Weakly-Supervised and Unsupervised Video Anomaly Detection

Yide Song

School of Computer Science, South China Normal University, Guangzhou, Guangdong 510631, China

yide.song@m.scnu.edu.cn

Abstract. As surveillance technology is continuously improving, an ever-increasing number of cameras are being deployed everywhere. Relying on manual detection of anomalies through cameras may be unreliable and untimely. Therefore, the application of deep learning in video anomaly detection is being extensively studied. Anomaly Detection (AD) refers to identifying events that deviate from the desired actions. This article discusses representative unsupervised and weakly-supervised learning methods applied to various data types. In these machine learning methods, Generative Adversarial Network, Auto Encoder, Recurrent Neural Network, etc. are broadly adopted for AD. Some renowned and new datasets are reviewed. Furthermore, we also proposed several future directions of research in video anomaly detection.

Keywords: Anomaly Detection, Novelty Detection, Video Object Detection, Weakly-Supervised Learning.

1. Introduction

Anomaly detection is defined as a problem of finding patterns in data that do not conform to expected behavior. Video anomaly detection (VAD) is currently an active research field due to the public's concern with security. VAD is regarded to localize anomalous events in videos spatially or temporally. It is suitable for security applications. With the increasing number of security cameras, it is time-consuming to identify anomalies in the videos by manual work. Moreover, the fact that normal events happen far more often than abnormal ones and the anomalies are defined based on the context which contains many uncertainties make the task challenging. A typical solution to detect anomalies is to utilize an unsupervised learning model which is trained on normal data by learning the patterns and trends of anomalies. Events are recognized to be abnormal if the trained model treats them as outliers.

Traditional methods [2-8] of VAD requires a multitude of handcrafted features and those features are often task-specific. These methods rely heavily on human intervention. For example, securities in the supermarket need to stare at multiple screens all the time to alert others to the danger. Hence, a series of deep learning (DL) based VAD methods and frameworks are proposed and achieve better performance and generality.

In the real-world application, a smashing number of videos are recorded by surveillance cameras and labeling all these raw videos are laborious. Hence, unsupervised anomaly detection is a viable solution for VAD which generally learns the representation of normal videos by reconstructing the features. The trained model will have a larger reconstruction error on unseen anomalous images and videos.

Some researchers proposed generative models like Generative Adversarial Networks (GAN) [9-14] and AutoEncoder (AE) [11], [15-17] to capture the normality underlying the given data. The underlying intuition behind is that anomalies are harder to be reconstructed. AnoGAN [9] is a typical generative method and EBGAN [10] borrows the concept of energy. Clustering is a widely used technique [11, 18, 19] to push the anomalies away from normal instances and minimize the distance between normal instances. To use features of neighboring frames, some proposed the Recurrent Neural Network (RNN)-based model to store key information in the model even if taking images as input. Others consider directly feeding video sequences into the network to better extract features [11,

15, 20-24]. Some ordering methods like [25] are based on observable ordinal variables associated with the ordering of abnormality. However, it may fail to generalize to unknown anomalies which differ from abnormal features in the training dataset because the model is well fit to detect the labeled anomalies.

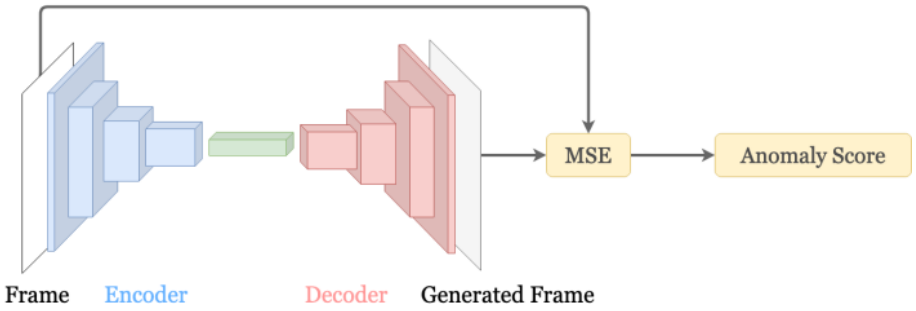

Figure 1. General Architecture of AutoEncoder

Table 1. Comparison of Methods Reviewed. DA and Sup stand for Data Augmentation and Supervision. MAE, MSE, CE, AE and MIL are short for Mean Absolute Error, Mean Squared Error, Cross-Entropy, AutoEncoder and Multiple Instance Learning

