Journal of Computer Science

Research Article

Anomaly Detection Based on Vision Transformer Model and Texture Features

Mohammed Lahraichi, Abdelhafid Berroukham and Khalid Housni

Laboratory of Research in Informatics L@RI, Department of Computer Science, Faculty of Sciences, Ibn Tofail University, Kenitra, Morocco

Article history
Received: 20-01-2025
Revised: 04-04-2025
Accepted: 26-02-2025

Corresponding Author:
Mohammed Lahraichi
Laboratory of Research in Informatics L@RI, Department of Computer Science, Faculty of Sciences, Ibn Tofail University, Kenitra, Morocco
Email: lahraichi.mohamed@gmail.com

Abstract: Anomaly detection is one of the video surveillance applications, which aims to detect and analyze abnormal behaviors and risky situations in order to prevent accidents. Various deep learning models have been previously developed for this purpose, such as CNN, RNN, and Vision Transformer, each one has its strengths and weaknesses based on the quality of input data. This paper proposes a novel approach based on the texture characteristics of input frames. In order to enrich the input data of the vision transformer model, and enhance feature extraction for the detection of anomaly, we combine the original image with its texture extracted using Local Binary Pattern(LBP), and fed them into a fine-tuned pre-trained Vision Transformer, enabling the automatic classification of video frames into abnormal and normal categories. The results demonstrate the effectiveness of our approach in identifying risky situations in video sequences.

Keywords: Anomaly Detection, Deep Learning, Vision Transformer, LBP

Introduction

The detection of risky road situations in video sequences or from surveillance cameras presents a considerable challenge, due to the inherent ambiguity of "risky situation". Various factors contribute to the complexity and diversity of visual features, including the quality of video sequences, illumination, occlusion, shadows, and moving cameras. An event is often considered as anomaly, if it happens rarely or suddenly (Popoola & Wang, 2012; Sabokrou *et al.*, 2018).

Anomaly detection involves identifying frames within a sequence that display events that substantially differ from normal situations, such as stampedes or traffic accidents (Pang *et al.*, 2020).

The detection of risky situations in public spaces, or in road traffic, is a vast field of research, and various approaches have been proposed to address this challenge. However, they have some limitations, particularly the reliance on labeled datasets containing normal events (Pang *et al.*, 2020). This dependency restricts their applicability because it requires human intervention to continuously retrain the system.

Recent deep learning algorithms have shown significant success in the anomaly detection field, They may be systematically grouped into four main classes: Error reconstruction algorithms (Zhao *et al.*, 2024), scoring (Liu *et al.*, 2018), algorithms based on the prediction of future frames, and classifier-based detection (Sabokrou *et al.*, 2018).

Approaches Based on Reconstruction Error

Multiple approaches employ a reconstruction error algorithm for anomaly detection. It uses the assumption that the normal samples show smaller error, while abnormal samples demonstrate significantly elevated error values (Zhao *et al.*, 2024).

Deep learning methods generally train an autoencoder to reconstruct normal events with a small similarity error. These approaches use an autoencoder deep learning model to accurately reconstruct a normal frame with minimal error.

However, the detection of abnormal events is not always guaranteed because the reconstruction error value for abnormal frames is not consistently higher (Liu *et al.*, 2018).

To overcome this limitation, the authors of (Hasan *et al.*, 2016), proposed a technique based on standard spatiotemporal local characteristics, for training autoencoders to recognize normal patterns.

The authors of paper (Wang & Yang, 2022) present a Convolutional Recurrent AutoEncoder (CR-AE), that combines a Convolutional LSTM network with a Convolutional AutoEncoder. They extract the output features from each Conv-LSTM layer's hidden state.

Thereafter, the input and testing video clips with higher reconstruction errors that were identified as anomalies were reconstructed using a convolutional decoder.

In order to perform anomaly detection, the paper (Chirikiri & Seo, 2024) proposes a method that introduces an autoencoder during preprocessing and uses the reconstruction error result as an additional input for the detector, and uses Grad-CAM (Selvaraju *et al.*, 2020) to focus on regions containing anomalous objects.

The approach in Chen *et al.* (2022) presents an anomaly detection system based on pre-trained transformer networks for feature reconstruction. However, its computational inefficiency limits real-time deployment. To enhance Vision Transformer (ViT) performance, the work in Mishra *et al.* (2021) introduces a reconstruction-based model incorporating patch embedding for improved anomaly detection and localization.

