Anomaly detection in surveillance videos using transformer based attention model

Kapil Deshpande
mit2020040@iiita.ac.in
Indian Institute of Information Technology Allahabad
Prayagraj, Uttar Pradesh, India

Narinder Singh Punn
pse2017002@iiita.ac.in
Indian Institute of Information Technology Allahabad
Prayagraj, Uttar Pradesh, India

Sanjay Kumar Sonbhadra
sanjaykumarsonbhadra@soa.ac.in
Shiksha 'O' Anusandhan
Odisha, Bhubaneswar, India

Sonali Agarwal
sonali@iiita.ac.in
Indian Institute of Information Technology Allahabad
Prayagraj, Uttar Pradesh, India

ABSTRACT

Surveillance footage can catch a wide range of realistic anomalies. This research suggests using a weakly supervised strategy to avoid annotating anomalous segments in training videos, which is time consuming. In this approach only video level labels are used to obtain frame level anomaly scores. Weakly supervised video anomaly detection (WSVAD) suffers from the wrong identification of abnormal and normal instances during the training process. Therefore it is important to extract better quality features from the available videos. With this motivation, the present paper uses better quality transformer-based features named Videoswin Features followed by the attention layer based on dilated convolution and self attention to capture long and short range dependencies in temporal domain. This gives us a better understanding of available videos. The proposed framework is validated on real-world dataset i.e. ShanghaiTech Campus dataset which results in competitive performance than current state-of-the-art methods. The model and the code are available at https://github.com/kapildeshpande/Anomaly-Detection-in-Surveillance-Videos.

CCS CONCEPTS

• **Computing methodologies → Knowledge representation and reasoning**; **Computer vision**.

KEYWORDS

Video anomaly detection, Weakly supervised, Videoswin features, Attention layer

ACM Reference Format:
Kapil Deshpande, Narinder Singh Punn, Sanjay Kumar Sonbhadra, and Sonali Agarwal. 2022. Anomaly detection in surveillance videos using transformer based attention model. In *International Workshop on Urban Computing, August 15, 2022, Washington, DC*. ACM, New York, NY, USA, 6 pages. https://doi.org/xx.xxxx/xxxxxxx.xxxxxxx

KDD 2022, August 14-18, 2022, Washington DC, USA

ACM ISBN xxx-x-xxxx-xxxx-x/xx/xx.
https://doi.org/xx.xxxx/xxxxxxx.xxxxxxx

1 INTRODUCTION

Video anomaly detection has gained a lot of attention due to its applications in surveillance systems. The cost of deploying surveillance systems has reduced significantly in recent years but it still requires human intervention in detecting anomalous events like fighting, abusing, stealing, etc. Considering the additional cost of human labor and the loss of productive time, the development of intelligent algorithms for video anomaly detection is required. The vague nature of anomaly and the unavailability of annotated data makes anomaly detection difficult. There are various unsupervised [3, 31] and weakly supervised [13, 21, 22] solutions present. Generally, unsupervised anomaly detection tries to learn the distribution of normal events and mark outliers as anomalies. Since it is impossible to learn all possible normal events distribution therefore this model is highly biased and fails in the case of real-world events. It tries to combat the above problem by using both normal and anomalous events while training.

Weakly supervised approaches require significantly less effort as compared to supervised learning because it requires only video-level labels instead of frame-level. However the major challenge of weakly supervised approaches is in identifying the abnormal snippets from the anomalous videos, this is because: the abnormal videos contain a large number of normal snippets and the abnormal events can have only slight differences from normal events. All these above issues can be resolved by using multiple instance learning (MIL) [26] where the training set is divided into the same numbers of abnormal and normal snippets. Two bags are created i.e. a normal bag which contains snippets from a normal video and an abnormal bag that contains snippets from an abnormal video. The snippet with the maximum anomaly score is selected from each bag and the loss is back propagated. Although this method partially addresses the previous issues but it also introduces the following problems: The highest anomaly score can be from normal bag instead of abnormal bag and when anomaly videos have multiple anomaly in the single video it fails to leverage additional anomalies because it only take the snippet with highest anomaly score from the abnormal bag.

