Real-World Anomaly Detection in Video Using Spatio-Temporal Features Analysis for Weakly Labelled Data with Auto Label Generation

Original Scientific Paper

Rikin J. Nayak
V T Patel Dept of E & C Engg, Chandubhai S Patel Institute of Technology,
Charotar University of Science and Technology, Changa, Ta-Petlad, Anand, Gujarat 388421, India
rikinnayak@gmail.com, 16drec006@charusat.edu.in

Jitendra P. Chaudhari
Charusat Space Research and Technology Center, V T Patel Dept of E & C Engg,
Chandubhai S Patel Institute of Technology, Charotar University of Science and Technology, Changa,
Ta-Petlad, Anand, Gujarat 388421, India jitendrachaudhari.ec@charusat.ac.in

Abstract *– Detecting anomalies in videos is a complex task due to diverse content, noisy labeling, and a lack of frame-level labeling. To address these challenges in weakly labeled datasets, we propose a novel custom loss function in conjunction with the multi-instance learning (MIL) algorithm. Our approach utilizes the UCF Crime and ShanghaiTech datasets for anomaly detection. The UCF Crime dataset includes labeled videos depicting a range of incidents such as explosions, assaults, and burglaries, while the ShanghaiTech dataset is one of the largest anomaly datasets, with over 400 video clips featuring three different scenes and 130 abnormal events. We generated pseudo labels for videos using the MIL technique to detect frame-level anomalies from video-level annotations, and to train the network to distinguish between normal and abnormal classes. We conducted extensive experiments on the UCF Crime dataset using C3D and I3D features to test our model's performance. For the ShanghaiTech dataset, we used I3D features for training and testing. Our results show that with I3D features, we achieve an 84.6% frame-level AUC score for the UCF Crime dataset and a 92.27% frame-level AUC score for the ShanghaiTech dataset, which are comparable to other methods used for similar datasets.*

Keywords*: anomaly detection, spatio-temporal analysis, 3d convolutional neural network, multi-instance learning*

1. INTRODUCTION

Anomaly detection, or the identification and classification of data patterns that deviate from normal patterns, is a crucial aspect of intelligent visual surveillance systems. The deployment of CCTV cameras has become widespread and more affordable,which has resulted in increased research attention on video-based anomaly detection. Deployment of CCTV cameras is particularly important for ensuring security in public areas such as railway stations, hospitals, and military bases. With the increasing availability of powerful computing resources, Artificial Intelligence and Deep learning have been integrated into smart video surveillance systems to efficiently process and analyze vast amounts of video data. In this paper, we employ a multi-instance learning technique to address the challenges of anomaly detection in videos, using a custom loss function under weakly supervised learning. We assess the effectiveness of our approach on two different datasets, we compare various feature extraction techniques and performance metrics.

Artificial Intelligence and Deep Learning have significantly enhanced smart video surveillance systems. The effectiveness of these approaches relies on substantial processing power, large datasets, and advanced resources, which have become increasingly accessible due to powerful GPUs and high-RAM systems. Although convolutional neural networks (CNNs) excel at processing spatial information in images, they face limitations when analyzing temporal information in videos. Recurrent neural networks, such as Long Short-Term Memory (LSTM) networks, can address this challenge by modeling sequence information in video data, where each frame depends on its predecessors.

Various studies have explored anomaly detection [1-5], employing two main approaches based on the availability of labeled data. The traditional method, suited for situations where labeled data is unavailable, trains the model using known normal data. Alternatively, if labeled data is available, it can be used to train the model and predict abnormal classes for future test data. According to D. Elliott (2010), after 12 hours, a sin-

gle person can miss up to 80% of the activity between two cameras. This highlights the importance of effective anomaly detection systems. Deep learning outperforms other methods when the available dataset is large [6]. Anomaly detection or outlier detection is useful in various applications, including detecting illegal traffic flow [7], retinal damage [8], and IoT big-data anomaly detection [9]. However, deep learning-based methods often face difficulties in anomaly detection because of the complex structure of the data and the narrow boundary between normal and abnormal data.

In this paper, we make the following contributions:

1. We utilize multi-instance learning technique with auto label generation loss to tackle the challenge of anomaly detection in videos, particularly when video-level labels are available, but anomalies occur at the frame level.
2. We introduce a custom loss function for use in weakly supervised learning, designed to enable more effective extraction of discriminative features and thereby improve anomaly detection performance.
3. We incorporate a mean squared error function on auto-generated labels, which aids in separating interclass features and increasing intraclass feature closeness.
4. Our experiments, conducted without sparsity and temporal smoothness constraints, show that our proposed model is robust and effective. We evaluate our model on two benchmark datasets, UCF Crime and Shanghaitech, using various feature extraction techniques and comparing proposed loss functions in different environments.
5. We demonstrate that the I3D feature extractor outperforms the C3D feature extractor in our experiments, and we assess the model's performance using the area under the curve (AUC) metric.