Score-Based Approaches

The main idea of score-based approaches (Liu *et al.*, 2018; Xie *et al.*, 2023) is to compute a score to identify whether a video frame contains anomalies or not.

Sultani *et al.* (2018) present an effective anomaly detection approach that uses both abnormal and normal videos to train their model. Therefore, to reduce the complexity of labeling abnormal data during training, they used weakly labeled training videos.

In the proposed approach, the authors develop a ranking model for anomaly detection that automatically generates high anomaly scores for abnormal videos, where videos are considered as instances in Multiple Instance Learning (MIL) and normal videos as bags. MIL is a deep learning mechanism that regroups training data into a bag, which contains a collection of instances.

Pang *et al.* (2020) address the problem of labelling data by learning anomaly scores, without explicitly labeling the video frames. To achieve this, they introduce an approach utilizing self-trained deep learning for ordinal regression to identify anomalies in video frames. It receives a set of unlabeled videos and then conducts an initial detection step to produce a collection of abnormal and normal data. These collections serve as a training dataset for the ResNet-50 model (He *et al.*, 2016) and an end-to-end fully connected network, where ResNet50 is a pre-trained model to extract appearance-based characteristics from images. The network is composed of a hidden layer of 100 units and an output layer with a linear unit.

Li *et al.* (2022) introduce a transformer framework based on Multi-Scale Learning (MSL) to estimate anomaly scores at the snippet level and compute anomaly probabilities at the video level within a weakly supervised learning setting for video anomaly detection.

Approaches Based on the Prediction of Future Frames

This approach learns a model capable of predicting future frames of the video sequence, based on the assumption that a normal frame is predictable.

The authors in Tang *et al.* (2020) propose an end-to-end network to predict future frames and compute reconstruction error. The prediction of future frames leads to higher reconstruction errors for abnormal events, In order to improve the quality of prediction of future frames from normal data, the proposed model uses two connected U-Net blocks in the generator for reconstruction of the output frames generated from the former block.

And the authors of Jin *et al.* (2022) apply a transformer-based approach for anomaly detection in aerial videos, by employing a transformer encoder to learn the representations of features from the video sequence, followed by a decoder that predicts the subsequent frame.

Classifiers-Based Approaches

The work of Sabokrou *et al.* (2018) formulated the detection of anomaly as a classification task and introduced a method that identifies anomalies in videos by examining the deep network layers' outputs. They have used temporal information in Fully Convolutional Neural Networks (FCNN). The FCN integrates a new convolutional layer that trains the kernels on the training video, into a pre-trained CNN using an AlexNet model (Krizhevsky *et al.*, 2017). The network is defined to perform two key tasks: Learning feature representation and identifying anomalies. This approach has yielded a good accuracy rate, but it still contains several limitations. It generates false positives in crowd scenes and when pedestrians are walking in various directions.

Vision Transformer (ViT) (Dosovitskiy *et al.*, 2021) has recently become a novel architecture in the field of computer vision. which was firstl applied to analyze and process Natural Language(NLP) (Vaswani *et al.*, 2017). Moreover, based on its success, the ViT has been adopted in various Computer Vision applications such as image classification (Chen *et al.*, 2022) and object detection (Carion *et al.*, 2020).

In order to enrich the input data of vision transformers used in Berroukham *et al.* (2023) and improve the anomaly detection quality, this study proposes a novel approach that combines the original frame with its texture and feeds them to Vision Transformer (ViT) models for robust anomaly detection. The integration leverages the deep learning capabilities of Vision Transformers and the texture analysis strength of LBP, offering enhanced detection accuracy and performance.

The empirical results demonstrate that the integration of the original frame and its texture outperforms traditional deep learning methods.

This study is structured as follows: We present in section two an overview of the vision transformer and the Local binary pattern algorithm, section trois presents our proposed approach, we give the discussion of the obtained results in the fourth section and finally, we present the conclusion.

