International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Enhanced Video Anomaly Detection Using Weakly-Supervised LSTM Framework and Comparative Analysis of I3D and ViT Feature Extraction Techniques

Preet Kanwal[1*], Shylaja S S[2], Prasad B Honnavalli[3]

Submitted:13/03/2024 **Revised**: 28/04/2024 **Accepted**: 05/05/2024

Abstract: Anomaly detection in video surveillance is a crucial yet challenging task, especially when anomalies exhibit a small difference compared to their normal counterparts. The nature of this problem becomes even more complex when relying on weakly-supervised approaches with video-level labels. In this study, we leverage a weakly supervised framework, treating each video as a sequence of instances and propose a novel method to detect anomalies that utilizes LSTM-based models to effectively capture temporal-dependencies. Furthermore, to assess the effectiveness of the model in detecting anomalies, we compared the performance of two feature extraction techniques - Inflated 3D ConvNet (I3D) and Vision Transformer (ViT). Extensive experiments conducted on RTX 4090 GPU and large-scale benchmark dataset - UCF-Crime demonstrate that our model achieves better anomaly detection performance (AUC : 90% with I3D and 86% with ViT) compared to existing state-of-the-art methods. The comparative analysis of the I3D and ViT feature extraction methods provide insights into their applicability to different types of video anomalies.

Keywords: *LSTM, Multiple-Instance Learning, Video Anomaly Detection, Weakly-Supervised*

1. Introduction

WITH extensive applications in autonomous surveillance systems [14, 31, 40, 48] , where the timely and precise identification of abnormal events is crucial for public safety and security, Video Anomaly Detection(VAD) has become a crucial field of study [4, 26]. Unusual occurrences such as shoplifting, accident and violence present significant hazards, and their prompt identification and timely detection can mitigate potential harm [12]. However, the very nature of the data poses inherent difficulty in defining what constitutes an "anomaly" against a backdrop of typical normal activities [1, 14, 17, 47].

The weakly-supervised approach to Video Anomaly detection which uses video level labels [31, 40, 45, 48, 49], has gained traction as compared to a fully supervised learning approach that requires frame-level labels and unsupervised methods such as one-class classifiers (OCCs), that uses only normal videos for training [14, 15, 19, 23, 28, 29, 46]. Supervised approaches require extensive manual annotation of video frames and OCCs struggle with generalization to unseen anomalies and suffer from high false alarm rates.

Weakly-supervised VAD is a promising approach that strikes a balance between the effort required to annotate and the detection accuracy. However, it still faces significant challenges owing to the very nature of the problem : a) The majority of the content in the long video is normal and only a small portion may contain an anomaly. Hence the dataset was highly imbalanced. b) Anomalies that exhibit only slight deviations from normal events pose a significant difficulty. c) The definition of an anomaly depends on the context, for example : playing football on the ground is normal but doing the same in the classroom is an anomaly.

To address the aforementioned challenges, recent studies have explored the use of Multiple Instance Learning (MIL) frameworks where each video is treated as a bag of snippets and the model is trained to identify abnormal snippets. However, existing MIL-based approaches can select non-representative abnormal snippets and lack robust mechanisms to capture temporal dependencies in video data [9, 19, 23, 32, 39, 40, 48, 49].

To address these limitations, we propose a novel approach that leverages Long Short-Term Memory (LSTM) networks to better model the temporal dependencies and detect the anomalous events effectively. We also compare the performance of two advanced feature extraction techniques-Inflated 3D ConvNet (I3D) and Vision Transformer (ViT)-in detecting the anomalies. We aim to improve the ability of the model to detect the anomalies by integrating the LSTM network with these feature extraction methods. For long surveillance videos, the proposed approach enhances the robustness of the system.

[1] *Associate Professor, Dept. of CSE, PES University, Bengaluru – 560085, India*
ORCID ID : 0000-0002-7490-0090
Email : preetkanwal@pes.edu* (Corresponding author)
[2] *Professor, Dept. of CSE, PES University, Bengaluru – 560085, India*
ORCID ID : 0000-0003-2628-8973
Email : shylaja.sharath@pes.edu
[3] *Professor, Dept. of CSE, PES University, Bengaluru – 560085, India*
ORCID ID : 0009-0008-5650-1804
Email : prasadhb@pes.edu

We also throw light on the data cases, where both I3D-based-LSTM network and ViT-based-LSTM network fail to classify correctly.

