Dual-Modality Deep Feature-based Anomaly Detection for Video Surveillance

Parth Lalitkumar Bhatt
Dept. of Computer Science
Lakehead University
Thunder Bay, Canada
bhattp@lakeheadu.ca

Dhruva Shah
Dept. of Computer Science
Lakehead University
Thunder Bay, Canada
dshah33@lakeheadu.ca

Christopher Silver
Dept. of Electrical and Computer Engineering
Lakehead University
Thunder Bay, Canada
crsilver@lakeheadu.ca

Wandong Zhang
Dept. of Electrical and Computer Enggineering
Western University
London, Canada
wzhan893@uwo.ca

Thangarajah Akilan
Dept. of Software Engineering
Lakehead University
Thunder Bay, Canada
takilan@lakeheadu.ca

***Abstract*—Detecting anomalies in videos is not only crucial but also an intriguing task in surveillance systems. It is a sequential modeling problem in nature that requires careful selection of spatial and temporal dependent patterns from a sequence of frames. There are several research works from traditional approaches to modern deep learning-based techniques introduced to address this problem. However, there is a huge demand for research and development to ameliorate the performance of the existing solutions. In response to that, this study proposes an improved video anomaly detection model using deep features extracted from a dual-modality input representation. The proposed model demonstrates effectiveness in the benchmark–UCF crime dataset by achieving the best AUC of 87.52%, which is $\approx$ 12.3% improvement compared to a baseline. The application aspect of this work includes strengthening the security measures in common places, viz. airports, banks, public transits, schools, and shopping complexes by detecting aberrational or suspicious activities in surveillance videos.**

***Index Terms*—anomaly detection, deep learning, dual-modality, video surveillance.**

I. Introduction

Anomaly detection in videos is a challenging task due to the ambiguous nature of what constitutes an anomaly, and the significant amount of information stored in spatiotemporal data frames. However, with an upsurge in crime that threatens public safety there is an increasing demand for smart surveillance systems to detect and respond to anomalous events in real-time. In recent years, deep learning-based video anomaly detection has emerged as a promising solution. Thus, the market of video analytics is expected to grow at a Compound Annual Growth Rate (CAGR) of 21.5%. However, the inherent technical challenges, such as meager data (i.e., ambiguous