To address all these issues the proposed solution used the Robust Temporal Feature Magnitude (RTFM) learning model inspired by Tian et al. [27]. This model relies on the temporal feature magnitude i.e. l2 norm of features for anomaly detection, where normal snippets are represented by low magnitude features, while abnormal snippets are represented by high magnitude features. This model try to maximize δ_{score} which denotes the difference between the mean of l2 norm of top K features from the abnormal and normal bag where K is the number of abnormal snippets in an abnormal video. This solves the previously discussed problem because: It's more likely to choose anomalous snippets from abnormal videos instead of normal videos. It can utilize multiple anomalies in the anomaly videos which will result in better utilization of training data.

It is important to extract better features from the available videos to avoid the wrong identification of abnormal and normal instances during the training process. Researchers have been inspired to employ video transformers as feature extractors to handle anomaly detection tasks as a result of their recent success with video classification tasks. Therefore the proposed model uses a transformer based features named Videoswin Features [15] which have consistently outperformed the CNN based models like I3D [6], C3D [29], etc. The feature extraction is followed by an attention layer based on dilated convolution to capture most relevant long and short range dependencies [1, 2, 23, 24]. The proposed solution is validated on a real-world dataset i.e. ShanghaiTech Campus dataset which results in competitive performance than current state-of-the-art methods. The major contribution of the present research work are described below:

- To improve the understanding of given videos, a newer transformer based feature extraction model is used named videoswin transformer.
- To highlight relevant features an attention layer based on dilated convolution and self attention that captures long and short range temporal dependencies.
- A comparative study with current state-of-the-art approaches is conducted to examine the effects of the proposed model on the open source ShanghaiTech dataset. The proposed model achieved competitive performance (AUC score) than current state-of-the-art methods.

The rest of the paper is divided into various sections, where Section 2 covers the prevailing work in anomaly detection, followed by proposed methodology in Section 3. Section 4 presents the experimental analysis and finally concluding remarks are presented in Section 5.

2 RELATED WORK

Traditional video anomaly detection uses unsupervised learning [16, 35] algorithms where it tries to learn the distribution of normal events and mark outliers as anomalies. Since it is impossible to learn all possible normal events distribution therefore this model is highly biased and fails in the case of real-world events. Other methods use one-class classification [3, 31] assuming only normal labeled data is available. Some approaches rely on tracking [4, 33] to model people's regular movement and identify deviations as anomalies. Since it is tough to acquire accurate information of tracks, numerous strategies for avoiding tracking and learning global motion patterns, such as topic modeling [9], have been used, context-driven method [12] social force models [20], histogram-based methods [7], motion patterns [25], Hidden Markov Model (HMM) on local spatiotemporal volumes[11], and mixtures of dynamic textures model [12]. These techniques learn distributions of normal motion patterns from training videos of normal behaviors and discover low likely patterns as anomalies. After the initial success of the sparse representation and dictionary learning methodologies, researchers employed sparse representation [17, 36] to learn the dictionary of patterns. Where anomalous patterns have high reconstruction errors during testing. After the initial success of deep learning in image classification, many techniques for video action classification [10, 28] have been developed.

Alternatively, some approaches rely on data reconstruction utilizing generative models to learn normal sample representations by (adversarial) reducing the reconstruction error [30, 34, 38]. These methods presume that undetected abnormal videos/images can often be poorly reconstructed, and samples with large reconstruction errors are considered anomalies. These techniques may overfit the training data and fail to distinguish abnormal from normal events due to a lack of prior knowledge about anomaly.

To solve the above problems Sultani et al. [26] introduced a weakly supervised solution that can learn anomaly patterns by using both normal and anomaly videos using MIL-based models with CNN as the backbone for feature extraction. However, it fails to separate noise present in the positive bag and this can lead to normal snippets being mistaken as abnormal. In this context, to clear the noise present in the positive bag Zhong et al. [37] proposed to use a graph convolution neural network. Although it partially solved the problem, it was computationally heavy. The RTFM model [27] which solves the above problem by using the l2 norm-based ranking loss function. Although it still uses CNN as the backbone for feature extraction.

To capture consistency between successive frames, traditional attention approaches employ consecutive frames and transform them into handcrafted motion trajectories. Other methods such as stacked RNN [18], LSTM [13], convolutional LSTM [13], and GCN-based [37] can capture short-range fixed-order temporal dependencies but they either fail to capture long-range dependencies or they are computationally expensive. Following this context, the proposed attention layer uses dilated convolution based attention mechanism which captures short and long-range temporal dependencies and is computationally inexpensive as compared to other methods. The proposed method uses a more effective transformer-based model for feature extraction and a temporal attention layer for necessary feature enhancement, thus improving the model's overall performance.