The remainder of this paper is organized as follows : Section 2 reviews related work with greater focus on Weakly-Supervised VAD. Section 3 elaborates on the proposed methodology and the model architecture. Section 4 discusses the and results on UCF-Crime dataset and a comparison to other state-of-the-art approaches. Finally, we conclude the paper in Section 5 and discuss potential research directions for future work.

2. Related Work

Video Anomaly Detection (VAD) has garnered significant research interest owing to its critical applications in surveillance systems, transportation, public safety and security. In real-world scenarios, due to the unpredictability and rarity of anomalous events, the problem becomes inherently challenging leading to significant imbalance between normal and anomalous events. Traditional methods to VAD typically rely on fully supervised approaches that require huge effort in terms of manual annotation (frame-level labels) or unsupervised methods such as one-class classifiers (OCCs) that learn to detect anomalies only from normal data and suffer from high false alarms [10, 30]. Both methods suffer from poor generalization.

In contrast, weakly-supervised video anomaly detection (WSVAD) has gained attention in recent years where only video-level labels are provided. The predominant approach within WSVAD to detect anomalies leverages Multiple Instance Learning (MIL) frameworks. [31] was the first to model the WSVAD problem using a Multiple-Instance Learning (MIL) framework, aiming to train a regression model where each video is treated as a bag of instances (frames/segments/clips). A bag is labelled positive if it contains at least one abnormal snippet. For each video segment/instance anomaly score is predicted by a deep neural network-based ranking model. The ranking loss encourages high scores for segments that contain anomalies in the video. The ranking is enforced only on segments that have the highest score in the positive and negative bag instead of ranking every instance in the bag. The model is trained to maximize the distance between the scores of anomalous and normal segments, ensuring they are as distinct as possible. The features are extracted using a pre-trained model - Convolutional 3D (C3D) [33] and a 2-layer Fully connected network (FCNN).

For any classification problem, the popular choice for the loss function is a categorical/binary cross entropy or hinge loss etc. [31] introduced MIL ranking loss which is made up of hinge loss and incorporates sparsity and smoothness constraints. Sparsity constraint states that the anomaly occurs only for a short period, therefore scores of the video that contains anomaly will be sparse. Temporal smoothness states that the anomaly scores vary smoothly from one segment to another, which means we would not see a drastic difference in the scores of adjacent segments. Therefore in the case of [31], the loss function is defined as :

$$L(W) = \max(0, 1 - \max_{i \in B_a} f(V_i^a) + \max_{i \in B_n} f(V_i^n)) \quad (1)$$

$$+ \lambda_1 \sum_{i=1}^{n-1} (f(V_i^a) - f(V_{i+1}^a))^2 + \lambda_2 \sum_{i=1}^{n} f(V_i^a) \quad (2)$$

$$+ \lambda_3 \|W\|_F \quad (3)$$

where B_a and B_n represent the positive (anomalous) and negative (normal) bags, respectively, $f(V_i^a)$ is the anomaly score for the *i-th* instance in the positive bag, W are the model weights and λ_1, λ_2 and λ_3 are the regularization parameters for smoothness and sparsity constraints, respectively.

MIL-based anomaly detection methods easily get affected by label noise and have a high false alarm rate. For example, a normal segment may get a top anomaly score mistakenly in an anomalous bag. Since 2018, after [31] WSVAD has become the prevalent area of research [9, 20, 21, 32, 45, 48] and each of these researches focused on improving the MIL ranking framework in terms of feature extraction [11, 25, 27, 35, 39, 49], loss function [32, 38, 45], label denoising [43, 48] or model refinement [9, 18, 42].

Following [31], many researchers [38, 45] proposed score distance approaches and exhibited better performance. However, these methods face several challenges such as : a) fail to leverage abnormal video labels and hence perform poorly when the video-level labels are noisy b) face difficulty in capturing complex temporal dependencies in video data. MIL approaches face limitations in learning when using a few or single significant snippets and relying solely on regression output, rather than making decisions which are feature based. To mitigate these challenges, recent works incorporate more sophisticated frameworks that utilize temporal modeling and context-aware learning.

In order to reduce the label noise, [48] restructured the problem of MIL-based WSVAD as a binary classification task and applied a GCN to improve noisy prediction derived from labels at the video level. Although [48] achieved better results than [31], training GCN and MIL can produce unconstrained latent space where normal and abnormal features might exist across the feature space, leading to unstable performance and increased computational costs.