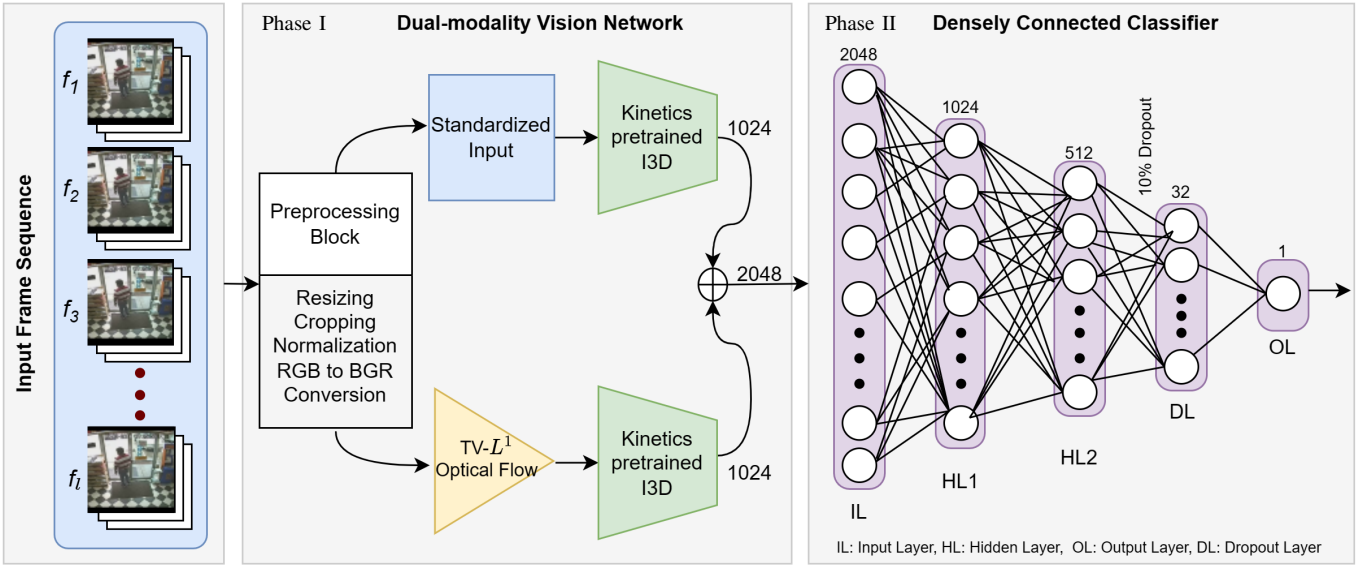

Fig. 1. A complete pipeline of the proposed model. Phase I: a feature extractor subnetwork using a dual-modality vision network, where regular RGB input frames and their respective optical flow information are fed to a pre-trained I3D to extract spatiotemporal cues, and the extracted deep features are concatenated. Phase II: a classification sub-network receives the concatenated deep features and predicts the status of the input video–anomaly vs. normal.

979-8-3503-2397-9/23/$31.00 ©2023 IEEE

data quality–labels are not well-defined or unknown, and lack of annotated samples), high computational complexity, real-time processing requirements, and privacy concerns make it harder to handle. To overcome these challenges, innovative approaches, including data regularization (e.g., data augmentation), model regularization (e.g., transfer learning), and advanced strategies are necessary to develop accurate and robust models that can be applicable to novel scenarios and environments. In this direction, this paper proposes a dual-modality deep feature-based framework using standard RGB and optical flow inputs with Inflated 3D Convolutional Neural Networks (I3D) [1], as shown in Fig. 1.

The rest of this paper is organized into four main sections: related work–Section II presents background information of key existing research works and segregates them into different types, proposed method–Section III elaborates on the solution built in this work, experimental analysis–Section IV provides in-depth quantitative and comparative analysis of the proposed model's performances, and Section V concludes the findings of this work with future directions.

II. Related Work

This section clusters the related works into three groups based on their learning approaches.

A. Transfer Learning

Transfer learning is an efficient technique for information-rich feature extraction [2, 3]. For instance, Sultani *et al.* [4] developed a model for classifying anomalous and normal occurrences in videos through Multiple Instance Learning algorithms (MIL). It creates bundles of videos from the same class and divides them into multiple instances. Using the features extracted via a pre-trained 3-D Convolutional Neural Network (3D-CNN) from these instances the anomaly ranking model is trained. Nazare *et al.* [5] investigated the classification strength of features extracted using a few state-of-the-art pre-trained CNNs in anomaly detection, including VGG-16, ResNet-50, Xception, and DenseNet-121. Although their work is considered to be an experimental study, it provides a benchmark for further research. On the other hand, Zahid *et al.* [6]'s framework is a classic example of a marriage between transfer learning and ensemble learning for video anomaly detection. The model integrates a 3D-CNN as a feature extractor vision network, and a bagging-based ensemble classifier consisting of three shallow Fully Connected (FC) Networks. Similarly, Vu *et al.* [7] implemented a multi-channel framework using multiple Conditional Generative Adversarial Networks (CGANs) to learn feature representations of appearance and motion. It benefits from unifying supervised and unsupervised learning approaches. Zhou *et al.* [8] also combine motion and appearance features under a unified framework wherein an intermediate process fuses the appearance and motion information of moving objects, while a fine-tuned CNN, more specifically a ResNet-50 is deployed to extract useful features from the fused data. Gandapur *et al.* [9] also exploit a RestNet-50 by integrating ResNet-50 and Convolutional GRU (ConvGRU) for video surveillance anomaly detection. Their model records a best performance of 82.22% accuracy on the UCF-Crime dataset.

B. Representative Learning

Representation learning is a collection of techniques that allow a model to automatically identify the representations required for feature detection or classification from raw input data. For instance, Hu *et al.* [10] proposed a deep incremental slow feature analysis (D-IncSFA) network that progressively learns abstract and global high-level representations of anomalous events from raw video sequences. The main merit of the D-IncSFA network is that it has feature extractor and anomaly detector functionality, allowing anomaly detection in a single step, but it fails to identify local anomalies. Zhu *et al.* [11] introduced a temporal cue-augmented model with an attention-based ranking mechanism that learns motion-aware long-range features capturing temporal relationships among input frames. Likewise, Wu and Liu *et al.* [12] investigated a causal convolutional model that learns Causal Temporal Relation (CTR) among features to enhance the representation. Gong *et al.* [13], on the other hand, came up with a Multi-level feature refinement approach, called Multi-scale Continuity-aware Refinement network (MCR). This model pays more attention to the continuity of anomalous instances in different temporal windows. Despite its convincing technique, its effectiveness is restricted to certain domains and records only 81% AUC in the UCF-Crime benchmark dataset. Recently, Patrikar *et al.* [14] leveraged representation learning, specifically using a Convolutional Long Short-Term Memory (C-LSTM) model to predict abnormal events in the future timestamps in surveillance videos.