Method	DA	Pretrained	Loss	Architecture	Sup.	Data Type
OR [26]	N	ResNet50	BCE	FCN	Unsup.	Video
MOCCA [20]	N	/	MSE	AE & LSTM	Unsup.	Image
SmithNet [23]	N	FlowNet2	GAN Loss	AE	Unsup.	Image
ALAD [42]	N	/	GAN Loss	GAN	Semi.	Image
ANGAN [31]	N	/	MAE&CE	GAN	Unsup.	Image
RTFM [26]	N	C3D, I3D	BCE-based Loss	FCN	Weakly.	Video
OCCNN [7]	N	ResNet-50, YOLOv3	Joint Loss[1]	CNN	Unsup.	Video
ALOCC [29]	N	/	MSE	AE & CNN	Semi.	Image
CutPaste [12]	Y	EfficientNet [34]	BCE	CNN & GDE[2]	Unsup.	Image
WETAS [11]	N	ResNet-34 [9]	BCE & Alignment Loss	CNN	Weakly.	Video
CTRFD [38]	N	I3D	BCE	CNN & FCN	Weakly.	Video
HF2-VAD [16]	N	/	MSE & CE	AE	Unsup.	Video
STAD [2]	Y	ResNet-18	MSE	CNN	Semi.	Image
GTAD [8]	Y	/	MSE	CNN	Unsup.	Image
PTM [40]	N	ResNet-18	MAE & CE	CNN & LSTM	Unsup.	Video
SSAGAN [10]	Y	/	CE	GAN	Unsup.	Video
EBGAN [43]	N	/	MSE	GAN	Unsup.	Image
OCGAN [27]	N	/	BCE	AE	Semi.	Image
MIRF [32]	N	C3D	MIL Ranking	FCN	Weakly.	Video
MSL [13]	N	C3D, I3D	BCE	CNN	Weakly.	Video
ROADMAP [37]	N	VGG16	MSE	CNN	Unsup.	Video
MIST [6]	Y	C3D, I3D	MIL Ranking	AE & CNN	Weakly.	Video
KCS [3]	N	/	MSE	U-Net	Unsup.	Video
CRFs [28]	N	ResNet-50	BCE & MSE	TRN[3]	Weakly.	Video
FenceGAN [22]	N	/	ℓ_2 & BCE	GAN	Semi.	Image
E^3 Outlier [36]	Y	WideResNet [41]	CE	CNN	Unsup.	Image

[1] The Joint Loss consists of Cross-Entropy and ℓ_1 Loss

[2] Gaussian Density Estimation

[3] Temporal Relational Network [44]

In real-world practice, researchers are seeking methods that work on limited labeled data because labeled video data are laborious to get. As the result, deep weakly supervised anomaly detection (WAD) [21-23], [26] and semi-supervised anomaly detection (SAD) [12, 13, 27, 28] are proposed to detect anomalies. WAD aims at building the anomaly detection model with data that is partially or incorrectly labeled. SAD aims at working on limited labeled data. These pseudo or noisy labels offer crucial information about anomalies which enhances the low-dimensional feature representation and accuracy. WAD and SAD are important if a small amount of anomaly data is given [29, 30]. It should also be noted that the anomaly classes are not exactly pre-defined. i.e., some classes in the application are unseen in training. However, due to these inaccurate or inexact data often cost little to grab in various scenarios [31], WAD and SAD play a crucial role in the practice of VAD.

Considerable researches are made and have gained a huge success on datasets like UCSD Ped1 Ped2 [4, 6], CUHK Avenue [5], ShanghaiTech [32] and UCF-Crime [23].

In this paper, we will focus on weakly-supervised learning and unsupervised learning. Mainstream methodologies will be presented.

2. Image Data Based Methods

2.1. Unsupervised Methods

Due to the high cost of fetching the labeled data, some researchers proposed anomaly detection based on unsupervised learning. In this section, some representative unsupervised methods for anomaly detection will be presented.

Generative Adversarial Network (GAN) is composed of two parts: the generator and the discriminator. The generator takes a vector randomly chosen from the learned distribution to generate a fake item and is trained to fool the discriminator to classify the fake items as real. The discriminator is trained to classify both true and fake items. The intuition behind GAN-based AD is that the anomalous image can be separated from the normal one.