To remedy the issue of weak labels [34], employed a

oneshot Siamese network for anomaly detection. [37] improved the detection capabilities by using a lightweight CNN with a residual attention LSTM framework, which efficiently captured spatial and temporal features in the video. [40], released a multi-modal violence dataset (XD-Violence) and proposed HL-Net (Holistic and localized method) to capture long-range dependencies and short-range interactions using GCN. [39] models abnormal concepts using dual-branch network with an aim to localize the spatio-temporal tube that encompasses the abnormal event. [36] extended their previous work by incorporating both appearance and motion features as a two stream mechanism, processed through a bidirectional LSTM network, eventually enhancing the model's accuracy in detecting and recognizing anomalies. [9] proposed MIST, where they fine-tuned a feature encoder based on generated pseudo-labels.

While [9, 31, 33] did pay some attention to temporal dependencies to consider multi-scale temporal information, [49] improved the MIL-framework by using a temporal enhanced network to obtain motion-sensitive features and integrated temporal context into MIL ranking model via attention blocks. However, for complex scenes, excluding appearance information and focusing only on temporal or motion information leads to an incomplete understanding. [44] focused on multi-scale temporal dependencies. [32] proposed robust model (RTFM) with an objective to enlarge feature magnitude value between normal and abnormal class and select topk segments to determine abnormal video segment. Current research uses the MTN network for feature aggregation introduced by [32]. [18] learned the temporal relations between sequences of multiple video instances and proposed an innovative multi-sequence learning method leveraging the transformer model. [6] introduced MGFN that uses Magnitude Contrastive loss to better differentiate between normal and abnormal features.

While several methods acknowledge the temporal correlation between video segments, they fail to consider temporal differences between normal and abnormal videos. These approaches overlook the importance of contextual relationships and motion, which is crucial in real-world scenarios. It is challenging to amplify the gap of features when dealing with weak-labels through a single-branched backbone.

2.1. Feature Extractors

For a model to effectively detect anomalies, the choice of feature extraction techniques play a critical role. Traditional approaches relied on handcrafted features. However, as deep learning has made substantial progress, it is now common to use deep neural network based feature extraction. Majorly there are two kinds of models that have been employed for this task : a) 3D-CNN based, b) Transformer based. WSVAD literature has predominantly utilized I3D to extract video features effectively from raw video frames. I3D models inflate 2D convolutional kernels to 3D and enable the model to process spatiotemporal data and learn from both motion and appearance cues in the video [3, 31]. On the other hand, Vision Transformers (ViT) have been adapted for computer vision tasks, inspired by their success in natural language processing. ViTs model long-range dependencies through self attention mechanisms and capture global contextual information, potentially offering superior performance in detecting subtle anomalies [2, 7, 8, 13, 41].

While I3D models excel in capturing detailed spatiotemporal features, they may struggle with processing long sequences due to

computational constraints[3]. In contrast, ViTs offer a more flexible and scalable approach, capable of handling long-range dependencies and complex contextual information, though they may require larger datasets and more computational power to achieve optimal results [8, 41].

Although several models are available, there exists a huge gap in terms of accuracy. There is a need to effectively and efficiently recognize anomalies. The comparative analysis of I3D and ViT presented in this paper adds to this growing body of knowledge and provides deeper insights into the applicability of these methods for different types of video anomalies. In this paper, we further explore these techniques by integrating them with LSTM networks to enhance the temporal modeling capabilities of the anomaly detection system.

3. Proposed Methodology

The architecture of our framework is shown in Fig. 1. The proposed pipeline extracts features from untrimmed video. The processed feature vectors along with the video level labels are passed as input to the deep learning model to perform a binary classification (anomaly -1, normal - 0). The various architectural modules are explained in the further subsections.

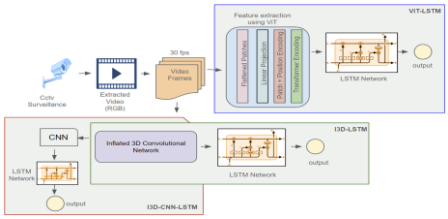

Fig. 1. High level Architecture Diagram. Two feature extractors have been explored - I3D and ViT. The feature vectors are then processed and results of the binary classification are compared for I3D-CNN-LSTM-Model, I3D-LSTM Model and ViT-LSTM Model.