3 PROPOSED METHOD

The proposed framework is divided into 3 stages as given in Fig. 1. As the exact location or frame-level labels are not provided for learning, the proposed solution follows a weakly supervised learning where the videos of different duration are divided into a fixed number of snippets containing the same number of frames. The proposed solution assumed that snippets obtained from an anomaly video contain at least one anomaly snippet, but the snippets from normal videos contain all normal snippets. In stage 1, A pre-trained videoswin model for feature extraction of the snippets. In stage 2 an attention layer is applied to the feature to capture relevant long and short range dependencies in the temporal domain. At last in

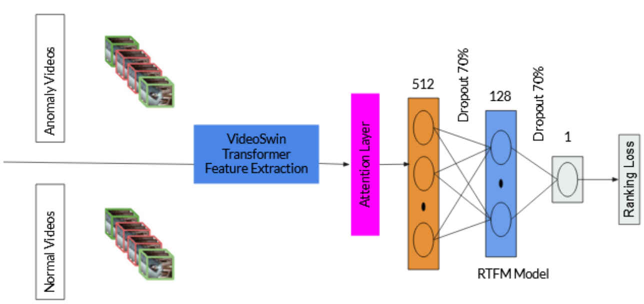

Figure 1: The proposed model architecture.

stage 3 RTFM model is used for anomaly detection on the features obtained from stage 3.

3.1 Stage 1 (Feature Extraction)

To extract features this paper uses videoswin transformer model which is trained on large-scale datasets like Kinetics [5] and ImageNet [8]. The use of a pretrained model allows us to extract better quality features. Traditional transformer models calculate self-attention with respect to all the elements present but in the case of images, it is computationally expensive to perform. To solve this issue, the swin transformer divides images into windows and calculates self-attention inside this window only. Now it slides the window on the images to get the self-attention value of the whole set of images more efficiently.

To begin the feature extraction process, the first step is to divide the videos into frames (of size let's say H × W). Now the set of these frames (T) makes the video for feature extraction where each video has RGB channels. This gives us the input dimension as N × C × T × H × W, where N is the batch size and C is the number of channels.

3.2 Stage 2 (Attention Layer)

The main objective of this stage is to learn the discriminative representation of normal and abnormal snippets by improving the quality of the feature map obtained from stage 1. This objective is achieved using an attention layer that can encode the long and short range dependencies in temporal domain on the feature map while drawing focus of the model towards most relevant features.

The proposed attention layer is shown in Fig. 2. Given an input feature map $F \in \mathbb{R}^{T \times D}$, it produces the output attention feature maps $F' \in \mathbb{R}^{T \times D}$. It consists of two modules, the one on the left is a short range module, it is used to capture short-term temporal dependencies and the one on the right is a long range module it is used to compute global temporal context.

To calculate the global temporal context, the pairwise temporal self attention is calculated which produces the feature map $M \in \mathbb{R}^{T \times T}$. It first applies the conv1D layer to reduce information to $F^c \in \mathbb{R}^{T \times D/4}$ where $F^c = \text{conv1D}(F)$, then it applies 3 conv1d layers separately. $F^{c1} = \text{conv1D}(F^c)$, $F^{c2} = \text{conv1D}(F^c)$, $F^{c3} = \text{conv1D}(F^c)$. It will combine these 3 conv1D layers with $F^{c4} = \text{conv1D}((F^{c1} * (F^{c2})^T) * F^{c3})$. A residual is added, which gives the final output, $M = F^{c4} + F^c$, where $M \in \mathbb{R}^{T \times T}$.