Vision Transformer Model Based on LBP Algorithm

Previous research has used RNN and LSTM (Sharma *et al.*, 2021) to detect and localize anomalies. However, these models have multiple limitations due to their sequential processing of data, leading to exploding or vanishing gradients when there are long-term dependencies between data. As a result, the Transformer model has largely supplanted LSTM due to its better performance in sequence-to-sequence tasks (Karita *et al.*, 2019).

Concepts of Vision Transformers

The theoretical foundations of ViT's architecture (Dosovitskiy *et al.*, 2021) are based on the concept of attention mechanisms, which is originally used in Natural Language Processing (NLP) (Vaswani eta al., 2017). ViT uses self-attention mechanisms at its core to process visual data. ViT dynamically analyzes the relevance of different image areas during the prediction task. Which means that ViT is excellent at capturing long-range dependencies, allowing it to consider the relationships and context between distinct regions in an image at the same time.

The mechanism of Self-attention enables ViT to capture the relationships within an image by dynamically assigning different levels of importance to various regions during processing. This flexibility is essential to understand the context and long-range dependencies in visual data.

Multi-head attention extends this concept by allowing ViT to perform attention operations multiple times in parallel, each one concentrating on different parts of the input images, where each head focuses on a specific region. This parallelization improves the model to capture multiple types of information and makes it easier to understand the different features of an image. This mechanism makes ViT able to successfully distinguish pertinent information, model dependencies, and hierarchical features. Thus, a powerful model for image understanding and analysis in diverse applications and an excellent tool for complex computer vision tasks.

Architecture of Vision Transformers

Figure (1) illustrates the original architecture of Vision Transformers (ViT) (Dosovitskiy *et al.*, 2021), which marks a revolutionary change from the traditional Convolutional Neural Networks (CNNs). ViT begins by splitting an image with a resolution of 224×224 into patches with fixed sizes 16×16, a total of 196 patches are created from one image, then it flattens the image patches, to create linear embeddings. These patch embeddings are combined with positional encodings to provide the localization of each patch in the image, which is considered as the input of the encoder, as you can see in the Figure (1). The encoder block is a series of Transformer encoder layers (typically 12), each one is composed of Multi-Head Self-Attention (MSA), feed forward neural networks, normalization block and residual connections.

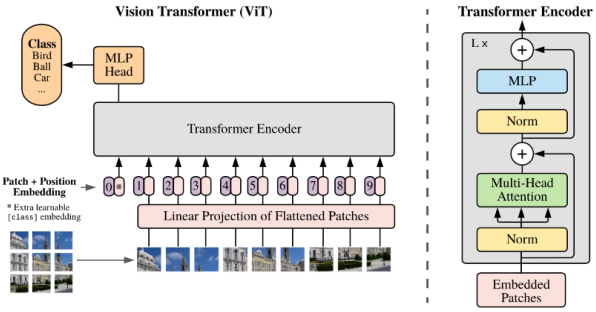

Fig. 1: Original architecture of ViT

Figure (2) illustrates the internal elements of each component of the encoder block in detail. The Self-Attention mechanisms enable ViT to weigh the relevance of different patches in the image dynamically, capturing both local and global relationships. The Multi-Head Attention mechanism parallelizes this process, to process various input data simultaneously, it has three inputs which are queries, keys, and values. Skip connections and layer normalization contribute to stable training, The skip connection is used at different places in the Transformer encoder, these connections are mainly to improve the flow of information to avoid vanishing gradients. Importantly, the output of the final Transformer layer is often aggregated and MLP heads are attached for specific tasks, such as image classification.

Figure (2) illustrates the internal elements of each component of the encoder block in detail. The Self-Attention mechanisms enable ViT to weigh the relevance of different patches in the image dynamically, capturing both local and global relationships. The Multi-Head Attention mechanism parallelizes this process, to process various input data simultaneously, it has three inputs which are queries, keys, and values. Skip connections and layer normalization contribute to stable training, The skip connection is used at different places in the Transformer encoder, these connections are mainly to improve the flow of information to avoid vanishing gradients. Importantly, the output of the final Transformer layer is often aggregated and MLP heads are attached for specific tasks, such as image classification.