3.1. Dataset

The experiments are performed on UCF-Crime dataset, which consists of 1900 (untrimmed) videos collected from various CCTV cameras, youtube videos etc. There are 950 normal and 950 anomaly videos in the dataset. The dataset includes 13 real-world anomalies that span a broad spectrum including Abuse, Arrest, Arson, Assault, Burglary, Explosion, Fighting, Road Accident, Robbery, Shooting, Stealing, Shoplifting, and Vandalism. The training set comprises 800 normal and 810 anomalous videos and the testing test contains 150 normal and 140 anomalous videos.

3.2. Feature Extraction and Processing

Each video V*i* is represented by frames extracted at 30 fps:

$$V_i = \{F_{i,1}, F_{i,2}, \ldots, F_{i,N_i}\}, where\ F_{i,j} \in R^{H \times W \times 3} \quad (4)$$

Here: - V*i* denotes the i-th video. - F*i,j* represents the j-th frame of video V*i* . - N*i* is the total number of frames in video V*i* . - H and W are the height and width of each frame. We have explored I3D and ViT based feature extraction.

Using I3D, we generate feature vectors of size $R^{S \times T \times F}$, where S denotes the number of frame segments analyzed, T indicates the temporal dimension of the video, and F represents the number of features extracted from each section at each time period.

The I3D model extracts spatiotemporal features from each video segment using 3D convolution:

$$Feature_{I3D}(V_i) = \{f_{i,1}, f_{i,2}, \ldots\ldots f_{i,S}\} \quad (5)$$

where each feature vector fi,k for segment k is defined as:

$$f_{i,k} = \phi_{I3D}(F_{i,j:j+T-1}), j = (k-1).T + 1 \quad (6)$$

Here: - ϕ_{I3D} is the feature extraction function of the I3D model. - $f_{i,k} \in R^F$ is the feature vector for the k-th segment. In our approach, we use ten-crop feature extraction. Each frame/clip is divided into 10 sections: top-right, top-left, bottom-right, bottom-left, and their horizontal flips. Each frame is represented by 10 sections, and 2048 features are extracted from each section. Further, we evenly segment each video into 32 temporal vectors, resulting in a final dimension of (32, 10, 2048).

On the other hand, using ViT for feature extraction enables its self-attention mechanism to capture global dependencies across different parts of an image, which is particularly beneficial for video analysis where understanding the broader context such as, the relationships among objects and the overall scene layout is essential. For ViT-based feature extraction, the model reshapes each video into a standardized format:

$$Feature_{ViT}(V_i) = \psi_{ViT}(V_i) \in R^{64 \times 1000} \quad (7)$$

where: - ψ_{ViT} represents the ViT feature extraction function. The output dimension 64 × 1000 standardizes feature representations across all videos.

3.3. Deep Learning Model

We leverage Long Short-Term Memory (LSTM) networks to model temporal dependencies and enhance the detection of anomalous events. For the I3D-CNN-LSTM-based model, the Conv3D layer output is reshaped and fed into two LSTM layers to learn long-term temporal dependencies. Finally, dense layers with ReLU activation and a sigmoid output layer are used for binary classification. The I3D-LSTM-based model consists of four LSTM layers with 100, 50, 20, and 10 units, respectively, designed to capture temporal dependencies in the input data with each layer except the last returning sequences. The final output layer is a Dense layer with a sigmoid activation function for binary classification. ViT-LSTM model begins with an LSTM layer of 100 units that returns sequences, which is tailored for the input shape of (64, 1000). This is followed by three more LSTM layers with 50, 20, and 10 units, respectively, all using ReLU activation functions. The final layer is a Dense layer with a single unit and a sigmoid activation function, ideal for binary classification tasks. The videos with output value > 0.5 are labelled anomalous. All the three models are compiled with the Adam optimizer, a learning rate of 0.0001, binary cross-entropy loss(Eq. 8).

$$Loss(y, \hat{y}) = -[y log(\hat{y}) + (1-y) log(1-\hat{y})] \quad (8)$$

where:

- y is the true label (0 or 1).
- $\hat{y}$ is the predicted probability of the positive class.

4. Results and Discussions

As VAD is a classification problem, a confusion matrix is calculated (Refer Fig. 2 based on UCF-Crime Test Dataset results (290 videos - 150 normal and 140 anomalous) to quantify the goodness of the model. However, the models usually predict the anomaly score between [0,1]. Therefore, there is always a need for a threshold to be defined in order to predict the correct class for the test video(s). If the anomaly score is above the threshold(0.5 in our case), it is considered an abnormal video, else a normal video. Because there exists data imbalance, the preferred choice of metrics is AUC-ROC curve which is a plot of (FPR, TPR). The range of ROC, AUC is [0,1]. Higher the value, better the classifier model.