The paper is organized as follows: Section 2 reviews related work and presents the problem statement, our proposed approach duscussed in Section 3, Section 4 discusses experimental results, and Section 5 concludes the paper.

2. RELATED WORK

Deep learning has proven to be a superior approach compared to traditional machine learning in several areas, particularly in image and video processing. Despite deep learning's superiority in various areas, detecting anomalies in image and video processing remains a challenging task, with many researchers making significant contributions to this field [10-14]. In [10], particle trajectories were utilized to model normal motion, and deviations from the norm were defined as anomalies. The author in [15] provides an in-depth analysis of deep anomaly detection in the medical domain. Researchers have also explored violence and aggression detection [16-18].

Feature learning is a conventional approach for inferring normality from data. However, due to difficulties associated with tracking objects in videos, many researchers have employed alternative methods such as motion pattern analysis using a histogram-based method [19], kernel density estimation methods [20], social force models [21], context-driven methods [22], and hidden Markov models [23]. These techniques offer different ways to address the difficulties of understanding motion and detecting deviations from normal patterns. During the testing phase, videos with lower probability are classified as anomalies, while normal videos are used for training. In [24], researchers focused on the problem of online detection of unusual events in videos using dynamic sparse coding. The main idea is that sparse representation can help us learn about normal behaviour in videos, which can then be used to detect unusual or abnormal events. Developing a video action classification model using deep learning has been proposed in [25]. However, video classification is more challenging than deep learning-based image classification due to the difficulty of obtaining annotations for training the model and the extensive efforts required to generate frame-level labels. To address the challenges posed by weakly labeled datasets, researchers have explored various approaches, as discussed in [26-29].

The RTFM(Robust Temporal Feature Magnitude learning) method [27] enhances detection by training a specialized function to recognize rare events and consider their timing, resulting in better accuracy and efficiency for detecting subtle anomalies. The MIST framework [28] focuses on using video-level annotations to refine important features, making the anomaly detection process more effective. Furthermore, the authors in [29] introduced the LAD database, a comprehensive collection of video sequences for anomaly detection, along with a multi-task deep neural network that leverages spatiotemporal features, achieving superior performance compared to existing methods in the field.

The author in [26] utilized a multi-instance learning (MIL) model to address the issue of weakly labeled datasets. Similar approaches have been employed for detecting anomalies, as discussed in [30-33]. The author in [31] proposes the Anomaly Regression Net (ARNet) framework for video anomaly detection, which only requires video-level labels in training and uses multiple-instance learning loss and centre loss for discriminative features. [32] proposes a weakly supervised deep temporal encoding-decoding solution using multiple instance learning for anomaly detection in surveillance videos and employs a new smoother loss function. [33] focuses on reducing false alarms in abnormal activity detection using 3D ResNet and deep multiple instance learning with a new ranking loss function, achieving the best performance on the UCF-Crime benchmark dataset. All three papers present novel approaches for video anomaly detection and achieve advanced results on challenging benchmark datasets.

Detecting anomalies with accuracy is a challenging task, primarily due to its subjective nature, which varies based on location and individual perspectives. Researchers have approached anomaly detection as a means of identifying low-probability patterns, as evident in studies conducted by [34-36]. In this research, we address the problem of anomaly detection as a regression issue and propose a customized loss function, coupled with multi-instance learning techniques.

Our proposed loss function aims to increase the gap between the normal and abnormal frames while minimizing computational complexity. This is achieved by removing the sparsity and temporal smoothness constraints typically present in similar techniques. The proposed methodology section will detail our approach to addressing the challenge of detecting anomalies with high accuracy.

2.1. PROBLEM STATEMENT

Our research tackles the challenge of frame-level anomaly detection in videos using the UCF Crime and Shanghai tech datasets. These datasets provide anomaly labels at the video level, complicating frame-level detection. To address this, we employ multi-instance learning (MIL) and split the dataset into two parts: one with normal frames and another with a mixture of normal and abnormal frames grouped under a single anomaly class. Our aim is to effectively detect anomalies at the frame level by utilizing MIL and a custom loss function that minimizes false anomaly detections. We will detail our techniques, their application to the datasets, and our experimental results in the subsequent sections. By enhancing frame-level anomaly detection, our research contributes to the field of video surveillance and has potential applications in security systems and public safety measures.

3. PROPOSED METHOD

This section of the paper aims to define the problem of anomaly detection in video, describe the feature extraction method, and provide a detailed description of the proposed loss function. To detect anomalies in the video, we utilize the UCF Crime and Shanghaitech datasets, which contain a range of videos of different lengths categorized as normal, explosive, burglary, fighting, and arrest. Similar to [26], anomaly detection is treated as a regression problem, where a sequence of frames serves as the input and an anomaly score between 0 and 1 is the output for each frame.