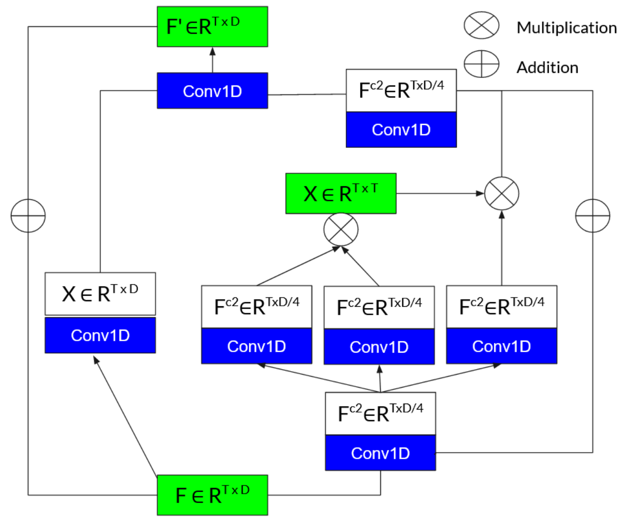

Figure 2: The proposed attention layer architecture

To calculate the short term temporal dependencies it applies the conv1D layer which gives it the output, $K = \text{conv1D}(F)$, where $F \in \mathbb{R}^{T \times D}$. The output M from the long range module is concatenated with the output K from the short range module and a residual connection is added to give us the final output, $F' = \text{concat}(M, K) + F^c$, where $F \in \mathbb{R}^{T \times D}$.

3.3 Stage 3 (Anomaly Detection)

The proposed anomaly detection model uses Robust Temporal Feature Magnitude Learning (RTFM) model, in which temporal feature magnitude i.e. l2 norm of video snippets are used for anomaly detection where normal snippets are represented by low magnitude features, while abnormal snippets are represented by high magnitude features. The proposed model assumes that anomalous snippets have a larger mean feature magnitude than normal snippets.

Let $\|x\|$ be the feature magnitude of snippets where x^+ means abnormal snippet and x^- means normal snippet, which are obtained by normal (X^+) and abnormal (X^-) videos. Model learns by trying to maximize the $\delta_{\text{score}}(X^+, X^-)$ which denotes the difference between the mean of l2 norm of topK features from the abnormal and normal bag where k is the number of abnormal snippets in abnormal video. To maximize the $\delta_{\text{score}}(X^+, X^-)$, the loss function (shown in 1) is optimized during backpropagation.

$$L(X^+, X^-) = max(0, m - mean(topK(\|X^+\|)) + mean(topK(\|X^-\|)) \tag{1}$$

where m is a constant predefined margin.

A binary cross-entropy based loss function is applied to learn the snippet classifier as shown in 2. It trains a snippet classifier with 0 and 1 class labels indicating normal and abnormal snippets respectively.

$$loss = -ylog(x) + (1 - y)log(1 - x) \tag{2}$$

where x is the mean of l2 norm of topK features, $x = mean(topK(\|X^+\|))$, and y is the binary value indicating actual class labels as normal or abnormal.

$$Smoothness = \Sigma f(v^i) - f(v^{i+1})^2 \tag{3}$$

Temporal smoothness is used between consecutive video snippets to vary anomaly score smoothly between video snippets.

$$Sparsity = \Sigma f(v_\text{i}) \tag{4}$$

Anomaly frequently happens over a brief period of time in real-world circumstances which leads to sparse anomaly scores of segments in the anomalous bag. To avoid this issue, a sparsity term is used.

$$\begin{aligned} Finalloss = \lambda 1 * \text{Eq.1} + \lambda 2 * \text{Eq.2}+ \\ \lambda 3 * \text{Eq.3} + \lambda 4 * \text{Eq.4} \end{aligned} \tag{5}$$

where λ's are the respective learning rates for the Eq.s.

4 EXPERIMENTS

4.1 Dataset Description

This paper uses a large-scale video anomaly detection dataset called the ShanghaiTech Campus dataset [14]. It includes video from fixed angle street surveillance cameras. It has 437 videos from 12 different backgrounds, with 130 anomalous and 307 normal videos. This is a popular benchmark dataset for anomaly detection tasks that uses both anomalous and normal data. To restructure the dataset into a weakly supervised training set, Zhong et al. [37] picked a sample of anomalous testing videos and turned them into training videos so that all 13 background scenes are covered by the training and testing set. To convert the dataset into weekly supervised, this paper used the same approach as used by Zhong et al. [37] and Tian et al. [27]. Fig. 3 shows the sample normal and abnormal clips from the dataset.