Through our extensive experiments we found that the I3D LSTM-based model has performed comparatively better

as compared to its ViT counterpart.

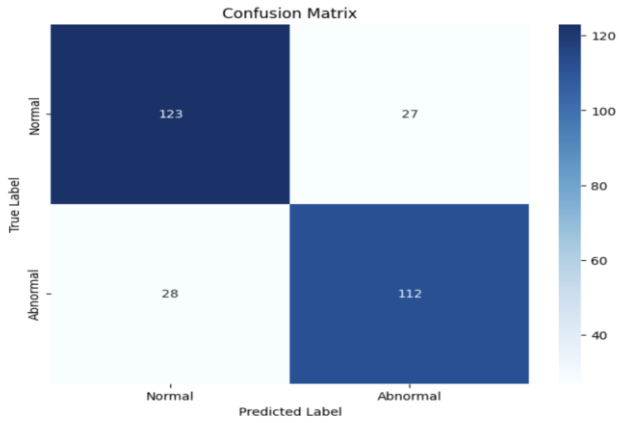

Fig. 2a. i3D-CNN-LSTM-CM

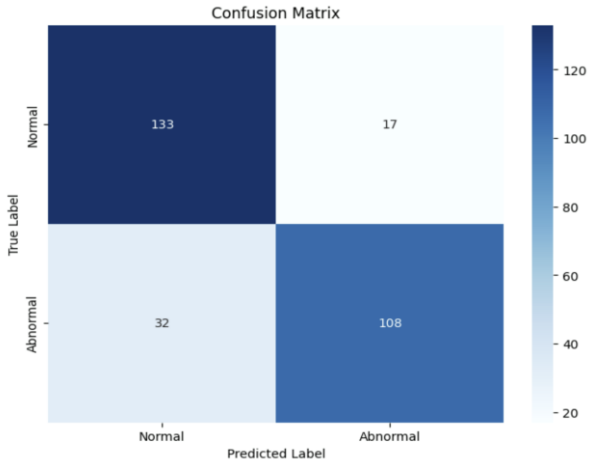

Fig. 2b. i3d-LSTM-CM

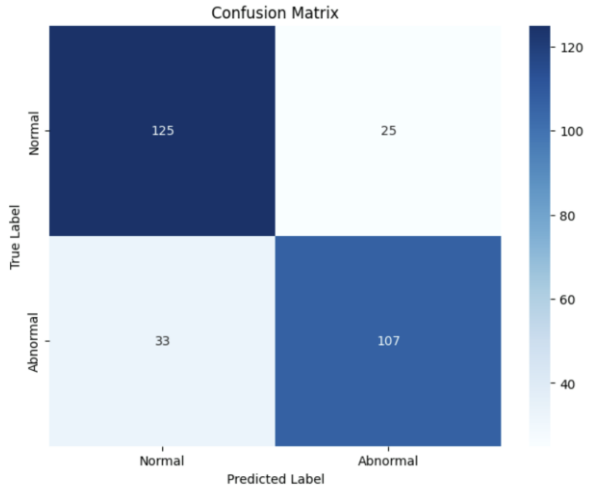

Fig. 2c. ViTLSTM-CM

Fig. 2. : Confusion Matrix of (a) I3D-CNN-LSTM, (b) I3D-LSTM Model, and (c) ViT-LSTM Model

This could be also due to the reason that the nature of anomalies in the dataset is varied and each anomaly has a very low proportion of videos in the dataset as compared to normal videos. Following previous work on WSVAD, we use receiver operating characteristic (ROC) curve and corresponding area under the curve (AUC) to evaluate the performance of our models. Using I3D-LSTMbased model produced an overall AUC of 90% as compared to overall AUC of 85.98% while using the ViT-LSTM-based model. Table 1 summarizes the results. Fig. 3 throws light on the AUC-ROC curve for each of the models.