C. Generative Learning

Generative learning is a probabilistic-based approach that specifies how an instance is formed and can produce fresh instances by sampling from collected historical data samples. Auto-encoders are one such generative model used to identify abnormalities in videos. In this way, Hasan *et al.* [15] developed an end-to-end 2-D Convolutional Autoencoder (CAE) to model regular frames (i.e., normal conditions) from local features. Similarly, Ionescu *et al.* [16] developed an unsupervised learning framework using multiple CAEs to encode object-centric motion and appearance information. The autoencoder-based models built in [17], [18], and [19] subsume a CAE and Convolutional LSTM (ConvLSTM) to capture spatiotemporal features. For instance, Duman and Erdem in [17], extracted speed and trajectory characteristics through a CAE-ConvLSTM architecture and Optical Flow (OF) from frame sequences. Nguyen [20] explored a novel method using UNet type Generative Adversarial Network (GAN) by formulating the anomaly detection task into video generation and object detection. Their model computes the difference between a generated frame and the actual next frame, to detect a probable anomalous condition in the video. The current research on video anomaly detection has several limitations. Firstly,

some approaches either ignore valuable motion information or use datasets with a limited number of anomalies and data. Secondly, the employed loss functions are not suitable for the perfect distinction between normal and anomalous events in videos. Thirdly, many existing methodologies include models with high computational complexity that limit their application for real-time video anomaly detection. To overcome these limitations, (i) new methods should be developed that incorporate motion flow information, (ii) large datasets for training, and (iii) advanced loss functions for better model generalization. By keeping these in mind, this work proposes a dual-modality deep feature-based approach that improves the video anomaly detection accuracy without a huge stall in computational speed.

III. Proposed Method

The proposed model consists of two stages–a dual-modality vision network for key feature extraction (Phase I), and a densely connected classifier (Phase II) as depicted in Fig. 1.

A. Dual-modality Vision Network

1) Pre-processing: The input video stream is processed through the six stages as outlined in the Algorithm 1 before being fed to the feature-extracting vision network.

Algorithm 1: Pre-processing

Input: Raw RGB Frame Sequence
if *Frame is not End-of-Sequence* **then**
- Resizing ← Resize the input frame to a fixed size to ensure uniformity across all frames;
- Cropping ← Apply center and random cropping to the frame;
- Normalization ← Normalize the pixel values of the cropped frame to ensure zero mean and unit variance;
- RGB to BGR Conversion ← Convert the RGB frame to BGR format as required by the feature extractor vision network;
- Optical Flow Computation ← Estimate optical flow on the input video using TV-L^1 algorithm [21] to capture temporal information;
- Optical Flow Normalization ← Normalize the estimated optical flow values to ensure consistency between dual-modality inputs;

end
Output: Input-ready frame to the vision network

2) Feature extraction: Spatiotemporal features from the dual-modality inputs– normalized BGR frame, and normalized TV-L^1 optical flow are extracted using a pre-trained Inflated 3D ConvNet (I3D) [1]. It was pre-trained on the large-scale human action classification dataset–Kinetics.

B. The Densely Connected Classifier

A five-layer FC neural network is optimally constructed as a binary classifier to detect the anomaly videos using the 2048-dimensional spatiotemporal features extracted by the vision network from the dual-modality inputs. The first four layers use ReLU activation function defined in (1) to introduce non-linearity and avoid the vanishing gradient problem. In addition, a 10% dropout regularization is applied to the 4^{th} layer to prevent overfitting. The last layer employs a Sigmoid activation given in (2) to estimate the probability of normal and abnormal events in the input videos.