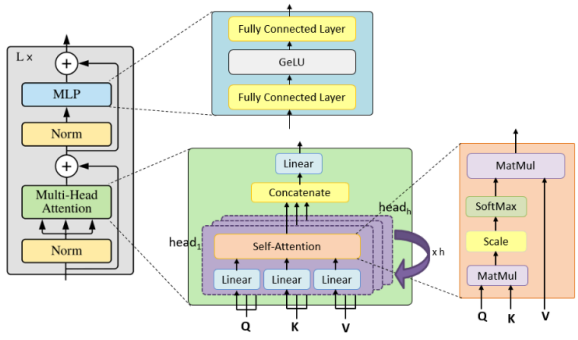

Fig. 2: The Vision Transformer encoder block

This architectural design empowers ViT to adapt dynamically to different image contents and captures complex spatial relationships, making it highly effective in a wide range of computer vision applications.

Vision Transformer encoder takes a one-dimensional dataset as input, composed of embedded image patches. Then, it reshapes the image $X \in R^{H \times W \times C}$ into a sequence of flattened 2D patches denoted $X_p \in R^{N \times (P \cdot P \cdot C)}$, where (H, W) represents the resolution of the frame, C represents the number of channels, (P, P) the resolution of each patch and the number of patches is denoted by N = HW/P^2. The ViT maintains a fixed patch embedding size D within all layers.

Eq. (1), presents the output of the initial patch, where Xclass is the class token, Xp the patch embedding, and Epos the position embedding.

There is a Multi-Head Attention layer (MSA) for every layer, the addition term at the end of Eq. (2) is equivalent to a residual connection that adds the input to the output of the MSA block.

The MLP block can be represented by Eq. (3), where MLP wraps a Normalization Layer (LN) in Eq.(4):

$$Z_0 = \left[X_{class} : X_p^1 E;\ X_p^2 E; \ldots;\ X_p^N E\right] + E_{pos} E \in \mathcal{R}^{\left(p^2 . C\right) . D},\ E_{pos} \in \mathcal{R}^{(N+1) . D} \quad (1)$$

$$Z_t^{'} = MSA\left(LA\left(Z_{t-1}\right)\right) + Z_{t-1} | t = 1 \ldots L \quad (2)$$

$$Z_t = MLP\left(LA\left(Z_t^{'}\right)\right) + Z_t^{'} | t = 1 \ldots L \quad (3)$$

$$y = LN\left(Z_L^0\right) \quad (4)$$

Local Binary Pattern(LBP)

LBP as proposed in Ahonen *et al.* (2006); Ojala *et al.* (2002), is a simple algorithm that encodes the image pixels to extract texture. LBP calculates the labels of the pixels by comparing the value of the central pixel with the value of each neighboring pixel as shown in Figure (3). The formula below shows how to calculate the LBP of a pixel $I\left(x_c, y_c\right)$:

$$I_{LBP}\left(x_c, y_c\right) = \sum_{p=-3}^{p=3} s\left(g_p - g_c\right) 2^p \left(g_p \vDash g_c\right), 0 \leq I_{LBP}\left(x_c, y_c\right) \leq 2^8 - 1 \quad (5)$$

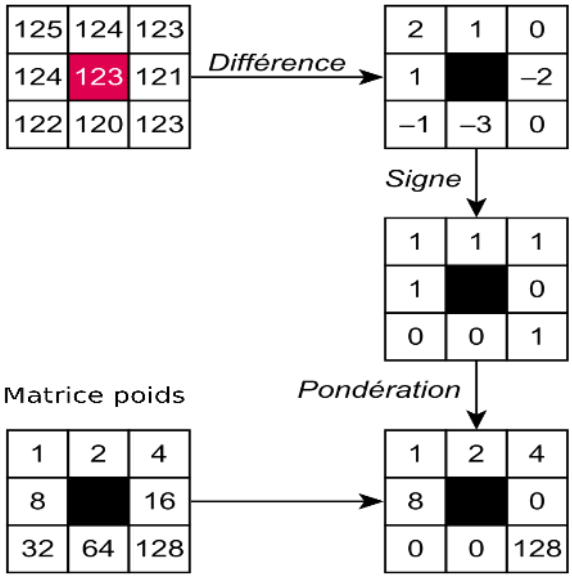

Fig. 3: Calculation of LPB for an image pixel

Where, g_c represents the value of the central pixel $(x_c, y_c)$and g_c corresponds to the values of the P neighboring pixels. The function $s\left(x\right)$ is defined as follows:

$$s\left(x\right) = \begin{cases} 1x \geq 0 \\ 0x < 0 \end{cases} \quad (6)$$

The advantage of the LBP operator in real-world applications is its robustness to the changes of intensity values caused by lighting variations.