Table 1. Test Results on UCF-Crime

Model	Accuracy	AUC	Loss
I3D-CNN-LSTM-Based	81.03	86.73	14.82
I3D-LSTM-Based	83.10	90.00	13.88
ViT-LSTM-Based	80.00	85.98	15.42

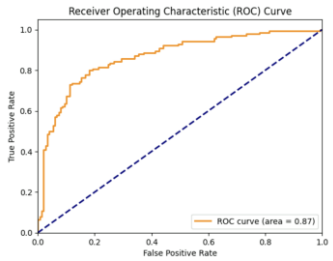

Fig 3a : i3D-CNN-LSTM-AUC-ROC

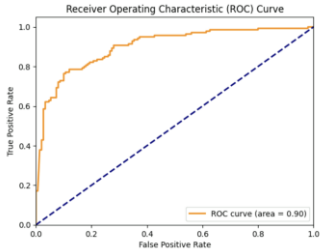

Fig 3a : i3D-LSTM-AUC-ROC

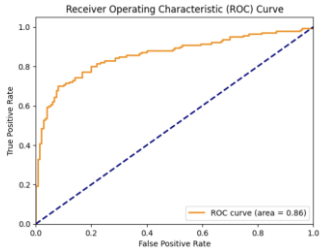

Fig 3c : ViT-LSTM-AUC-ROC

Fig. 3. : AUC-ROC Curve (a) I3D-CNN-LSTM, (b) I3D-LSTM Model, and (c) ViT-LSTM Model

To evaluate the practicality of the framework, a detailed analysis of the computational requirements was performed. ViT-LSTM has a higher computational requirement as it effectively captures global dependencies - the average inference time per video was 150 ms with 75% of GPU utilization on RTX 4090. For resource-constrained environments, I3D based models are more feasible as the average inference time per video was 120 ms with 60% GPU utilization on RTX 4090.

We further investigated and collected the misclassification statistics for the three prediction

pipelines using the same test dataset. Refer to Table 2 for the details. We identify the commonly mispredicted videos among all three models as the most challenging. These videos include: *Arrest039_x264, Explosion011_x264, Explosion016_x264, Normal_Videos_010_x264, Normal_Videos_478_x264, Normal_Videos_606_x264, Normal_Videos_887_x264, Normal_Videos_901_x264, Normal_Videos_925_x264, RoadAccidents004_x264, RoadAccidents125_x264, RoadAccidents133_x26, Robbery102_x264, Shooting048_x264, and Shoplifting033_x264.*

The normal videos mostly got misclassified due to factors such as night vision and individuals wearing hoodies, sudden changes in light intensity etc. Additionally, there aren't enough videos in the dataset that depict explosions inside the buildings when viewed from outside, while most of the explosion videos had a clear smoke visibility. The misclassification of Road Accident videos primarily occurred due to high-speed vehicles and scenes being captured from considerable distance. To address these challenges, a) a label denoising mechanism(e.g., Graph convolutional network) can be incorporated to refine noisy labels. b) use techniques like oversampling for underrepresented anomaly types to balance the dataset.

Table 2. Misclassification Statistics between the 3 Prediction Pipelines for the Same Test Set

Category	Test Set Samples	I3D-CNN-LSTM	I3D-LSTM	ViT-LSTM
Abuse	2	0	1	1
Arrest	5	2	1	4
Arson	9	0	1	1
Assault	3	0	0	0
Burglary	13	2	1	1
Explosion	21	2	2	5
Fighting	5	0	1	0
Road Accidents	23	9	11	5
Robbery	5	1	2	2
Shooting	23	4	4	5
Shoplifting	21	6	7	7
Stealing	5	1	1	0
Vandalism	5	1	0	2
Normal	150	27	17	25
Total	290	55	49	58

Fig. 4 provides a performance comparison of popular state-of-the-art(SOTA) models on the UCF-Crime Dataset, measured by area under the ROC curve (AUC %). The compared models are [5, 14, 16, 18, 22, 24, 27, 32, 43, 45, 48]

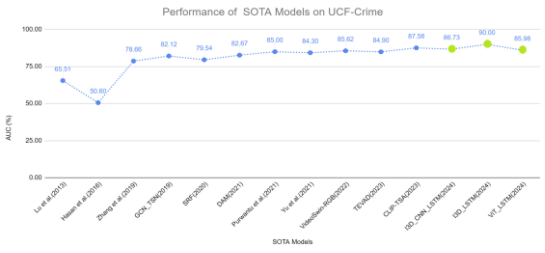

Fig. 4. : Model Performance compared among popular SOTA Models

5. Conclusion and Future Work

In this paper, we have addressed the complex challenge of video anomaly detection (VAD) using a weakly-supervised framework that leverages Long Short-Term Memory (LSTM) networks to capture temporal dependencies in video sequences. Our comparative analysis of two advanced feature extraction techniques, Inflated 3D ConvNet (I3D) and Vision Transformer (ViT), has revealed valuable insights into their efficacy. The results indicate that the I3D-based model outperforms the ViT-based model, achieving an Area Under the Curve (AUC) score of 90% compared to 86% with ViT. The proposed method provides a robust solution to handle imbalanced datasets and anomalies with minimal deviations from normal activities.