In this work, we present a deep learning method-based approach for detecting anomalies. We begin by converting the input video into a fixed-size array and then extract features using both three-dimensional convolutional features [37] and inflated three-dimensional (I3D) features [38]. Each video is then segmented into a fixed number of non-overlapping temporal segments, and each segment is treated as a "bag" instance for feature extraction. We extract 3D convolutional and I3D features from each video segment.

We utilized two types of pre-processed video data, namely C3D and I3D features, to extract features for our model. These models were chosen due to their efficiency in learning spatiotemporal features, which are crucial for further processing. C3D features consist of two-stream pre-processed video data with a feature dimension of 4096. On the other hand, I3D features are composed of RGB and optical-flow features, with a feature dimension of 2048 for each. During the training process, we concatenated the RGB and optical flow features to create a unified input. To visualize our proposed approach, we have included a diagram of the model with the custom loss function in Fig. 1.

The loss function of the support vector machine model is

$$L(w) = \frac{1}{k}\sum_{i=1}^{k} \max\left(0, 1 - y_i \mid w^T x_i + b\right) + \lambda ||w||_2^2 \quad (1)$$

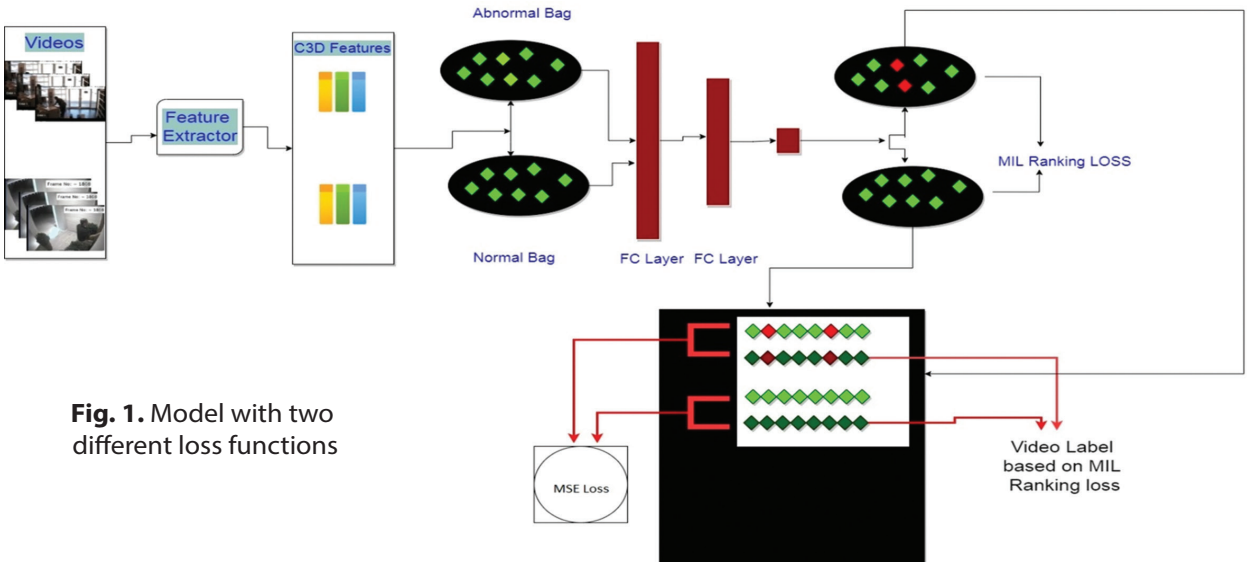


Fig. 1. Model with two different loss functions

where w is the weight vector, b is the bias term, and T denotes the transpose of the weight vector. The loss is accumulated over all training data points and is often combined with a regularization term to prevent overfitting. The loss function used in the model has two components: the hinge loss and the regularization term. During training, the learning parameter w is adjusted to minimize the hinge loss, which generates a positive loss for incorrectly classified features. In this supervised learning approach, the labels *Yi* and features *xi* are used along with the bias b to determine the loss. However, as the video frames lack annotation, this approach is not applicable. To address this issue, the MIL approach was adopted, as discussed in [26]. Under this approach, each video is divided into a bag, with the positive bag containing both normal and abnormal frames and the negative bag containing only normal frames. Similar to [26, 39], the maximum ($wTXi + b$) is considered for both types of bags. This approach allows the model to learn to identify abnormal frames without the need for individual frame annotations, which will modify equation 1 to

$$L(w) = \frac{1}{k}\sum_{i=1}^{k} \max\big(0,1 - \mathrm{y_i}\max_{\mathrm{i\ from\ bag}} (|\ \mathrm{w^T x_i} + \mathrm{b}|)\big) + \lambda||\mathrm{w}||_2^2 \quad (2)$$

Here, *Yi* is the bag-level label. where y is the true label of the bag (+1 for positive bags and -1 for negative bags), $f(x)$ is the decision function of the SVM, B is the set of instances in the bag, and $max(_{i\ from\ bag})$ $(|w^T x_i + b|)$ is the maximum predicted score for any instance in the bag.

