6479–6488.

[67] Shengyang Sun and Xiaojin Gong. 2023. Hierarchical semantic contrast for scene-aware video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 22846–22856.

[68] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W Verjans, and Gustavo Carneiro. 2021. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4975–4986.

[69] Boyang Wan, Yuming Fang, Xue Xia, and Jiajie Mei. 2020. Weakly supervised video anomaly detection via center-guided discriminative learning. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.

[70] Tian Wang, Meina Qiao, Zhiwei Lin, Ce Li, Hichem Snoussi, Zhe Liu, and Chang Choi. 2018. Generative neural networks for anomaly detection in crowded scenes. *IEEE Transactions on Information Forensics and Security* 14, 5 (2018), 1390–1399.

[71] Xuanzhao Wang, Zhengping Che, Bo Jiang, Ning Xiao, Ke Yang, Jian Tang, Jieping Ye, Jingyu Wang, and Qi Qi. 2021. Robust unsupervised video anomaly detection by multipath frame prediction. *IEEE transactions on neural networks and learning systems* (2021).

[72] Yang Wang, Tianying Liu, Jiaogen Zhou, and Jihong Guan. 2023. Video anomaly detection based on spatio-temporal relationships among objects. *Neurocomputing* 532 (2023), 141–151.

[73] Donglai Wei, Yang Liu, Xiaoguang Zhu, Jing Liu, and Xinhua Zeng. 2022. MSAF: Multimodal Supervise-Attention Enhanced Fusion for Video Anomaly Detection. *IEEE Signal Processing Letters* 29 (2022), 2178–2182.

[74] Jhih-Ciang Wu, He-Yen Hsieh, Ding-Jie Chen, Chiou-Shann Fuh, and Tyng-Luh Liu. 2022. Self-supervised Sparse Representation for Video Anomaly Detection. In *European Conference on Computer Vision*. Springer, 729–745.

[75] Peng Wu, Jing Liu, and Fang Shen. 2019. A deep one-class neural network for anomalous event detection in complex scenes. *IEEE transactions on neural networks and learning systems* 31, 7 (2019), 2609–2622.

[76] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. 2020. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *European conference on computer vision*. Springer, 322–339.

[77] Peng Wu, Xiaotao Liu, and Jing Liu. 2022. Weakly supervised audio-visual violence detection. *IEEE Transactions on Multimedia* (2022).

[78] Dan Xu, Yan Yan, Elisa Ricci, and Nicu Sebe. 2017. Detecting anomalous events in videos by learning deep representations of appearance and motion. *Computer Vision and Image Understanding* 156 (2017), 117–127.

[79] Cheng Yan, Shiyu Zhang, Yang Liu, Guansong Pang, and Wenjun Wang. 2023. Feature prediction diffusion model for video anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5527–5537.

[80] Zhiwei Yang, Jing Liu, Zhaoyang Wu, Peng Wu, and Xiaotao Liu. 2023. Video event restoration based on keyframes for video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14592–14601.

[81] Muchao Ye, Xiaojiang Peng, Weihao Gan, Wei Wu, and Yu Qiao. 2019. Anopcn: Video anomaly detection via deep predictive coding network. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1805–1813.

[82] Guang Yu, Siqi Wang, Zhiping Cai, En Zhu, Chuanfu Xu, Jianping Yin, and Marius Kloft. 2020. Cloze test helps: Effective video anomaly detection via learning to complete video events. In *Proceedings of the 28th ACM International Conference on Multimedia*. 583–591.

[83] Jongmin Yu, Younkwan Lee, Kin Choong Yow, Moongu Jeon, and Witold Pedrycz. 2021. Abnormal event detection and localization via adversarial event prediction. *IEEE Transactions on Neural Networks and Learning Systems* (2021).

[84] Jiashuo Yu, Jinyu Liu, Ying Cheng, Rui Feng, and Yuejie Zhang. 2022. Modality-Aware Contrastive Instance Learning with Self-Distillation for Weakly-Supervised Audio-Visual Violence Detection. In *Proceedings of the 30th ACM International Conference on Multimedia*. 6278–6287.

[85] Muhammad Zaigham Zaheer, Jin-ha Lee, Marcella Astrid, and Seung-Ik Lee. 2020. Old is gold: Redefining the adversarially learned one-class classifier training paradigm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14183–14193.

[86] Muhammad Zaigham Zaheer, Arif Mahmood, Marcella Astrid, and Seung-Ik Lee. 2020. Claws: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection. In *European Conference on Computer Vision*. Springer, 358–376.

[87] M Zaigham Zaheer, Arif Mahmood, M Haris Khan, Mattia Segu, Fisher Yu, and Seung-Ik Lee. 2022. Generative Cooperative Learning for Unsupervised Video Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14744–14754.

[88] Muhammad Zaigham Zaheer, Arif Mahmood, Hochul Shin, and Seung-Ik Lee. 2020. A self-reasoning framework for anomaly detection using video-level labels. *IEEE Signal Processing Letters* 27 (2020), 1705–1709.

[89] Chen Zhang, Guorong Li, Yuankai Qi, Shuhui Wang, Laiyun Qing, Qingming Huang, and Ming-Hsuan Yang. 2023. Exploiting completeness and uncertainty of pseudo labels for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16271–16280.

[90] Dasheng Zhang, Chao Huang, Chengliang Liu, and Yong Xu. 2022. Weakly Supervised Video Anomaly Detection via Transformer-Enabled Temporal Relation Learning. *IEEE Signal Processing Letters* (2022).

[91] Mengyang Zhao, Yang Liu, Jing Liu, and Xinhua Zeng. 2022. Exploiting Spatial-temporal Correlations for Video Anomaly Detection. In *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 1727–1733.

[92] Yiru Zhao, Bing Deng, Chen Shen, Yao Liu, Hongtao Lu, and Xian-Sheng Hua. 2017. Spatio-temporal autoencoder for video anomaly detection. In *Proceedings of the 25th ACM international conference on Multimedia*. 1933–1941.

[93] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H Li, and Ge Li. 2019. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1237–1246.

[94] Joey Tianyi Zhou, Jiawei Du, Hongyuan Zhu, Xi Peng, Yong Liu, and Rick Siow Mong Goh. 2019. Anomalynet: An anomaly detection network for video surveillance. *IEEE Transactions on Information Forensics and Security* 14, 10 (2019), 2537–2550.

[95] Shifu Zhou, Wei Shen, Dan Zeng, Mei Fang, Yuanwang Wei, and Zhijiang Zhang. 2016. Spatial–temporal convolutional neural networks for anomaly detection and localization in crowded scenes. *Signal Processing: Image Communication* 47 (2016), 358–368.

[96] Yi Zhu and Shawn Newsam. 2019. Motion-aware feature for improved video anomaly detection. *arXiv preprint arXiv:1907.10211* (2019).