Future work could explore further refinements at each stage of the pipeline to reflect the complexity of real-world surveillance systems for example, in the feature extraction process, the integration of additional contextual information, and handling multi-modal data. For example, sound patterns during anomalous events (e.g., gunshots or explosions) could complement visual cues. Combining these modalities through attention mechanisms will likely enhance the robustness of anomaly detection. With the advent of Industry 4.0, there is a need for deploying these models at the edge, which is extremely challenging given the very nature of the problem. To reduce computational demands while preserving performance, optimizations methods such as model compression, model pruning, quantization and knowledge distillation can be applied.

Author contributions

Conceptualization, P.K. and S.S.S., P.B.H; methodology, P.K., S.S.S. implementation, preparing figures, tables, P.K. validation, P.K., S.S.S. and P.B.H. writing - original draft preparation, P.K. writing - review and editing, P.K., S.S.S. supervision, S.S.S. and P.B.H. project administration, S.S.S and P.B.H. All authors have read and agreed to the published version of the manuscript.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] B. Antic and B. Ommer, "Video parsing for abnormality detection," *2011 International Conference on Computer Vision*, pp. 2415–2422, IEEE, 2011.

[2] G. Bertasius, H. Wang, and L. Torresani, "Is spacetime attention all you need for video understanding?" *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

[3] J. Carreira and A. Zisserman, "Quo vadis, action

recognition? A new model and the kinetics dataset," *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[4] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," *CoRR*, vol. abs/1901.03407, 2019.

[5] W. Chen, K. T. Ma, Z. J. Yew, M. Hur, and D. A.-A. Khoo, "Tevad: Improved video anomaly detection with captions," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5548–5558, 2023.

[6] Y. Chen, Z. Liu, B. Zhang, W. Fok, X. Qi, and Y.-C. Wu, "MGFN: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection," *arXiv preprint arXiv:2211.15098*, 2022.

[7] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[8] M. Fayyaz and J. Gall, "SCT: Set constrained temporal transformer for set supervised action segmentation," *CoRR*, vol. abs/2003.14266, 2020.

[9] J.-C. Feng, F.-T. Hong, and 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*, pp. 14009–14018, 2021.

[10] D. Gong et al., "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[11] Y. Gong et al., "Multiscale continuity-aware refinement network for weakly supervised video anomaly detection," *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2022.

[12] S. Gopalakrishnan, "A public health perspective of road traffic accidents," *Family Medicine and Primary Care*, 2012.

[13] K. Han et al., "A survey on vision transformer," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, pp. 1–1, 2020.

[14] M. Hasan et al., "Learning temporal regularity in video sequences," *Proceedings of IEEE Computer Vision and Pattern Recognition*, 2016.

[15] R. Hinami, T. Mei, and S. Satoh, "Joint detection and recounting of abnormal events by learning deep generic knowledge," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3619–3627, 2017.

[16] H. K. Joo, K. Vo, K. Yamazaki, and N. Le, "CLIPTSA: Clip-assisted temporal self-attention for weakly supervised video anomaly detection," *2023 IEEE International Conference on Image Processing (ICIP)*, pp. 3230–3234, IEEE, 2023.

[17] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1446–1453, IEEE, 2009.

[18] S. Li, F. Liu, and L. Jiao, "Self-training multisequence learning with transformer for weakly supervised video anomaly detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 24, 2022.

[19] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection: A new baseline," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6536–6545, 2018.

[20] Y. Liu, J. Liu, W. Ni, and L. Song, "Abnormal event detection with self-guiding multi-instance ranking framework," *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 01–07, IEEE, 2022.

[21] Y. Liu, J. Liu, X. Zhu, D. Wei, X. Huang, and L. Song, "Learning task-specific representation for video anomaly detection with spatio-temporal attention," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2190–2194, IEEE, 2022.

[22] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 fps in MATLAB," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2720–2727, 2013.