Our proposed loss function aims to maximize the distance between positive and negative bags, with only the maximum distance feature considered for each bag. The selection of the maximum feature is based on the assumption that each abnormal bag should contain at least one abnormal instance, while a normal bag should only contain normal instances. Building on this approach, we developed a custom loss function that combines multi-instance learning with the residual difference between actual and predicted labels to train the network. In this case, the actual label is determined through maximum selection in the MIL process. This can be explained by equation 3:

$$L_{MIL} = \frac{1}{kN}\sum_{kN} (^{Max}_{i\epsilon Ni} (yi * (w^t xi + b))) + \gamma - \frac{1}{KAN}\sum_{kAN} (^{Max}_{i\epsilon Abi} (yi * (w^t xi + b)) \quad (3)$$

Here, the first term represents the average sum of the maximum distanced feature from each normal video, and the second term represents the average sum of the maximum distanced feature from each abnormal video, and the third term represents the hyperparameter. The objective of this loss function is to maximize the difference between the abnormal and normal features, as represented by the first two terms of the equation. To further refine this approach, we introduce a custom loss function that combines multi-instance learning with the residual difference between the actual and predicted labels.

The following equation explains how pseudo-labels are generated for each instance:

$$L_{MSE} = \frac{1}{k}\sum_{k} (Y_{\mathrm{AUTO}} - Y_{MIL})^2 \quad (4)$$

Here, Y_{AUTO} is a label generated based on the distance of the feature from the line measured by L_{MIL}. Y_{AUTO} assigns a label of 1 (abnormal) to an instance if the maximum absolute value of the weighted sum for all instances in the bag is greater than a certain threshold. Otherwise, it assigns a label of 0 (normal). This method helps identify the most representative instances within each bag, which, in turn, assists in the training of the network to maximize the difference between normal and abnormal features. Y_{AUTO} is calculated as follows:

$$Y_{\mathrm{AUTO}} = 1 \text{ if } \max_{\mathrm{i\ from\ bag}}(\ |\mathrm{w^T x_i} + \mathrm{b}|\ ) \text{ else } 0 \quad (5)$$

Y_{MIL} is the actual distance calculated for each feature in the bag. The final loss function is the sum of equations (3) and (4).

$$L_{MIL\_MSE} = L_{MIL} + L_{MSE} \quad (6)$$

Combining multi-instance learning with the residual difference The custom loss function incorporates both the multi-instance learning component (L_{MIL}) and the residual difference between actual labels and pseudo-labels (L_{MSE}). This combination allows the model to better learn the relationship between the features and the labels, resulting in improved anomaly detection.

Using Y_{AUTO} as a pseudo-label helps the model learn better decision boundaries by leveraging the information from the most representative instances. This aids in training the model to effectively distinguish between normal and abnormal instances, improving its overall anomaly detection capability.

4. EXPERIMENTAL RESULTS

This section describes the use of C3D and I3D as feature extractors for video anomaly detection on the UCF Crime dataset and the ShanghaiTech dataset. C3D is a neural network that extracts spatiotemporal features from videos, while I3D is a modified version of C3D that achieves advance results in video recognition tasks. The proposed approach extracts features using pre-trained C3D and I3D networks and uses a one-class SVM classifier with a custom loss for anomaly detection. The one-class SVM classifier is a popular choice for anomaly detection as it is designed to distinguish between normal and abnormal instances. The experimental results show that I3D outperforms C3D in all evaluation metrics, and the system's performance improves with an increase in the number of frames used in feature extraction. The proposed approach achieves competitive results compared to state-of-the-art methods on the UCF Crime dataset and the ShanghaiTech dataset.

The UCF Crime dataset and the ShanghaTech dataset are both challenging and widely used benchmark datasets for video anomaly detection. The UCF Crime dataset consists of 1,900 real-world surveillance videos that encompass various crime types, such as theft, robbery, vandalism, and fights. This diverse dataset poses a challenge for models to accurately detect and classify different types of anomalous behaviours in realistic settings. On the other hand, the ShanghaTech dataset contains 437 high-resolution surveillance videos from diverse environments like streets, parks, and commercial areas, featuring anomalies such as jaywalking, loitering, and illegal parking. Its difficulty arises from the high variability in video content, camera angles, and lighting conditions, making it a robust dataset for evaluating video anomaly detection model performance across different scenarios.

4.1. C3D NETWORK

This section introduces a video anomaly detection approach utilizing C3D features extracted from the UCF Crime dataset, as outlined in [26]. The C3D features capture both the appearance and dynamics of moving objects for video action recognition. Each video is segmented into non-overlapping fixed-size segments to create a 4096x32 feature matrix. A neural network having four fully interconnected layers with 256, 64, and 16 neurons and a single output neurons is employed, using an Adagrad optimizer and a learning rate of 0.01. The performance is assessed by the area under the receiver operating characteristic (AUC-ROC) curve, enabling fair comparisons. This approach computes the ROC curve based on the frame-level anomaly score.