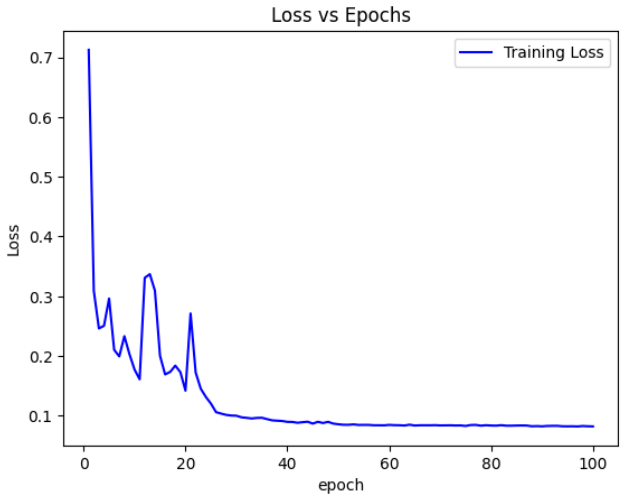

Fig. 2. Training progress of the classifier. The training loss is defined in (3).

TABLE I
Hyperparameter Setting of the Model Training Procedure

Hyperparameter	Value
Optimizer	Adam
Batch size	30
Loss function	MIL as defined in (3)
Learning rate	0.001
Weight decay	1×10^{-12}
Number of epochs	100

$$f(x) = \max(0, x), \tag{1}$$

where x is the input to the activation function and $f(x)$ is the output.

$$\sigma(x) = \frac{1}{1 + e^{-x}}, \tag{2}$$

where x is the input to the function and it can take any real value, and the output is the probability of an event to occur, in this case, an anomaly in the input video.

1) Model initialization: The Xavier method [22] is used to initialize the weights of the classifier subnetwork. It helps to keep the variance of activations and gradients consistent across layers, preventing issues, like vanishing and exploding gradients that often lead to unstable training.

2) Model training: Table I tabulates the hyperparameter configuration used to train the classifier subnetwork. The training procedure takes a mini-batch of samples ($N = 30$) in each iteration and the MIL-ranking loss function ($\mathcal{L}$) as defined in (3) with Adam optimizer to update the parameters.

The MIL-ranking loss function is taken form Sultani *et al.*[4] and Park *et al.*[23].

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\Big[\max\Big(0, 1 - y_{\mathrm{a}_i} + y_{\mathrm{n}_i}\Big) + \lambda_1 \sum_{j=1}^{\ell} y_{\mathrm{a}_{i,j}} + \lambda_2 \sum_{j=1}^{\ell-1}(y_{i,j} - y_{i,j+1})^2\Big], \tag{3}$$

where N, ℓ, and i denote the mini-batch of samples used in each iteration of training, sequence length, and the index of the current sample in the batch, respectively. Hence, y_{a_i} is the maximum predicted anomaly score for the i-th sample, y_{n_i} is the maximum predicted normal score for the i-th sample. The term $\max(0, 1 - y_{\mathrm{a}_i} + y_{\mathrm{n}_i})$ is the hinge loss, which makes sure that the anomaly scores are higher than the normal scores by at least a margin of 1. If this condition is not satisfied, the loss is increased by the amount that the margin is violated. The term $\sum_{j=1}^{\ell} y_{\mathrm{a}_{i,j}}$ is to produce sparse anomaly scores, which means that it should only assign high scores to a small number of segments in the input video, resulting in preventing false alarms. The term $\sum_{j=1}^{\ell-1}(y_{i,j} - y_{i,j+1})^2$ is for smoothing scores across adjacent video segments to eliminate abrupt score changes caused by noise or other artifacts in the video. λ_1 and λ_2 are hyperparameters that control the weight of the sparsity term, and the weight of the smoothness term, respectively. In this case, both λ_1 and λ_2 are set to 8×10^{-5}, based on empirical analysis. Also, in the context of the proposed work, training the model for 100 epochs is sufficient that ensures model convergence. Moreover, this choice of epoch count strikes a balance between computational efficiency and model performance, as demonstrated by the achieved 87.52% AUC on the benchmark UCF crime dataset, showcasing the model's capability to generalize well to real-world surveillance scenarios. The training progress of the classifier is shown in Fig. 2. Note that to avoid the issue of overfitting, early stopping regularization is employed with a patience of 10 and a minimum delta value of 0.001, with training loss set as a criterion.