AnoGAN [9] were proposed to identify anomalous images and segment anomalous regions within image data. AnoGAN has one generator and discriminator. The generator is trained on healthy data to learn the latent feature presentation of them using residual loss function. The discriminator uses an improved discrimination loss based on feature matching which is noted as $\mathcal{L}_D(\mathbf{z}_\gamma) = \sum \left|\mathbf{f}(\mathbf{x}) - \mathbf{f}\left(G(\mathbf{z}_\gamma)\right)\right|$, where $G(\mathbf{z}_\gamma)$ is the output of the generator and $f(\cdot)$ is the output of the intermediate layer of the discriminator which is used to specify the statistics of the input. The output of the model, anomaly score, is calculated as $\mathrm{A}(\mathbf{x}) = (1-\lambda)\cdot \mathrm{R}(\mathrm{x}) + \lambda \cdot \mathrm{D}(\mathrm{x})$ where $\mathrm{R}(\mathrm{x})$ is the residual score and $\mathrm{D}(\mathrm{x})$ is the discrimination score.

Energy-based Generative Adversarial Network (EBGAN) [10] is an energy-based model which assigns low energies to the desired sample while high energies to the incorrect ones. In the scenarios of VAD, the outliers will cause energy pull-ups. Zhao et al. treat the discriminator as an energy function that is used to calculate the reconstruction error of autoencoder architecture. The generator learns to produce samples in regions of the space where the discriminator assigns low energy.

On the contrary, Golan et al. [33] considered to utterly bypass the reconstruction error. The approach that learns to discriminate between a series of geometric transformations of normal images and it is shown in the experiment that the learning of these features is useful for AD.

There is also another form of unsupervised learning, which is clustering-based. Wang et al. [18] proposed $E^3\ Outlier$, a framework that uses surrogate supervision to create numerous pseudo-classes from original unlabeled data to learn better feature representation. The model learns to detect outliers unsupervised according to inlier priority. The inlier priority-based score is used to prioritize the reduction of inliers' loss during the training process and identify whether an input is an outlier. Xie et al. [34] proposed a data-driven approach to solve for the feature space and cluster memberships jointly. A parameterized non-linear mapping from the data space X to a lower-dimensional feature

space Z is optimized for the clustering objective. Deep Embedded Clustering (DEC) is used to iteratively refine clusters with an auxiliary target distribution derived from the current soft cluster assignment.

Instead of treating the network as a single block, Massoli et al. [16] propose Multilayer One-Class Classification for Anomaly Detection (MOCCA), which integrates multiple layers to learn the extracted representations of videos. In the meanwhile, the autoencoder is only used for the reconstruction task.

Recurrent architecture is also used for AD. In [17], SmithNet works well with image data, which extracts and accumulates the semantic information embedded in frames by the recurrent neural network. SmithNet specializes in the scene and the object texture of instant motion. The researchers investigate the motion features hidden in a single frame.

Li et al. proposed CutPaste [35], a framework focuses on a detailed form of AD. The one-class detect detection methods are applied to the datasets with various unknown anomalous patterns shown in high-resolution images. Li et al. divide AD into a two-stage task. One is self-supervised representation learning by several surrogate tasks. Another is the classification task to distinguish between anomalous and normal patterns by the generative classifier. Noteworthy, a data augmentation strategy, which cuts image segments and pastes them at an arbitrary location, is utilized to enhance the generalizability by manually producing spatial irregularity.

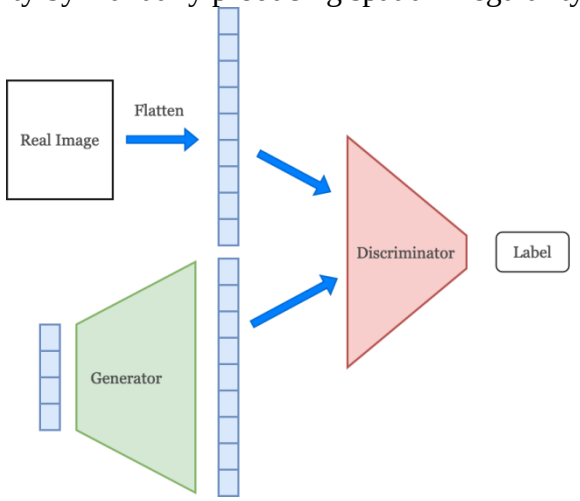

Figure 2. Flow of GAN in Video Anomaly Detection

2.2. Semi-Supervised Methods

Methods like [23] achieve AD in an offline manner which doesn't meet the real-time needs. Zenati et al. [14] provided optimization for GAN-based AD methods by jointly learning an encoder network during training.

Similarly, the adversarial learning method [28] proposed by Sabokrou et al. includes the auto-encoder in the classification. Inspired by GAN, Sabokrou et al. consider a framework that comprises two modules and train it in an end-to-end fashion. Specifically, the detector $\mathcal{D}$ learns the latent representation of positive samples and distinguishes those samples from the reconstructed ones. The reconstructor $\mathcal{R}$ learns to efficiently reconstruct the positive samples and for novelty samples, it cannot do it accurately. In general, the whole learning process of the one-class classifier needs no novelty class data.