Other important characteristics of LBP are its ability to extract discriminative features in a simple and easy manner and also the simplicity of its implementation (Ahonen *et al.*, 2006), which allows for real-time image analysis. Figure (4) shows an image texture calculated using LBP.

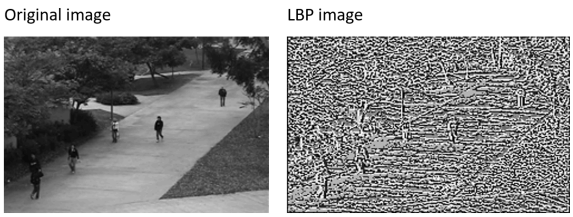

Fig. 4: LBP of an image

Materials and Methods

Proposed Model

Concepts of Vision Transformers

We adapt the Vit model, developed in Chen *et al.* (2022), and pre-train it to detect anomalous events in video sequences. The flowchart of our fine-tuned model is illustrated in Figure (5).

In Figure (5), the image resulting from integration of the video frame with its LBP $x \in R^{H\times W\times C}$ is partitioned into a series of one-dimensional patches, which are encoded and passed as input to the transformer encoder. (H, W) denotes the original frame resolution, C is the number of channels and P = 16 represents the resolution of each patch.

Each flattened patch x in the sequence is mapped into a latent vector space with a hidden size of D = 768. A learnable embedding class ($z_0^0 = X_{class}$) is then added to the sequence of embedded patches.

To maintain the order of spatial information, positional embeddings Epocs are added to the patch embeddings before being fed them as input into the transformer encoder (Eq. 1).

The output model Z_0^L is normalized to obtain the frame's feature representation y. This representation is then fed into a classifier to predict its corresponding class.

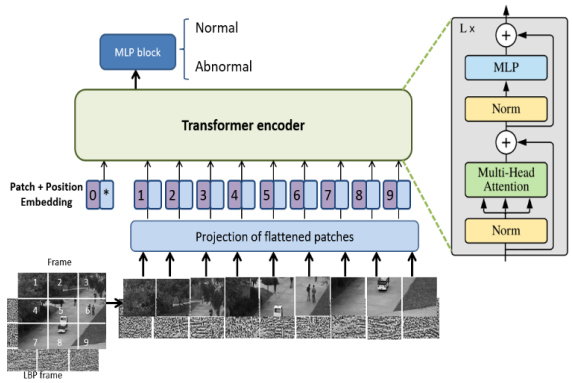

Fig. 5: Flowchart of the proposed model

Experiments Setup

Dataset

To evaluate our approach, we employ the UCSD anomaly detection dataset (Mahadevan *et al.*, 2010), which is regrouped into two distinct subsets (Peds1 and Peds2). The data is split into separate training and testing sets, each test clip contains at least one anomalous event, which primarily involves either irregular pedestrian movements or non-pedestrian objects moving across pedestrian areas (Figure 6).

The dataset contains various anomalies such as skaters, cyclists, small vehicles, and pedestrians leaving designated walkways. The two subsets differ in their frame dimensions and camera angles.

The frames of the UCSD dataset contain the original frames and their masks delineating the object presenting the anomaly. Our approach consists of classifying the frames as normal and abnormal, so we chose to label the dataset frames as normal with the label 1 and abnormal with 0. In this study, we resize all frames to 224×224.

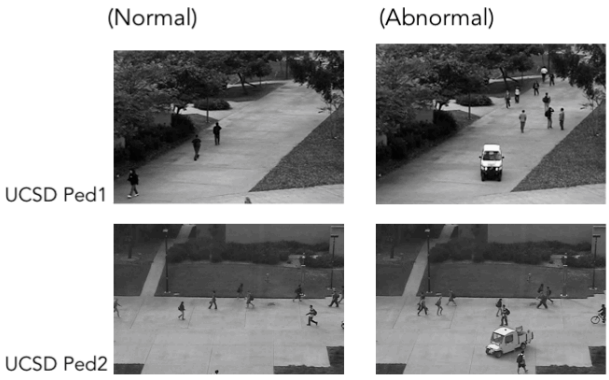

Fig. 6: Sample frames from the UCSD dataset

Evaluation Metrics

We assessed the performance of our approach using standard classification metrics such as the recall, precision, accuracy, F1-score, and confusion matrix. The formula of each metric is illustrated in Table (1). Where TP represents true positive, TN true negative, FP false positive, and FN false negative.