IV. Experimental Analysis

A. Dataset

The model training and testing are conducted using the benchmark–UCF-Crime [4], which is a large-scale dataset with 1900 untrimmed surveillance videos, totalling a 128 hours of recordings from indoor and outdoor cameras. It has a near-balanced set of samples for normal and anomalous events in both the training and testing sets. The anomalous events include thirteen different anomalies, viz. abuse, arrest, arson, assault, road accident, burglary, explosion, fighting, robbery, shooting, stealing, shoplifting, and vandalism. The dataset contains only frame-level annotations. The training set has 800 normal and 810 anomalous videos, while the testing set has 150 normal and 140 anomalous videos, with all 13 anomalies occurring in various temporal locations, and some videos containing multiple anomalies.

TABLE II
Performance Comparison of the Methods on UCF-Crime Benchmark Dataset. The Best Performances are in Blue Ink.

Method	Features	AUC (%)	% of IMP.
Sultani *et al.*[4]	I3D-RGB	77.92	Baseline
MCR [13]	I3D-MCR	81.00	↑3.9
RTFM [24]	I3D-RGB	84.30	↑8.2
CTR [12]	I3D-CTR	84.89	↑8.9
MGFN [25]	I3D-RGB	86.98	↑11.6
Ours	I3D-RGB + I3D-OF	87.52	↑12.3

IMP - Improvement

B. Evaluation metrics

The Area Under the Curve (AUC) defined in (4) is a commonly used evaluation metric for binary classification problems like in this work. It is calculated as the area under the Receiver Operating Characteristic (ROC) curve, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR).

$$AUC = \int_{-\infty}^{\infty} ROC(\tau)d\tau, \tag{4}$$

where $ROC(\tau)$ represents the ROC curve at a given threshold value τ, and $d\tau$ denotes the differential element in the integral.

C. Comparative analysis

Table II compares the performance of the proposed model with existing recent works, where [4] is considered as the baseline as it also depends on the deep features generated by I3D. While our model outperforms the best existing model, it achieved a 12.3% of improvement over the baseline model.

V. Conclusion and Future Work

The video anomaly detection problem remains an ongoing challenge, requiring new research directions to achieve advanced surveillance systems for enhanced public safety. Improvements are inevitable to the existing solutions. This study proposes a dual-modality deep feature-based model by exploiting regular RGB frames and optical flow information for accurate anomaly detection. An ablation analysis on a benchmark dataset proves the effectiveness of the proposed solution. The future research direction of the proposed model includes improving its robustness to unseen anomalies, enhancing anomaly localization and interpretability, and exploring privacy-preserving methods for ethical surveillance applications. Additionally, investigating domain adaptation, real-time optimization, and multimodal fusion techniques would further advance the model's performance and applicability in diverse surveillance scenarios. Overall, this work highlights the potential of dual-modality deep feature-based anomaly detection for video surveillance and provides a promising direction for future research.

VI. Acknowledgement

This research was enabled in part by support provided by the Digital Research Alliance of Canada (https://alliancecan.ca/en).

References

[1] Joao Carreira and Andrew Zisserman. *Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset*. 2018. arXiv: 1705.07750 [cs.CV].

[2] Thangarajah Akilan et al. "Fusion of transfer learning features and its application in image classification". In: *IEEE 30th Canadian Conf. on Electrical and Compu. Engineering*. 2017, pp. 1–5.

[3] Wandong Zhang et al. "HKPM: A Hierarchical Key-Area Perception Model for HFSWR Maritime Surveillance". In: *IEEE Trans. on Geoscience and Remote Sensing* 60 (2022), pp. 1–13.

[4] Waqas Sultani, Chen Chen, and Mubarak Shah. "Real-world anomaly detection in surveillance videos". In: *Proc. of the IEEE Conf. on compu. vis. and pattern recogni.* 2018, pp. 6479–6488.

[5] Tiago S Nazare, Rodrigo F de Mello, and Moacir A Ponti. "Are pre-trained CNNs good feature extractors for anomaly detection in surveillance videos?" In: *arXiv preprint arXiv:1811.08495* (2018).

[6] Yumna Zahid, Muhammad Atif Tahir, and Muhammad Nouman Durrani. "Ensemble learning using bagging and inception-V3 for anomaly detection in surveillance videos". In: *2020 IEEE Intl. Conf. on Image Processing (ICIP)*. IEEE. 2020, pp. 588–592.