The above-mentioned GAN-based methods may not handle the high-dimensional data well. Fence GAN [12] is proposed to solve this. Fence GAN exploits GAN's ability to make the generated samples lie at the boundary of the distribution drawn by real ones instead of overlapping them. The encirclement loss and dispersion loss are introduced during learning to enclose the real data with the generated ones.

However, when training the GAN, many previous works only focus on learning the latent representation of normal samples but ignore that the abnormal samples should be poorly reconstructed.

One-Class GAN(OCGAN) [13], a two-fold latent space learning method is proposed to solve this. OCGAN consists of four components: an auto-encoder, two discriminators and a classifier. The auto-encoder is injected with noise. To avoid over-fitting and improve generalizability. One discriminator called latent discriminator is trained to differentiate between latent representations of real samples and those samples from uniform distribution $U(-1,1)^d$. The aim of this is to force the latent representation of real samples to be distributed uniformly. Another discriminator is called a visual discriminator which is used to differentiate between images of known classes and samples generated from the decoder. The role of the classifier is to label the reconstructions of in-class samples as positive and the generated fake ones as negative.

Bergmann et al. [27] turn the anomaly detection problem into a feature regression problem by implicitly learning the distribution of training features. Their teacher-students model is trained in an end-to-end fashion. The teacher is a feature extractor that generates surrogate labels for students. The student model is trained to mimic teachers' output given the labels optimized according to the anomaly scores on the pixel level. Then, the regression error and students' predictive variance can be used to produce the dense anomaly maps for drawing the anomalous regions in original images.

3. Video Data Based Methods

In the era of big data, tons of video data are recorded and marking them is laborious and unrealistic. The existed image-based methods mentioned above don't take the video-specific features into consideration, which will cause the model unreliable, not to mention there's also a need for real-time application. Beyond that, training those models has a high computational cost which is not worth it. In the following section, numerous representative methods will be reviewed. Some are weakly-supervised and others are unsupervised.

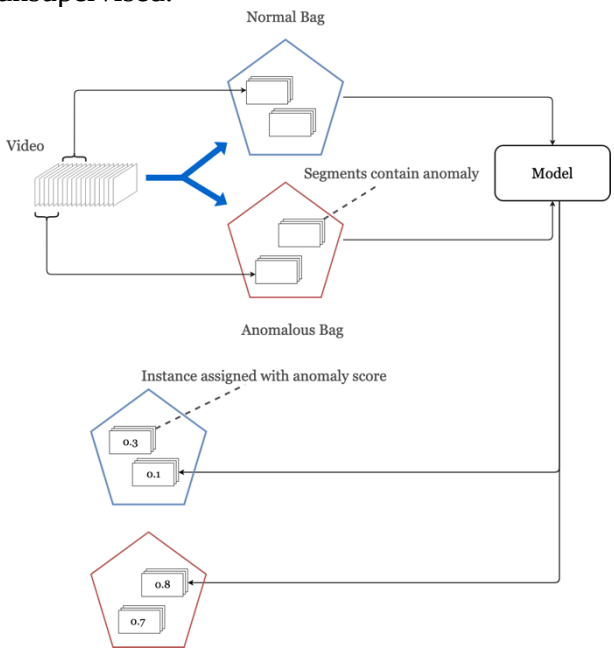

Figure 3. Weakly-Supervised Method for Multiple Instance Learning

3.1. Weakly Supervised Learning

Weakly supervised methods aim at leveraging the limited or inaccurately labeled anomalous video. The majority of the videos are normal while only a small number of them contain anomalies, which cause the imbalance of the data. Furthermore, the inexact anomaly labels are often inexpensive to grab in certain scenarios.

To use video-level instead of clip-level data, Sultani et al. [23] present a deep multiple instances ranking framework, which utilizes the weakly labeled training videos. Sultani et al. divide normal and anomalous video segments into two bags. Anomalous video segments must contain anomalous

events but when such events occur and the length of it is not deterministic. The Multiple Instance Learning (MIL) loss with sparsity and temporal smoothness is proposed to better localize anomalies. The loss function encourages high anomalous scores for anomalous video segments. Due to the segment-level labels are not available, the ranking objective function will be: $\max_{i\in\mathcal{B}_a} f(\mathcal{V}_a^i) > \max_{i\in\mathcal{B}_n} f(\mathcal{V}_n^i)$ to enforce ranking only on the two segments with highest anomaly score in the positive and negative bags. However, the segment with the highest score in the negative bag may be incorrectly considered an anomalous segment, which should be avoided. Also, the anomalous segments in the negative bag should be sparse and the difference between adjacent video segments should not be large. To consider the problems outlined above, the modified MIL loss is noted as:

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

Where $\mathcal{B}_a$ and $\mathcal{B}_n$ are abnormal and normal bag, $\mathcal{V}_a$ and $\mathcal{V}_n$ are anomalous and normal video segments. By training on this, the model will predict high scores for anomalous segments in positive bags.