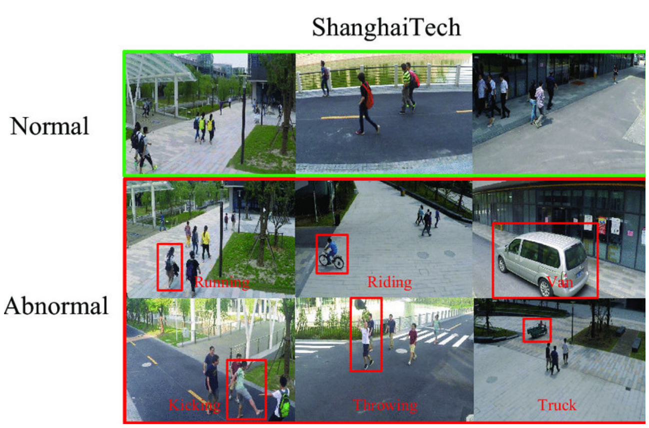

Figure 3: ShanghaiTech dataset normal and abnormal clips

4.2 Evaluation Metric

To measure the model's performance, this paper used frame-level receiver operating characteristics (ROC) as well as its area under the curve (AUC) score, following the previous methods [26, 27, 37]. AUC score is a measure of separability. It represents the model's ability to discriminate between classes. An AUC score of 1 means the model can separate both classes perfectly. AUC score of 0 means the model is reciprocating the results means it predicts positive class as negative and vice-versa. AUC score of 0.5 means the model has no class separation capacity. The ROC curve is plotted at all possible classification thresholds by calculating the values of TPR and FPR at every threshold from 0 to 1.

4.3 Implementation Details

For feature extraction from pre-trained videoswin transformer model on Kinetics dataset, each video is divided into frames of size 224 x 224, for video is divided into T = 32 temporal segments where each segment is 16 frames long, this gives us ten crop features of dimensions 32 x 1024. Cropping snippets into the four corners, center and their flipped form is referred as ten cropping.

In Eq. 1 the margin, $m = 100$ and the value of $k = 3$. The 3 fully connected (FC) layers in the RTFM model have 512, 128 and 1 nodes respectively where 1^{st} and 2^{nd} FC layer is followed by a ReLU activation function and the last layer is followed by sigmoid function. A dropout function is added after every layer with rate = 0.7. The model is trained using adam optimizer with weight decay of 0.005 and learning rate of 0.001 with batch size = 32 for 500 epochs. Each mini batch has 32 samples chosen at random from normal and abnormal videos. For fare comparison, this paper used the same benchmark setup used by Sultani et al.[26], Zhong et al. [37] and Tian et al. [27].

4.4 Result Analysis

The results are reported on the ShanghaiTech Campus dataset [14]. Where two backbone models for feature extraction are used namely I3D [6] and videoswin [15]. Comparisons with the previous weakly supervised solutions are given in Table 1 and visually presented in Fig. 4. Furthermore, the inference drawn is analysed with the help of ROC curves as given in Fig. 5.

The usage of videoswin features leads to better performance than I3D features because of improved video understanding. MIL model specially performed better with videoswin features, it even outperformed the RTFM model which was previously performing better than MIL with I3D features. To compare the attention layer introduced, this paper added a different attention layer to previous methods namely LSTM, CBAM [32], RTFM's Attention Layer [27]. The results obtained are given in Table 2. With the introduction of the proposed attention layer a better AUC score of around 1% is acquired. The usage of LSTM and CBAM results in decreased

Table 1: The comparative analysis of video anomaly detection models. The best outcomes are shown in bold font.

Method	Feature	AUC
MIL	I3D	92.3
MIL	Videoswin	96.9
RTFM	I3D	93.0
RTFM	Videoswin	96.4
Proposed Model	I3D	93.7
Proposed Model	**Videoswin**	**97.9**