Table 1: Evaluation metrics

Criterion	Definition
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Recall	$\frac{TP}{TP+FN}$
Precision	$\frac{TP}{TP+FP}$
F1-Score	$\frac{2xTP}{2xTP+FP+FN}$

Hyper-Parameter Setting

We have trained our model on the UCSD Anomaly Detection dataset in five epochs, the batch size is 8, The optimization of the model is assured using Adam optimizer (Kingma, 2018), the learning rate is fixed to 2×10^{-5} and the encoder has L = 12 layers and 30% of each class in the training set is for validation.

The Vit model was fine-tuned and their weights were initialized using the pre-trained model based on the ImageNet-21K dataset (Vaswani *et al.*, 2017). To train, validate, and test our pretrained model, we use the PyTorch framework on an NVIDIA Tesla T4 GPU.

Experimentation

The loss and accuracy curves obtained during the training and validation phases are presented in Figure (7). The model started with a loss of 0.75 in the first epoch and reduced it to attain 0.02. The model achieves a good value of 0.993 for accuracy.

Figure (2) presents the architecture of our Vision Transformer Encoder, which consists of two components: A self-attention block and a multilayer perceptron block, We generate attention maps from the ViT model to visualize the detected spatial markers.

DOI: 10.3844/jcssp.2025.1613.1620

As illustrated in Figure (8), we display samples of several frames combined with LBP, along with their corresponding attention maps. These attention maps reveal that the ViT using frame and its LBP, is highly effective at quickly identifying the key elements in the frame sequence, which correspond to the anomalous objects in our work.

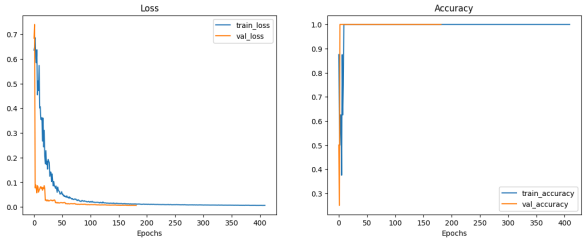

Fig. 7: Loss and accuracy curve for both the training and validation steps

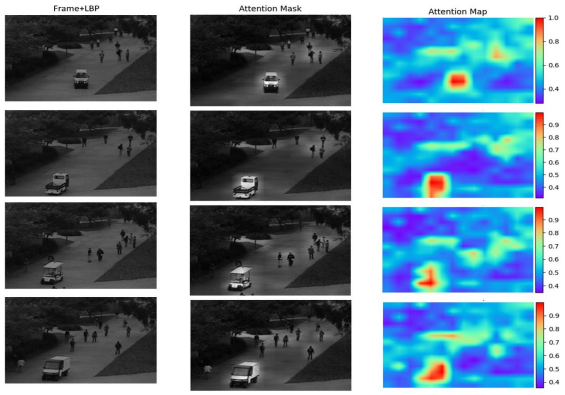

Fig. 8: Original frame+LBP image, attention mask, and attention map

Results and Discussion

Based on the confusion matrix as shown in Figure (9), the vit (Berroukham *et al.*, 2023) has correctly classified 59% of the abnormal frames and 38% of the normal frames, but it has misclassified 3% of normal frames, The model achieves an accuracy of 97%. By comparison, the proposed model based on Vit and LBP, showed more effectiveness in classifying frames as normal and abnormal, than the Vit model, as it achieved an accuracy of 99%.

Table (2) shows the quantitative results of the proposed and other models, The suggested method significantly reduces false positives, achieving a high precision of 0.97, compared to the ViT model's precision of 0.92. In terms of F1-score and accuracy, the proposed model demonstrates superior performance in anomaly detection. This enhanced performance is attributed to the improvement of the input data by incorporating texture features into the original frame.

The anomaly detection results are shown in Figure (10), where the model successfully identifies whether a frame contains an anomaly or not. Various frames exhibiting anomalies, such as motorcycles, pedestrian skis, and cars, have been detected.