Sometimes, the model cannot work well with the samples including more than one abnormal snippet and the weak video-level labels don't provide training signals that necessarily separate the normal and anomalous samples. To resolve it, Tian et al. [26] deploy a method named Robust Temporal Feature Magnitude learning (RTFM) where features with low magnitude are identified as normal (negative) samples. Assuming that abnormal samples have the greater mean feature magnitude, the model trains a classifier utilizing k samples with top anomalous scores. The method successfully incorporates long and short-range temporal dependencies seamlessly with a pyramid of dilated convolutions (PDC) and a temporal self-attention module (TSA)

Feng et al. [20] propose a multiple-instance self-training framework (MIST) to efficiently refine task-specific representations that help to discriminate between different types of videos. MIST is composed of a multiple-instance pseudo label generator and a self-guided attention-boosted feature encoder. The generator applies a sparse continuous sampling strategy to generate reliable clip-level pseudo labels. The encoder is capable of extracting task-specific representations and focusing on anomalous regions automatically.

Transformer-based Multi-Sequence Learning (TMSL) network is proposed, which is effective in learning both snippet-level anomaly scores and video-level anomaly probability. The MSL method does not use a single instance as the optimization unit but a sequence containing multiple instances as an optimization unit and the ranking loss is calculated to select sequences during training. The network is composed of a multi-layer convolutional Transformer encoder to encode extracted snippet features, a video classifier to predict video-level anomaly scores, and a snippet regression to predict snippet-level anomaly scores. Because the task is to predict fine-grained anomaly scores, a two-stage self-training strategy is adopted to refine the anomaly scores.

To gain a more discriminative model utilizing local and global features and their short-range correlations, Purwanto et al. [19] integrate a self-attention module containing conditional random fields (CRFs) with CNN. The relation-aware feature extractor extending the temporal relational network (TRN) is also included to extract multiscale CNN features. Furthermore, the contrastive MIL scheme enables the model to maximize the margin between normal and abnormal samples.

Lee et al. innovatively considers the segment-level anomaly detection with the dynamic length of videos. WETAS [21] is proposed to incorporate sequential pseudo-labels with the input instances based on the Dynamic Time Warp Alignment (DTW). In MIL, the continuity of detected points should be emphasized for further interpretation of results.

In [24], the temporal relation and discrimination by learned features are tackled by two tasks. One consists of using causal temporal relation (CTR) module which only captures the current and

historical video frames within a short range of time. The CTR module simultaneously digs the temporal relationships of semantic similarity and positional prior. Another comprises two auxiliary modules: the compactness module (CP) and the dispersion module (DP). The modules are applied to normal and abnormal videos respectively to best separate normal and abnormal features and minimize the distance between normal features in anomalous and normal videos.

3.2. Unsupervised Learning

Liu et al. proposed HF^2-VAD [15], a framework incorporating flow reconstruction and frame prediction. The framework comprises two essential modules: ML-MemAE-SC (Multi-Level Memory modules in an Autoencoder with Skip Connections) and CVAE (Conditional Variational Autoencoder). ML-MemAE-SC is used to memorize inner patterns of normal videos for optical flow reconstruction such that the anomalous videos will come with larger errors. CVAE captures the correlation between the video frame and optical flow to predict the next frame from several previous frames.

In [36], VAD is achieved by four sub-tasks (i.e., proxy tasks). The object in the video will be extracted utilizing a pre-trained YOLOv3 model before being sent to proxy tasks. The first task is predicting the arrow of time, which is discriminating between onward and backward moving objects. The second task is to predict the irregularity of motion (whether the object is captured in consecutive frames). The third task is reconstructing the appearance of the objects in the middlebox given preceding and succeeding frames. Task 4 is named model distillation. YOLOv3 and ResNet-50 pre-trained on ImageNet are used as teachers. By jointly addressing the tasks, we can finally get the anomaly score of the video for AD.

Pang et al. [25] turn VAD into a surrogate two-class ordinal regression task. This end-to-end approach enables the feature extractor to learn AD-tailored features in training data. It first utilizes existing generic (not video-specific) AD to produce pseudo labels and then stack differentiable feature representation for end-to-end training. The output is then passed through pre-trained ResNet-50 to get the anomaly scores which are used to iteratively optimize and improve the detection performance. Ultimately, the model offers the localization of the identified anomalies within the corresponding images.