4.2. I3D NETWORK

This experiment adopts the Inflated 3D (I3D) model, pre-trained on the Kinetics dataset, as the feature extraction network. The I3D network output for each video includes RGB and optical flow features, which are concatenated, producing a 2048x32 feature output size. A four-layer fully connected neural network with 128, 32, and 16 units and a single output layer is used. Training is conducted with the Adagrad optimizer and a 0.01 learning rate. Tables 1 and 2 display the results of our custom loss function.

Table 1 highlights the effectiveness of incorporating I3D features into the model for video anomaly detection. The I3D features-based approach achieves an AUC score of 84.66, surpassing other methods in the comparison, thus demonstrating its superiority. Tests were also conducted using C3D features and I3D with only RGB features. Table 2 summarizes the corresponding AUC, F1, and EER scores, providing insights into the performance of different feature sets in video anomaly detection and emphasizing the advantages of I3D features.

Our experiments, conducted using the open-source code by Sultani et al. [26], are based on established research and methods. A confusion matrix in Table 3 adds context and understanding to our findings, detailing the rates of true and false predictions, enabling readers to evaluate the model's effectiveness in detecting video anomalies comprehensively.

Overall, our results in Tables 1, 2, and 3 strongly support I3D features for video anomaly detection. The high AUC score, F1 score, and EER emphasize the effectiveness of our approach compared to others. Incorporating I3D features yields the best performance, as indicated by the highest AUC score. These findings have important implications for future research. Fig. 2 and 3 display results for various test dataset videos. Fig. 3 illustrates the anomaly score graph for abnormal frames when the model generates higher scores compared to normal frames. This figure presents the results for two specific video instances: a) Stealing079_x264 and b) Stealing047_x264. The visual representation in Fig. 3 provides insights into the model's ability to accurately detect and distinguish abnormal behavior, such as theft, from regular activities, further showcasing the effectiveness of the model in video anomaly detection tasks.

Table 1. AUC Score of Comparison on UCF Crime dataset with Various methods

Method	Features	AUC (%)
Hasan et al [35]	C3D RGB	50.6
Lu et al [40]	C3D RGB	65.51
Sultani et al. [26]	C3D RGB	75.41
MIST [28]	C3D RGB	81.40
	I3D RGB	82.30
Zhang et al.[41]	C3D RGB	78.66
J. Zhong et al [42]	C3D	81.08
	TSNRGB	82.12
	TSNOptical Flow	78.08
Proposed	C3D RGB	76.004
	I3D RGB	81.55
	I3D RGB + Optical Flow	84.66

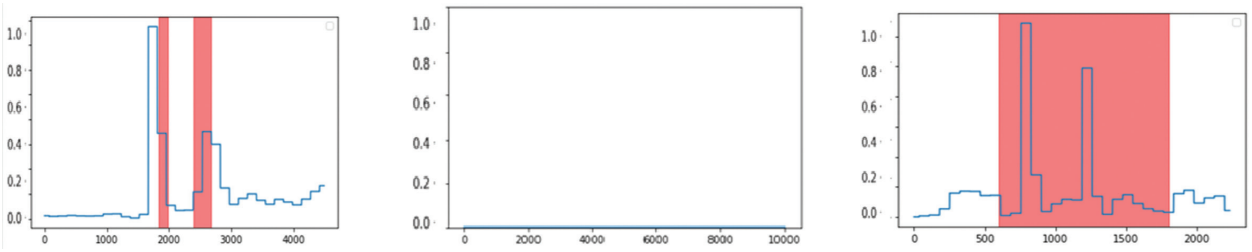
Table 2. AUC Score of Comparison on UCF Crime dataset : C3D vs I3D

Features	AUC (%)	F1 Score	EER
C3D	76.00	25.61	30.75
I3D RGB + Optical Flow	84.66	35.63	22.59