[7] Tuan-Hung Vu et al. "Multi-Channel Generative Framework and Supervised Learning for Anomaly Detection in Surveillance Videos". In: *Sensors* 21.9 (2021), p. 3179.

[8] Joey Tianyi Zhou et al. "Anomalynet: An anomaly detection network for video surveillance". In: *IEEE Trans. on Information Forensics and Security* 14.10 (2019), pp. 2537–2550.

[9] Maryam Qasim Gandapur and Elena Verdú. "ConvGRU-CNN: Spatiotemporal Deep Learning for Real-World Anomaly Detection in Video Surveillance System". In: *Intl. Journal of Interactive Multimedia and Artificial Intelligence* (2023).

[10] Xing Hu et al. "Video anomaly detection using deep incremental slow feature analysis network". In: *IET Compu. vis.* 10.4 (2016), pp. 258–267.

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

[12] Peng Wu and Jing Liu. "Learning Causal Temporal Relation and Feature Discrimination for Anomaly Detection". In: *IEEE Trans. on Image Processing* 30 (2021), pp. 3513–3527.

[13] Yiling Gong et al. "Multi-Scale Continuity-Aware Refinement Network for Weakly Supervised Video Anomaly Detection". In: *2022 IEEE Intl. Conf. on Multimedia and Expo (ICME)* (2022).

[14] Devashree R. Patrikar and Mayur Rajaram Parate. "Anomaly Detection by Predicting Future Frames using Convolutional LSTM in Video Surveillance". In: *2023 2nd Intl. Conf. on Paradigm Shifts in Communications Embedded Systems, Machine Learning and Signal Processing (PCEMS)*. 2023, pp. 1–6.

[15] Mahmudul Hasan et al. "Learning temporal regularity in video sequences". In: *Proc. of the IEEE Conf. on Compu. vis. and pattern recogni.* 2016, pp. 733–742.

[16] Radu Tudor Ionescu et al. "Object-centric auto-encoders and dummy anomalies for abnormal event detection in video". In: *Proc. of the Conf. on Compu. Vis. and Pattern recogni.* 2019, pp. 7842–7851.

[17] Elvan Duman and Osman Ayhan Erdem. "Anomaly detection in videos using optical flow and convolutional autoencoder". In: *IEEE Access* 7 (2019), pp. 183914–183923.

[18] Karishma Pawar and Vahida Attar. "Application of deep learning for crowd anomaly detection from surveillance videos". In: *11th Intl. Conf. on Cloud Comput., Data Science & Engineer.* IEEE. 2021, pp. 506–511.

[19] Anitha Ramchandran and Arun Kumar Sangaiah. "Unsupervised deep learning system for local anomaly event detection in crowded scenes". In: *Multimedia Tools and Applications* 79.47 (2020), pp. 35275–35295.

[20] Khac-Tuan Nguyen et al. "Anomaly detection in traffic surveillance videos with gan-based future frame prediction". In: *Proc. of the 2020 Intl. Conf. on Multimedia Retrieval*. 2020, pp. 457–463.

[21] Christopher Zach, Thomas Pock, and Horst Bischof. "A duality based approach for realtime tv-l 1 optical flow". In: *Pattern Recogni.: 29th DAGM Symposium, Heidelberg, Germany, September 12-14, 2007. Proceedings 29*. Springer. 2007, pp. 214–223.

[22] Xavier Glorot and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks". In: *Proc. of the Thirteenth Intl. Conf. on Artificial Intelligence and Statistics*. Ed. by Yee Whye Teh and Mike Titterington. Vol. 9. Proceedings of Machine Learning Research. Chia Laguna Resort, Sardinia, Italy: PMLR, 2010, pp. 249–256.

[23] Jaeyoo Park, Junha Kim, and Bohyung Han. "Learning to adapt to unseen abnormal activities under weak supervision". In: *The Asian Conf. on Compu. vis.* 2020.

[24] Yu Tian et al. *Weakly-supervised Video Anomaly Detection with Robust Temporal Feature Magnitude Learning*. 2021. arXiv: 2101.10030 [cs.CV].

[25] Yingxian Chen et al. *MGFN: Magnitude-Contrastive Glance-and-Focus Network for Weakly-Supervised Video Anomaly Detection*. 2022. arXiv: 2211.15098 [cs.CV].