Huang et al. [37] propose a self-supervised attentive generative adversarial network (SSAGAN). The method enhances existing deep generative model (DGM) methods by enlarging the gap of anomaly scores. There are majorly two parts to the method. One is introducing a self-supervised rotation degree detection task that enables the generator to learn semantic information inside the normal frames. Another is adding the self-attention mechanism to SSAGAN which captures a wide range of contextual information to improve prediction. It comprises the self-attentive predictor, the vanilla discriminator, and the self-supervised discriminator. The self-attentive generator first receives numerous consecutive frames to predict the future frame and the generated frame will be rotated four degrees $(0°, 90°, 180°, 270°)$. The vanilla prediction errors and auxiliary rotation detection loss are utilized to train a more discriminative model. The vanilla discriminator is to perform the binary classification while the auxiliary self-supervised discriminator is to handle the task like rotation degree detection.

Xu et al. [38] propose a probabilistic temporal modeling framework containing three modules: Probabilistic Annotation Modeling (PAM) to generate temporal label distributions, Temporal Label Aggregation (TLA) to refine that distribution and Dense Probabilistic Localization (DPL) to generate supervision signal for training. PAM applies a probability distribution to model the determined temporal location. Points in the distribution signify the possibility of unintentional behavior occurring. DPL contains three different modules: (1) Probabilistic dense classification (Pdc): quantifies the probability of action transition between intentional and unintentional behavior. (2) Probabilistic temporal detection (Ptd): quantifies the temporal boundary of unintentional behavior. (3) Probabilistic regression (Pr): localizes the temporal location of behavior transitions by calculating the cumulative probability differences

Chang et al. [11] innovatively combines multiple techniques for AD. K-means cluster for joint optimization of K-means cluster and representation contatenation of spatial and motion features. Maximization of the length of temporal information by calculating the residual between the first and the last individual frame. Variance attention module to weigh the importance of features in a sliding window fashion. Combination of reconstruction error and cluster distance for anomaly evaluation and conclude it into a pixel-level score.

Wang et al. propose ROADMAP [39]. The ROADMAP contains multipath ConvGRUs that emphasizes on prioritize the attention to static background part and real anomalies. Furthermore, ConvGRUs are used to extract informative parts of different scales and learn the temporal relation of neighboring frames.

4. Real-World Datasets

In this section, commonly used datasets for video anomaly detection are listed.

Table 2. Overview of Commonly Used Datasets

Name	Anomaly	Labeled	Type	Size
UCSD Ped 1 [6]	28.60%	Y	Video	~14000 frames
UCSD Ped 2 [4]	35.90%	Y	Video	~5600 frames
UMN Unusual Crowd Activity	/	N	Video	~7710 frames
CUHK Avenue [5]	12.46%	Y	Video	30652 frames
Shanghai Tech Campus dataset [32]	5.38%	Y	Video	317398 frames
UCF Crime [23]	/	Y	Video	~13769300 frames
OOPS [40]	/	Y	Video	~50 hours

UCSD Ped 1 [6]: The dataset contains clips of groups of people walking away from or towards the camera which is mounted at an elevation, overlooking pedestrian walkways.

UCSD Ped 2 [4]: The dataset contains scenes with pedestrian movement parallel to the camera plane.

UMN Unusual Crowd Activity: The dataset includes scenes with crows running in one direction or dispersing from a central point.

CUHK Avenue [5]: The dataset contains videos captured in CUHK campus avenue.

ShanghaiTech Campus dataset [32]: The dataset contains videos that have a wide range of scene and view angles.

UCF Crime [23]: The dataset consists of untrimmed real-world surveillance videos with realistic anomalies such as fighting, road accident, etc.

OOPS [40]: The Oops dataset covers numerous unintentional actions like physical and social errors, errors in planning and execution, etc.

5. Future Directions

5.1. Better and Complex Datasets

A multitude of models are developed and achieve notable training and testing accuracy. However, many existing datasets mentioned above cover only a small range of scenarios. UMN dataset's snippets have a very limited complexity because the objects in it are only people. The videos in UCSD Ped1 and Ped2 are captured from fixed surveillance cameras and the resolutions are low. There's a need to obtain more complicated datasets like the OOPS, which is much closer to real-world scenarios.

5.2. Optimization for Computation

Self-supervised learning and Semi-supervised learning have achieved fantastic results on VAD. Nowadays, the resolution of video data is higher than before, and the videos are cheap to store. To

gain an effective model for downstream learning tasks, a significant amount of computation on these video data must be made. Hence, it is urgent to find viable ways to better utilize these video data.