[23] W. Luo, W. Liu, and S. Gao, "A revisit of sparse coding based anomaly detection in stacked RNN framework," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 341–349, 2017.

[24] S. Majhi, S. Das, and F. Bremond, "DAM: Dissimilarity attention module for weakly-supervised video anomaly detection," *2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–8, IEEE, 2021.

[25] S. Majhi, R. Dash, and P. K. Sa, "Two-stream CNN architecture for anomalous event detection in real-world scenarios," *International Conference on Computer Vision and Image Processing*, pp. 343–

353, Springer, 2020.

[26] G. Pang, C. Shen, L. Cao, and A. Van Den Hengel, "Deep learning for anomaly detection: A review," *ACM Computing Surveys (CSUR)*, vol. 54, no. 2, pp. 1–38, 2021.

[27] 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," *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 173–183, 2021.

[28] M. Ravanbakhsh, M. Nabi, H. Mousavi, E. Sangineto, and N. Sebe, "Plug-and-play CNN for crowd motion analysis: An application in abnormal event detection," *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1689–1698, IEEE, 2018.

[29] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, and N. Sebe, "Abnormal event detection in videos using generative adversarial nets," *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 1577–1581, IEEE, 2017.

[30] L. Ruff et al., "Deep one-class classification," *Proceedings of the 35th International Conference on Machine Learning*, vol. 80, pp. 4393–4402, PMLR, 2018.

[31] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6479–6488, 2018.

[32] Y. Tian et al., "Weakly-supervised video anomaly detection with robust temporal feature magnitude learning," *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4975–4986, 2021.

[33] D. Tran et al., "Learning spatiotemporal features with 3D convolutional networks," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4489–4497, IEEE, 2015.

[34] A. Ullah et al., "One-shot learning for surveillance anomaly recognition using Siamese 3D CNN," *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020.

[35] W. Ullah et al., "Intelligent dual stream CNN and echo state network for anomaly detection," *Knowledge-Based Systems*, vol. 253, p. 109456, 2022.

[36] W. Ullah et al., "Artificial intelligence of things-assisted two-stream neural network for anomaly detection in surveillance big video data," *Future Generation Computer Systems*, vol. 129, pp. 286–297, 2022.

[37] W. Ullah et al., "An efficient anomaly recognition framework using an attention residual LSTM in surveillance videos," *Sensors*, vol. 21, no. 8, p. 2811, 2021.

[38] B. Wan et al., "Weakly supervised video anomaly detection via center-guided discriminative learning," *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2020.

[39] J. Wu et al., "Weakly-supervised spatio-temporal anomaly detection in surveillance video," *The Thirtieth International Joint Conference on Artificial Intelligence*, 2021.

[40] [40] P. Wu et al., "Not only look, but also listen: Learning multimodal violence detection under weak supervision," *European Conference on Computer Vision (ECCV)*, 2020.

[41] J. Yin et al., "LiDAR-based online 3D video object detection with graph-based message passing and spatiotemporal transformer attention," *CoRR*, vol. abs/2004.01389, 2020.

[42] S. Yu et al., "Cross-epoch learning for weakly supervised anomaly detection in surveillance videos," *IEEE Signal Processing Letters*, vol. 28, pp. 2137–2141, 2021.

[43] M. Z. Zaheer et al., "A self-reasoning framework for anomaly detection using video-level labels," *IEEE Signal Processing Letters*, vol. 27, pp. 1705–1709, 2020.

[44] D. Zhang et al., "Weakly supervised video anomaly detection via transformer-enabled temporal relation learning," *IEEE Signal Processing Letters*, 2022.

[45] J. Zhang et al., "Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection," *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 4030–4034, 2019.

[46] Y. Zhang et al., "Video anomaly detection based on locality sensitive hashing filters," *Pattern Recognition*, vol. 59, pp. 302–311, 2016.

[47] B. Zhao, F.-F. Li, and E. P. Xing, "Online detection of unusual events in videos via dynamic sparse coding," *CVPR 2011*, pp. 3313–3320, IEEE, 2011.

[48] J.-X. Zhong et al., "Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection," *Proceedings of the IEEE*

Conference on Computer Vision and Pattern Recognition, pp. 1237–1246, 2019.

[49] Y. Zhu and S. Newsam, "Motion-aware feature for improved video anomaly detection," *arXiv preprint arXiv:1907.10211*, 2019.