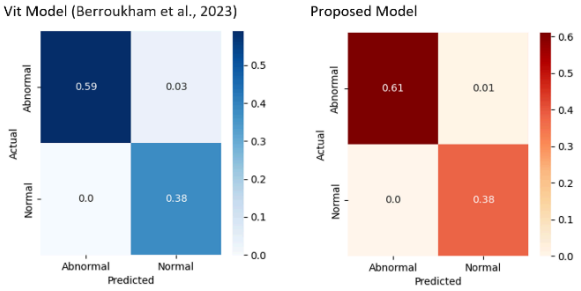

Fig. 9: Confusion matrix of Vit and proposed models

Table 2: Evaluation metrics of VIT and VIT-LBP models

Metrics	Spatiotemporal Autoencoder with Dynamic Map (Feng *et al.*, 2022)	Vit model (Berroukham *et al.*, 2023)	Proposed model
Precision	0.87	0.92	0.97
Recall	0.97	1	1
F1-score	0.92	0.96	0.98
Accuracy		0.97	0.99

Fig. 10: Results of our pretrained anomaly detection model

Conclusion

This study introduces an approach based on a fine-tuned Vision transformer model to classify video frames as either normal or abnormal. As input, the Vision Transformer model uses the combination of a frame and its texture extracted using LBP which provides a powerful technique for anomaly detection. The experimental results show that the proposed method outperforms the deep learning-based approaches in various anomaly detection scenarios.

Acknowledgment

We want to express our thanks and gratitude to all those who helped us throughout this study.

Funding Information

The authors have not received any financial support or funding to report.

Author's Contributions

All authors equally contributed to this study.

Ethics

The authors confirm that this manuscript has not been published elsewhere and that no ethical issues are involved.

References

Ahonen, T., Hadid, A., & Pietikainen, M. (2006). Face Description with Local Binary Patterns: Application to Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(12), 2037-2041. https://doi.org/10.1109/tpami.2006.244

Berroukham, A., Housni, K., & Lahraichi, M. (2023). Fine-Tuning Pre-trained Vision Transformer Model for Anomaly Detection in Video Sequences. *Proceedings of the 6th International Conference on Big Data and Internet of Things*, 279-289. https://doi.org/10.1007/978-3-031-28387-1_24

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-End Object Detection with Transformers. *Computer Vision -- ECCV 2020*, 213-229. https://doi.org/10.1007/978-3-030-58452-8_13

Chen, H., Li, C., Wang, G., Li, X., Rahaman, M. M., Sun, H., & Hu, W. (2022). GasHis-Transformer: A multi-scale visual transformer approach for gastric histopathological image detection. *Pattern Recognition*, *130*, 108827. https://doi.org/10.1016/j.patcog.2022.108827

Chen, L., You, Z., Zhang, N., Xi, J., & Le, X. (2022). UTRAD: Anomaly detection and localization with U-Transformer. *Neural Networks*, *147*, 53-62. https://doi.org/10.1016/j.neunet.2021.12.008

Chirikiri, Y., & Seo, M. (2024). Anomaly Detection Using Reconstruction Error and GradCAM. *2024 IEEE 13th Global Conference on Consumer Electronics (GCCE)*, 793-795. https://doi.org/10.1109/gcce62371.2024.10760385

Cordonnier, J.-B., Mahendran, A., Dosovitskiy, A., Weissenborn, D., Uszkoreit, J., & Unterthiner, T. (2021). Differentiable Patch Selection for Image Recognition. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr46437.2021.00238

Feng, J., Wang, D., & Zhang, L. (2022). Crowd Anomaly Detection via Spatial Constraints and Meaningful Perturbation. *ISPRS International Journal of Geo-Information*, *11*(3), 205. https://doi.org/10.3390/ijgi11030205

Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A. K., & Davis, L. S. (2016). Learning Temporal Regularity in Video Sequences. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 733-742. https://doi.org/10.1109/cvpr.2016.86

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778). https://doi.org/10.1109/cvpr.2016.90

Jin, P., Mou, L., Xia, G.-S., & Zhu, X. X. (2022). Anomaly Detection in Aerial Videos With Transformers. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1-13. https://doi.org/10.1109/tgrs.2022.3198130