5.3. Adaptive Methods

Since the definition of anomaly may changes as time goes on. The instances, which are previously recognized as anomalies may not be anomalous anymore. Therefore, there is a need to find methods to adapt the models to the new definition of anomaly.

5.4. Loitering Anomalies

Many existing methods based on video data take the motion features into account. However, some anomalies in the real world do not have discriminative motion features compared with normal events. Models may fail to detect this kind of anomaly if no special treatments are done.

References

[1] W. Luo et al., "Video Anomaly Detection with Sparse Coding Inspired Deep Neural Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 3, pp. 1070–1084, Mar. 2021, doi: 10.1109/TPAMI.2019.2944377.

[2] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust Real-Time Unusual Event Detection using Multiple Fixed-Location Monitors," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 3, pp. 555–560, Mar. 2008, doi: 10.1109/TPAMI.2007.70825.

[3] 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, Jul. 2013, doi: 10.1016/j.patcog.2012.11.021.

[4] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly Detection and Localization in Crowded Scenes," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 1, pp. 18–32, Jan. 2014, doi: 10.1109/TPAMI.2013.111.

[5] C. Lu, J. Shi, and J. Jia, "Abnormal Event Detection at 150 FPS in MATLAB," in 2013 IEEE International Conference on Computer Vision, Dec. 2013, pp. 2720–2727. doi: 10.1109/ICCV.2013.338.

[6] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Jun. 2010, pp. 1975–1981. doi: 10.1109/CVPR.2010.5539872.

[7] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in 2009 IEEE Conference on Computer Vision and Pattern Recognition, Jun. 2009, pp. 935–942. doi: 10.1109/CVPR.2009.5206641.

[8] V. Saligrama and Zhu Chen, "Video anomaly detection based on local statistical aggregates," in 2012 IEEE Conference on Computer Vision and Pattern Recognition, Jun. 2012, pp. 2112–2119. doi: 10.1109/CVPR.2012.6247917.

[9] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery," in Information Processing in Medical Imaging, 2017, pp. 146–157. doi: 10.1007/978-3-319-59050-9_12.

[10] J. Zhao, M. Mathieu, and Y. LeCun, "Energy-based Generative Adversarial Network," 2016, doi: 10.48550/ARXIV.1609.03126.

[11] Y. Chang et al., "Video anomaly detection with spatio-temporal dissociation," Pattern Recognition, vol. 122, p. 108213, Feb. 2022, doi: 10.1016/j.patcog.2021.108213.

[12] P. C. Ngo, A. A. Winarto, C. K. L. Kou, S. Park, F. Akram, and H. K. Lee, "Fence GAN: Towards Better Anomaly Detection," in 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), Nov. 2019, pp. 141–148. doi: 10.1109/ICTAI.2019.00028.

[13] P. Perera, R. Nallapati, and B. Xiang, "OCGAN: One-Class Novelty Detection Using GANs With Constrained Latent Representations," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2019, pp. 2893–2901. doi: 10.1109/CVPR.2019.00301.

[14] H. Zenati, M. Romain, C.-S. Foo, B. Lecouat, and V. Chandrasekhar, "Adversarially Learned Anomaly Detection," in 2018 IEEE International Conference on Data Mining (ICDM), Nov. 2018, pp. 727–736. doi: 10.1109/ICDM.2018.00088.

[15] Z. Liu, Y. Nie, C. Long, Q. Zhang, and G. Li, "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, 2021, pp. 13588–13597.

[16] F. V. Massoli, F. Falchi, A. Kantarci, S. Akti, H. K. Ekenel, and G. Amato, "MOCCA: Multilayer One-Class Classification for Anomaly Detection," IEEE Transactions on Neural Networks and Learning Systems, pp. 1–11, 2021, doi: 10.1109/TNNLS.2021.3130074.

[17] T.-N. Nguyen, S. Roy, and J. Meunier, "SmithNet: Strictness on Motion-Texture Coherence for Anomaly Detection," IEEE Transactions on Neural Networks and Learning Systems, pp. 1–14, 2021, doi: 10.1109/TNNLS.2021.3116212.

[18] S. Wang et al., "Effective End-to-end Unsupervised Outlier Detection via Inlier Priority of Discriminative Network," in Advances in Neural Information Processing Systems, 2019, vol. 32.

[19] D. Purwanto, Y.-T. Chen, and W.-H. Fang, "Dance with Self-Attention: A New Look of Conditional Random Fields on Anomaly Detection in Videos," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 173–183.