Table 3. Confusion Matrix

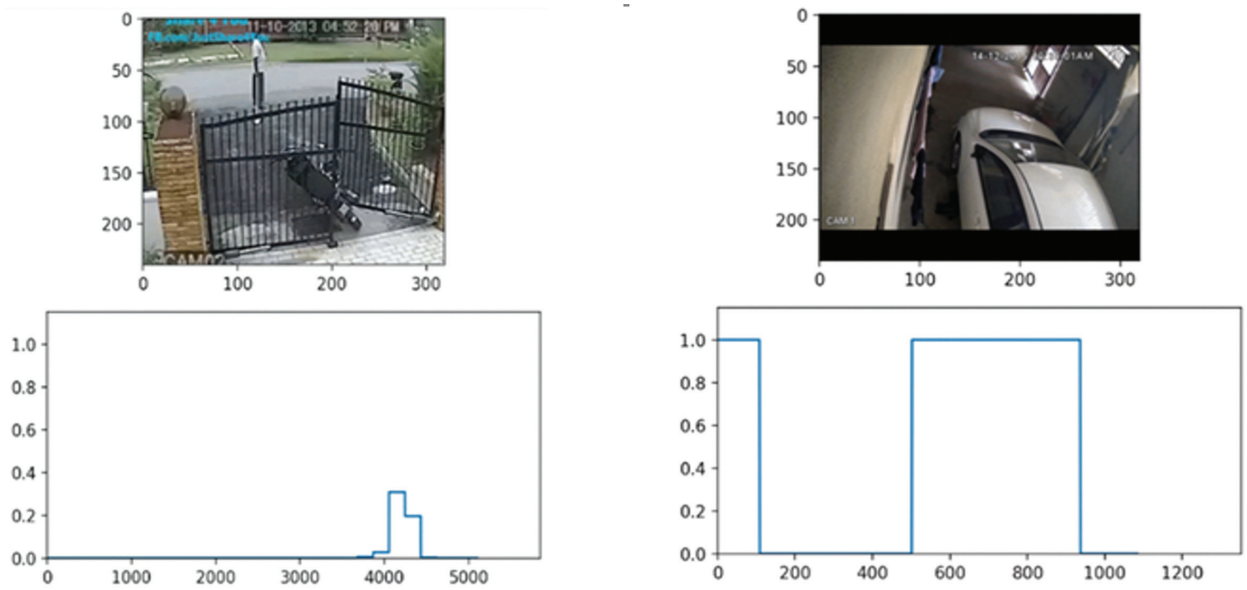
	Predicted: Normal	Predicted: Abnormal
Actual: Normal	82.85 % (851346)	17.14 % (176131)
Actual: Abnormal	33.03 % (27860)	66.96 % (56471)

The model accurately predicts vandalism and stealing anomalies but doesn't generate an alert for normal videos. Due to varying conditions and challenges, the model isn't flawless at anomaly detection. In Fig. 4, a) the model occasionally fails to generate alerts when anomalies are present, and b) the model generates false alerts in the absence of visible anomalies. Generally, our network produces higher scores for abnormal video segments.



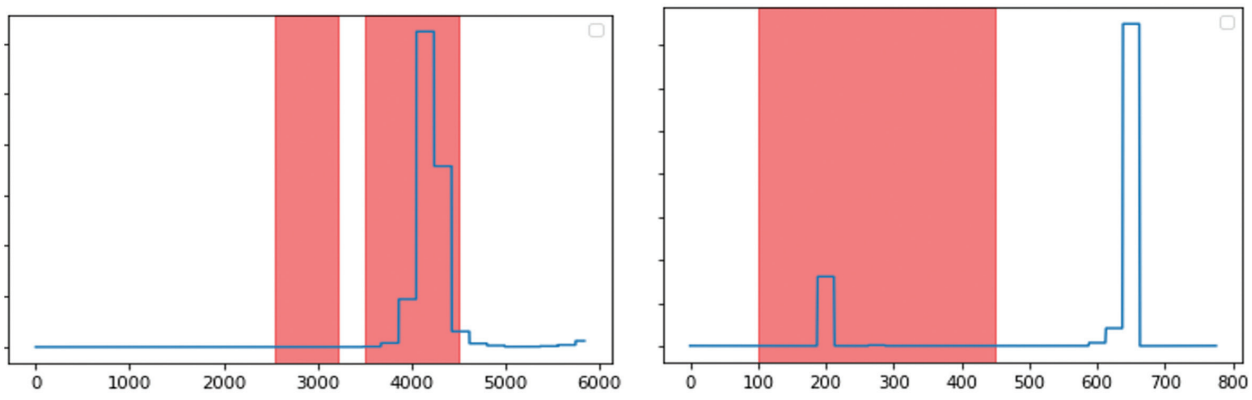
(a) (b) (c)

Fig. 2. Visualization of testing Results for a). Vandalism028, b). Normal877, c). Stealing059 Red portion shows actual abnormal frames Blue line shows anomaly score for given frames



(a) (b)

Fig. 3. Results of anomaly score graph for abnormal frames when model generates higher score compare to normal frames a) Stealing079_x264 b) Stealing047_x264



(a) (b)

Fig. 4. Wrong prediction Results a) Stealing079_x264 here Model is not predicting anomaly for first window b) Explosion027_x264 Model is generating wrong alarm even anomaly is not there in video

The ShanghaTech dataset, a challenging benchmark dataset, is employed to further evaluate our model's performance. In this case, we consolidate all normal videos into a single normal class and all abnormal videos into a single abnormal class. Following the dataset split suggested in [42], we facilitate binary categorization. The dataset consists of 238 training videos and 199 test videos. Table 4 presents the AUC score results using I3D features for the ShanghaTech dataset. Our proposed model outperforms the other methods listed in Table 4, achieving an impressive AUC score of 92.27. This performance on the ShanghaTech dataset, known for its difficulty, further validates the effectiveness of our model in anomaly detection tasks.