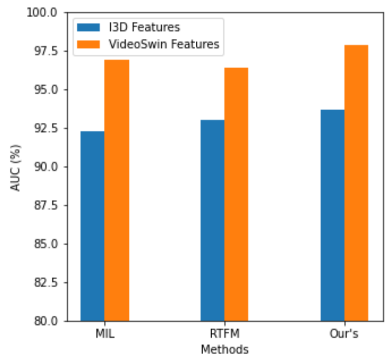

Figure 4: Comparison I3D vs Videoswin Features

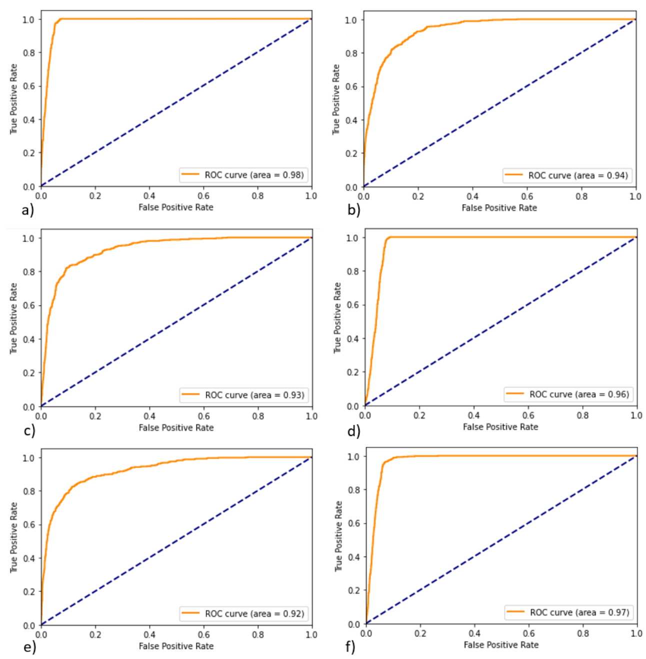

Figure 5: ROC curves of the a) proposed model using videoswin backbone, b) proposed model using I3D backbone, c) RTFM model using I3D backbone, d) RTFM model using videoswin backbone, e) MIL model using I3D backbone, and f) MIL model using videoswin backbone

performance because the models fail when high dimensional feature maps are given as input [19].

5 CONCLUSION

In this research work, a weakly supervised strategy is proposed. It uses better quality features extracted from videoswin transformer model, followed by an attention layer to encode the long and short range dependencies in the temporal domain. The use of the robust temporal feature magnitude (RTFM) model makes this approach better than multiple instance learning (MIL) based techniques because it learns more discriminative features than the MIL model and it exploits abnormal data more easily. It is found from experiments that the use of better quality features and an improved attention layer resulted in improved performance of the model. In future, more experiments can be performed by exploring different strategies to minimize the noise present in the positive bag.

Table 2: The comparative analysis of various attention layers on the video anomaly detection models. The best outcomes are shown in bold font.

Method	**Feature**	**AUC**
MIL + LSTM	I3D	89.0
MIL + LSTM	Videoswin	96.6
RTFM + LSTM	I3D	89.0
RTFM + LSTM	Videoswin	96.6
MIL + CBAM	I3D	88.0
MIL + CBAM	Videoswin	96.9
RTFM + CBAM	I3D	87.5
RTFM + CBAM	Videoswin	96.2
RTFM + No Attention	I3D	91.0
RTFM + No Attention	Videoswin	97.1
Proposed Model	**Videoswin**	**97.9**

REFERENCES

[1] Sonali Agarwal and GN Pandey. 2010. SVM based context awareness using body area sensor network for pervasive healthcare monitoring. In *Proceedings of the First International Conference on Intelligent Interactive Technologies and Multimedia*. 271–278.

[2] Prachi Agrawal, Narinder Singh Punn, Sanjay Kumar Sonbhadra, and Sonali Agarwal. 2021. Impact of Attention on Adversarial Robustness of Image Classification Models. In *2021 IEEE International Conference on Big Data (Big Data)*. IEEE, 3013–3019.

[3] Arslan Basharat, Alexei Gritai, and Mubarak Shah. 2008. Learning object motion patterns for anomaly detection and improved object detection. In *2008 IEEE conference on computer vision and pattern recognition*. IEEE, 1–8.

[4] Arslan Basharat, Alexei Gritai, and Mubarak Shah. 2008. Learning object motion patterns for anomaly detection and improved object detection. In *2008 IEEE conference on computer vision and pattern recognition*. IEEE, 1–8.

[5] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. 2019. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987* (2019).

[6] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.