[20] J.-C. Feng, F.-T. Hong, and W.-S. Zheng, "MIST: Multiple Instance Self-Training Framework for Video Anomaly Detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14009–14018.

[21] D. Lee, S. Yu, H. Ju, and H. Yu, "Weakly Supervised Temporal Anomaly Segmentation with Dynamic Time Warping," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 7355–7364.

[22] S. Li, F. Liu, and L. Jiao, "Self-Training Multi-Sequence Learning with Transformer for Weakly Supervised Video Anomaly Detection," p. 9.

[23] W. Sultani, C. Chen, and M. Shah, "Real-World Anomaly Detection in Surveillance Videos," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Jun. 2018, pp. 6479–6488. doi: 10.1109/CVPR.2018.00678.

[24] P. Wu and J. Liu, "Learning Causal Temporal Relation and Feature Discrimination for Anomaly Detection," IEEE Transactions on Image Processing, vol. 30, pp. 3513–3527, 2021, doi: 10.1109/TIP.2021.3062192.

[25] G. Pang, C. Yan, C. Shen, A. van den Hengel, and X. Bai, "Self-Trained Deep Ordinal Regression for End-to-End Video Anomaly Detection," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2020, pp. 12170–12179. doi: 10.1109/CVPR42600.2020.01219.

[26] Y. Tian, G. Pang, Y. Chen, R. Singh, J. W. Verjans, and G. Carneiro, "Weakly-supervised video anomaly detection with robust temporal feature magnitude learning," in Proceedings of the IEEE/CVF international conference on computer vision (ICCV), 2021, pp. 4975–4986.

[27] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "Uninformed Students: Student-Teacher Anomaly Detection with Discriminative Latent Embeddings," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2020, pp. 4182–4191. doi: 10.1109/CVPR42600.2020.00424.

[28] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli, "Adversarially Learned One-Class Classifier for Novelty Detection," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Jun. 2018, pp. 3379–3388. doi: 10.1109/CVPR.2018.00356.

[29] G. Pang, L. Cao, L. Chen, and H. Liu, "Learning representations of ultrahigh-dimensional data for random distance-based outlier detection," in Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, Jul. 2018, pp. 2041–2050. doi: 10.1145/3219819.3220042.

[30] G. Pang, C. Shen, and A. van den Hengel, "Deep anomaly detection with deviation networks," in Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, Jul. 2019, pp. 353–362. doi: 10.1145/3292500.3330871.

[31] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in 2018 IEEE/CVF conference on computer vision and pattern recognition, Jun. 2018, pp. 6479–6488. doi: 10.1109/CVPR.2018.00678.

[32] W. Liu, W. Luo, D. Lian, and S. Gao, “Future Frame Prediction for Anomaly Detection A New Baseline,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6536–6545.

[33] I. Golan and R. El-Yaniv, “Deep anomaly detection using geometric transformations,” in Advances in neural information processing systems, 2018, vol. 31.

[34] J. Xie, R. Girshick, and A. Farhadi, “Unsupervised deep embedding for clustering analysis,” in Proceedings of the 33rd international conference on international conference on machine learning - volume 48, 2016, pp. 478–487.

[35] C.-L. Li, K. Sohn, J. Yoon, and T. Pfister, “CutPaste: Self-Supervised Learning for Anomaly Detection and Localization,” in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2021, pp. 9659–9669. doi: 10.1109/CVPR46437.2021.00954.

[36] M.-I. Georgescu, A. Barbalau, R. T. Ionescu, F. S. Khan, M. Popescu, and M. Shah, “Anomaly Detection in Video via Self-Supervised and Multi-Task Learning,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 12742–12752.

[37] C. Huang et al., “Self-Supervised Attentive Generative Adversarial Networks for Video Anomaly Detection,” IEEE Transactions on Neural Networks and Learning Systems, pp. 1–15, 2022, doi: 10.1109/TNNLS.2022.3159538.

[38] J. Xu, G. Chen, N. Zhou, W.-S. Zheng, and J. Lu, “Probabilistic Temporal Modeling for Unintentional Action Localization,” IEEE Transactions on Image Processing, vol. 31, pp. 3081–3094, 2022, doi: 10.1109/TIP.2022.3163544.

[39] X. Wang et al., “Robust Unsupervised Video Anomaly Detection by Multipath Frame Prediction,” IEEE Transactions on Neural Networks and Learning Systems, pp. 1–12, 2021, doi: 10.1109/TNNLS.2021.3083152.

[40] D. Epstein, B. Chen, and C. Vondrick, “Oops! Predicting Unintentional Action in Video,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 919–929.