Table 4. AUC Score of Quantitative Comparison on Shanghai Tech dataset

Method	Features	AUC(%)
Zhong et al.	TSNRGB	84.44
Zhong et al.	TSNOptical-Flow	84.13
Sultani et al [26]	C3D	86.30
AR-NET	I3D conc (RGB + Optical Flow)	91.24
Proposed (I3D RGB + Optical Flow)	I3D conc(RGB + Optical Flow)	**92.27**

5. CONCLUSION

This article presents a novel approach for detecting anomalies in videos using multi-instance learning and a custom dynamic loss function called LMIL_MSE. The loss function is calculated using the mean square error and is influenced by how well interclass features are separated. Pseudo-label generation is also used to improve the quality of the training data. Our experiments on the UCF Crime and Shanghai Tech datasets show that our approach outperforms previous methods in detecting video anomalies. To extract features from the video data, we used both C3D and I3D networks. We found that the I3D features yielded the highest AUC score in our experiments. Additionally, we utilized a multi-instance learning approach to improve the detection of anomalies in the video data.

Our research offers a promising solution to the challenging problem of detecting video anomalies. By achieving better results than previous methods, our approach has the potential to enhance the accuracy and reliability of video anomaly detection systems in real-world applications.

6. REFERENCES:

[1]. C. C. Aggarwal, "An introduction to outlier analysis", Outlier analysis, Springer Cham., 2017, pp. 1-34.

[2]. A. Deng, B. Hooi. "Graph neural network-based anomaly detection in multivariate time series", Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, 2-9 February 2021, pp. 4027-4035.

[3]. A. Boukerche, L. Zheng, O. Alfandi, "Outlier detection: Methods, models, and classification", ACM Computing Surveys, Vol. 53, No. 3, 2020, pp. 1-37.

[4]. A. Zimek, E. Schubert, H.-P. Kriegel. "A survey on unsupervised outlier detection in high-dimensional numerical data", Statistical Analysis and Data Mining, Vol. 5, No. 5, 2012, pp. 363-387.

[5]. R. Chalapathy, S. Chawla, "Deep learning for anomaly detection: A survey", arXiv:1901.03407, 2019.

[6]. A. Bahnsen, "Building AI Applications Using Deep Learning", https://albahnsen.com/2017/06/06/building-ai-applications-using-deep-learning/ (accessed: 2023)

[7]. X. Xie, C. Wang, S. Chen, G. Shi, Z. Zhao, "Real-time illegal parking detection system based on deep learning", Proceedings of the International Conference on Deep Learning Technologies, 2017, pp. 23-27.

[8]. T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery", Proceedings of Information Processing in Medical Imaging: 25th International Conference, Boone, NC, USA, 25-30 June 2017, pp. 146-157.

[9]. M. Mohammadi, A. Al-Fuqaha, S. Sorour, M. Guizani, "Deep Learning for IoT Big Data and Streaming Analytics: A Survey", IEEE Communications Surveys & Tutorials, Vol. 20, No. 4, 2018, pp. 2923-2960.

[10]. S. Wu, B. E. Moore, M. Shah, "Chaotic invariants of Lagrangian particle trajectories for anomaly detection in crowded scenes", Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 2010, pp. 2054-2060.

[11]. D. Xu, E. Ricci, Y. Yan, J. Song, N. Sebe. "Learning deep representations of appearance and motion for anomalous event detection", arXiv:1510.01553, 2015.

[12]. B. Antić, B. Ommer, "Video parsing for abnormality detection", Proceedings of the International Conference on Computer Vision, Barcelona, Spain, 6-13 November 2011, pp. 2415-2422.

[13]. V. Singh, S. Singh, P. Gupta, "Real-Time Anomaly Recognition Through CCTV Using Neural Networks", Procedia Computer Science, Vol. 173, 2020, pp. 254-263.

[14]. Md Sharif, L. Jiao, C. W. Omlin, "Deep Crowd Anomaly Detection: State-of-the-Art, Challenges, and Future Research Directions", arXiv:2210.13927, 2022.

[15]. G. Litjens et al. "A survey on deep learning in medical image analysis", Medical Image Analysis, Vol. 42, 2017, pp. 60-88.

[16]. S. A. A. Akash, R. S. S. Moorthy, K. Esha, N. Nathiya, "Human Violence Detection Using Deep Learning Techniques", Journal of Physics: Conference Series, Vol. 2318, No. 1, 2022, p. 012003.

[17]. H. Gupta, S. T. Ali, "Violence Detection using Deep Learning Techniques", Proceedings of the International Conference on Emerging Techniques in Computational Intelligence, Hyderabad, India, 2022, pp. 121-124.