[7] Xinyi Cui, Qingshan Liu, Mingchen Gao, and Dimitris N Metaxas. 2011. Abnormal detection using interaction energy potentials. In *CVPR 2011*. IEEE, 3161–3167.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[9] Timothy Hospedales, Shaogang Gong, and Tao Xiang. 2009. A markov clustering topic model for mining behaviour in video. In *2009 IEEE 12th International Conference on Computer Vision*. IEEE, 1165–1172.

[10] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 1725–1732.

[11] Louis Kratz and Ko Nishino. 2009. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 1446–1453. https://doi.org/10.1109/CVPR.2009.5206771

[12] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. 2014. Anomaly Detection and Localization in Crowded Scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 1 (2014), 18–32. https://doi.org/10.1109/TPAMI.2013.111

[13] Wen Liu, Weixin Luo, Zhengxin Li, Peilin Zhao, Shenghua Gao, et al. 2019. Margin Learning Embedded Prediction for Video Anomaly Detection with A Few Anomalies.. In *IJCAI*. 3023–3030.

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

[15] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2021. Video swin transformer. *arXiv preprint arXiv:2106.13230* (2021).

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

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

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

[19] Snehashis Majhi, Ratnakar Dash, and Pankaj Kumar Sa. 2020. Temporal Pooling in Inflated 3DCNN for Weakly-supervised Video Anomaly Detection. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. IEEE, 1–6.

[20] Ramin Mehran, Alexis Oyama, and Mubarak Shah. 2009. Abnormal crowd behavior detection using social force model. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 935–942. https://doi.org/10.1109/CVPR.2009.5206641

[21] Guansong Pang, Longbing Cao, Ling Chen, and Huan Liu. 2018. Learning representations of ultrahigh-dimensional data for random distance-based outlier detection. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 2041–2050.

[22] Guansong Pang, Chunhua Shen, and Anton van den Hengel. 2019. Deep anomaly detection with deviation networks. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 353–362.

[23] Narinder Singh Punn and Sonali Agarwal. 2022. CHS-Net: A Deep Learning Approach for Hierarchical Segmentation of COVID-19 via CT Images. *Neural Processing Letters* (2022), 1–22.

[24] Narinder Singh Punn and Sonali Agarwal. 2022. RCA-IUnet: a residual cross-spatial attention-guided inception U-Net model for tumor segmentation in breast ultrasound imaging. *Machine Vision and Applications* 33, 2 (2022), 1–10.

[25] Imran Saleemi, Khurram Shafique, and Mubarak Shah. 2008. Probabilistic modeling of scene dynamics for applications in visual surveillance. *IEEE transactions on pattern analysis and machine intelligence* 31, 8 (2008), 1472–1485.

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

[27] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W Verjans, and Gustavo Carneiro. 2021. Weakly-supervised video anomaly detection with contrastive learning of long and short-range temporal features. (2021).

[28] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 4489–4497.

[29] Du Tran, Lubomir D Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2014. C3D: generic features for video analysis. *CoRR, abs/1412.0767* 2, 7 (2014), 8.

[30] Shashanka Venkataramanan, Kuan-Chuan Peng, Rajat Vikram Singh, and Abhijit Mahalanobis. 2019. Attention guided anomaly detection and localization in images. *arXiv preprint arXiv:1911.08616* (2019).

[31] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. 2014. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1386–1393.

[32] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*. 3–19.

[33] Shandong Wu, Brian E. Moore, and Mubarak Shah. 2010. Chaotic invariants of Lagrangian particle trajectories for anomaly detection in crowded scenes. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2010), 2054–2060.

[34] Dan Xu, Elisa Ricci, Yan Yan, Jingkuan Song, and Nicu Sebe. 2015. Learning deep representations of appearance and motion for anomalous event detection. *arXiv preprint arXiv:1510.01553* (2015).

[35] Bin Zhao, Li Fei-Fei, and Eric P. Xing. 2011. Online detection of unusual events in videos via dynamic sparse coding. *CVPR 2011* (2011), 3313–3320.

[36] Bin Zhao, Li Fei-Fei, and Eric P. Xing. 2011. Online detection of unusual events in videos via dynamic sparse coding. *CVPR 2011* (2011), 3313–3320.

[37] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H Li, and Ge Li. 2019. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1237–1246.

[38] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Dae ki Cho, and Haifeng Chen. 2018. Deep Autoencoding Gaussian Mixture Model for Unsupervised Anomaly Detection. In *ICLR*.