Karita, S., Chen, N., Hayashi, T., Hori, T., Inaguma, H., Jiang, Z., Someki, M., Soplin, N. E. Y., Yamamoto, R., Wang, X., Watanabe, S., Yoshimura, T., & Zhang, W. (2019). A Comparative Study on Transformer vs RNN in Speech Applications. *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 449-456. https://doi.org/10.1109/asru46091.2019.9003750

Kingma, D. P. (2018). Adam: A Method for Stochastic Optimization. *ArXiv:1412.6980*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet Classification With Deep Convolutional Neural Networks. *Communications of the ACM*, *60*(6), 84-90. https://doi.org/10.1145/3065386

Li, S., Liu, F., & Jiao, L. (2022). Self-Training Multi-Sequence Learning with Transformer for Weakly Supervised Video Anomaly Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(2), 1395-1403. https://doi.org/10.1609/aaai.v36i2.20028

Liu, W., Luo, W., Lian, D., & Gao, S. (2018). Future Frame Prediction for Anomaly Detection - A New Baseline. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6536-6545. https://doi.org/10.1109/cvpr.2018.00684

Mahadevan, V., Li, W., Bhalodia, V., & Vasconcelos, N. (2010). Anomaly detection in crowded scenes. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1975-1981. https://doi.org/10.1109/cvpr.2010.5539872

Mishra, P., Verk, R., Fornasier, D., Piciarelli, C., & Foresti, G. L. (2021). VT-ADL: A Vision Transformer Network for Image Anomaly Detection and Localization. *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*, 1-6. https://doi.org/10.1109/isie45552.2021.9576231

Ojala, T., Pietikainen, M., & Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(7), 971-987. https://doi.org/10.1109/tpami.2002.1017623

Pang, G., Yan, C., Shen, C., van den Hengel, A., & Bai, X. (2020). Self-Trained Deep Ordinal Regression for End-to-End Video Anomaly Detection. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12170-12179. https://doi.org/10.1109/cvpr42600.2020.01219

Popoola, O. P., & Wang, K. (2012). Video-Based Abnormal Human Behavior Recognition-A Review. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *42*(6), 865-878. https://doi.org/10.1109/tsmcc.2011.2178594

Sabokrou, M., Fayyaz, M., Fathy, M., Moayed, Zahra., & Klette, R. (2018). Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes. *Computer Vision and Image Understanding*, *172*, 88-97. https://doi.org/10.1016/j.cviu.2018.02.006

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, *128*(2), 336-359. https://doi.org/10.1007/s11263-019-01228-7

Sharma, S., Sudharsan, B., Naraharisetti, S., Trehan, V., & Jayavel, K. (2021). A fully integrated violence detection system using CNN and LSTM. *International Journal of Electrical and Computer Engineering (IJECE)*, *11*(4), 3374. https://doi.org/10.11591/ijece.v11i4.pp3374-3380

Sultani, W., Chen, C., & Shah, M. (2018). Real-World Anomaly Detection in Surveillance Videos. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT. https://doi.org/10.1109/cvpr.2018.00678

Tang, Y., Zhao, L., Zhang, S., Gong, C., Li, G., & Yang, J. (2020). Integrating prediction and reconstruction for anomaly detection. *Pattern Recognition Letters*, *129*, 123-130. https://doi.org/10.1016/j.patrec.2019.11.024

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all You Need. *Advances in Neural Information Processing Systems*, *30*, 6000-6010.

Wang, B., & Yang, C. (2022). Video Anomaly Detection Based on Convolutional Recurrent AutoEncoder. *Sensors*, *22*(12), 4647. https://doi.org/10.3390/s22124647

Xie, X., Li, Z., Huang, Y., & Wu, D. (2023). D.: A weakly supervised anomaly detection method based on deep anomaly scoring network. *Signal, Image and Video Processing*, *17*(8), 3903-3911. https://doi.org/10.1007/s11760-023-02619-7

Zhao, L., Chai, Y., Zhang, Q., & Karimi, H. R. (2024). Self-supervised anomaly detection based on foreground enhancement and autoencoder reconstruction. *Signal, Image and Video Processing*, *18*(1), 343-350. https://doi.org/10.1007/s11760-023-02756-z