[18]. S. Sadiq, A. Mehmood, S. Ullah, M. Ahmad, G. S. Choi, B.-W. On, "Aggression detection through deep neural model on Twitter", Future Generation Computer Systems, Vol. 114, 2021, pp. 120-129.

[19]. X. Cui, Q. Liu, M. Gao, D. N. Metaxas, "Abnormal detection using interaction energy potentials", Proceedings of the CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011, pp. 3161-3167.

[20]. G.-h. Ji, X.-h. Zhang, X.-y. Cheng, "Pedestrians Detection Based on the Integration of Human Features and Kernel Density Estimation", IOP Conference Series: Materials Science and Engineering, Vol. 490, No. 4, 2019, p. 042027.

[21]. X. Yang et al. "Deep social force network for anomaly event detection", IET Image Processing, Vol. 15, No. 14, 2021, pp. 3441-3453.

[22]. Y. Zhu, N. M. Nayak, A. K. Roy-Chowdhury, "Context-aware activity recognition and anomaly detection in video", IEEE Journal of Selected Topics in Signal Processing, Vol. 7, No. 1, 2012, pp. 91-101.

[23]. L. Kratz, K. NishiNo, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20-25 June2009, pp. 1446-1453.

[24]. B. Zhao, L. Fei-Fei, E. P. Xing, "Online detection of unusual events in videos via dynamic sparse coding", Proceedings of the CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011, pp. 3313-3320.

[25]. A. Karpathy, G. Toderici, S. Shetty, Thomas Leung, R. Sukthankar, L. Fei-Fei, "Large-scale video classification with convolutional neural networks", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23-28 June 2014, pp. 1725-1732.

[26]. W. Sultani, C. Chen, M. Shah, "Real-world anomaly detection in surveillance videos", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18-23 June 2018, pp. 6479-6488.

[27]. Y. Tian, G. Pang, Y. Chen, R. Singh, J. W. Verjans, G. Carneiro, "Weakly-supervised video anomaly detection with robust temporal feature magnitude learning", Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10-17 October 2021, pp. 4975-4986.

[28]. J.-C. Feng, F.-T. Hong, W.-S. Zheng, "MIST: Multiple instance self-training framework for video anomaly detection", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20-25 June 2021, pp. 14009-14018.

[29]. B. Wan, W. Jiang, Y. Fang, Z. Luo, G. Ding, "Anomaly detection in video sequences: A benchmark and computational model", IET Image Processing, Vol. 15, No. 14, 2021, pp. 3454-3465.

[30]. B. Wan, Y. Fang, X. Xia, J. Mei, "Weakly supervised video anomaly detection via center-guided discriminative learning", Proceedings of the IEEE International Conference on Multimedia and Expo, London, UK, 6-10 July 2020, pp. 1-6.

[31]. A. M. Kamoona, A. K. Gostar, A. Bab-Hadiashar, R. Hoseinnezhad. "Multiple instance-based video anomaly detection using deep temporal encoding–decoding", Expert Systems with Applications, Vol. 214, 2023, p.119079.

[32]. S. Dubey, A. Boragule, M. Jeon, "3d resnet with ranking loss function for abnormal activity detection in videos", Proceedings of the International Conference on Control, Automation and Information Sciences, Chengdu, China, 23-26 October 2019, pp. 1-6.

[33]. S. Dubey, A. Boragule, J. Gwak, M. Jeon, "Anomalous event recognition in videos based on joint learning of motion and appearance with multiple ranking measures", Applied Sciences, Vol. 11, No. 3, 2021, p.1344.

[34]. T. Pevný, "Loda: Lightweight on-line detector of anomalies", Machine Learning, Vol. 102, 2016, pp. 275-304.

[35]. M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, L. S. Davis, "Learning temporal regularity in video sequences", Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, 27-30 June 2016, pp. 733-742.

[36]. G. Stein, U. Seljak, B. Dai, "Unsupervised in-distribution anomaly detection of new physics through conditional density estimation", arXiv:2012.11638, 2020.

[37]. D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, "Learning spatiotemporal features with 3d convolutional networks", Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7-13 December 2015, pp. 4489-4497.

[38]. J. Carreira, A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset", IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21-26 July 2017, pp. 6299-6308.

[39]. S. Andrews, I. Tsochantaridis, T. Hofmann, "Support vector machines for multiple-instance learning", MIT Press, 2002, pp. 577-584.

[40]. C. Lu, J. Shi, J. Jia, "Abnormal event detection at 150 fps in matlab", Proceedings of the IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 1-8 December 2013, pp. 2720-2727.

[41]. J. Zhang, L. Qing, J. Miao, "Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection", Proceedings of the IEEE International Conference on Image Processing, Taipei, Taiwan, 22-25 September 2019, pp. 4030-4034.

[42]. J. Zhong et al. "Graph Convolutional Label Noise Cleaner: Train a Plug-And-Play Action Classifier for Anomaly Detection", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 2019 pp. 1237-1246.