Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

Crowd Localization and Anomaly Detection Using Video Anomaly Scoring Network

[1]Sivalingan H and [2]Dr. N. Anandakrishnan
[1]Research Scholar, Providence College for Women, Coonoor.
[2]Head & Assistant Professor, Department of Computer Science & Applications,
Providence College for Women, Coonoor.
[1]nescnr@gmail.com, [2]anandpjn@gmail.com

Article Info
Page Number: 825-837
Publication Issue:
Vol. 72 No. 1 (2023)

Abstract
Anomaly detection has become core topic in deep learning and computer vision. With the increase in surveillance cameras, it is important foridentifying the difference between the available normal data and the unusual happening in the Video Surveillance. Even though many methods are implemented by using LSTM, convolutional network and vision transformers they have high computational time, low resolution and poor anomaly accuracy.So, this work proposesa new technique named VST (Video Swin Transformer) -Anomaly Scoring Network (VASN) in high image resolution for predicting every abnormal action in each frame.University of Minnesota (UMN) dataset are pretrained in this model and then semi- supervised basedanomaly scoring network is added to find the anomaly behaviour in the video clips. The localization of crowd for target detection and then features of the target are extracted to detect the anomaly score, when comparing the other existing techniques, the work shows better performance with the score value 98.2 in the anomaly detection.

Article History
Article Received: *15 October 2022*
Revised: *24 November 2022*
Accepted: *18 December 2022*

Keywords: Video Swin Transformer, Anomaly Scoring Network, Video Surveillance

1. Introduction

Anomaly detection is the identification of rare events, items, or observations which are suspicious because they differ significantly from standard behaviours or patterns. There are three main classes of anomaly detection techniques: unsupervised, semi-supervised, and supervised. In this method semi-supervised anomaly detection algorithm is used, whichwork with a data set that is partially flagged. It will then build a classification algorithm on just that flagged subset of data, and use that model to predict the status of the remaining data. There are many applications in anomaly detection like network anomaly, online anomalydetection, systems log, Image anomaly detection. In this work, detecting anomaly in video surveillanceis the major goal where the anomalies are predicted by detecting the change in the frames by using one of the methods in deep learning.

Currently one of the best methods for image recognition in video surveillance is transformers which plays major role in computer vision.Itbeginsto demonstrate revolutionary performance improvements for a variety of computer vision tasks. This indicates that CV (computer vision) and NLP (natural language processing) modelling could potentially be unified under the Transformer architecture. This trend is highly beneficial for both fields: 1) it can facilitate joint modelling of visual and textual signals; 2) it can enable better sharing of modelling knowledge between the two fields, which would accelerate progress in both fields.

Transformer's general modelling capabilities come from two aspects. On one hand, Transformer can be seen as performing on a graph. The graph is fully connected, and the relationships between nodes are learned in a data-driven way. On the other hand, Transformer has good generality in establishing the relationships between graph nodes through a philosophy of validation: no matter how heterogeneous the nodes are, their relationships can be established by projecting them into an aligned space to calculate similarity.

This trend began with the introduction of Vision Transformer (ViT) [18,19], which globally models spatial relationships on non-overlapping image patches with the standard Transformer encoder. The great success of ViT on images has led to investigation of Transformer-based architectures for video-based recognition tasks in the initial attempts at Transformer-based video recognition, a factorization approach is also employed, via a factorized encoder [20] or factorized self-attention. This has been shown to greatly reduce model size without a substantial drop in performance.

The implementation of the video Swin transformer[1]this approach is done by adaption of spatiotemporal of Swin Transformer [17], which was recently introduced as a general-purpose vision backbone for image understanding. Swin Transformeris incorporated inductive bias for spatial locality, as well as for hierarchy and translation invariance, which strictly follows the hierarchical structure of the original Swin Transformer, but extends the scope of local attention computation from only the spatial domain to the spatiotemporal domain. As the local attention is computed on non-overlapping windows, the shifted window mechanism of the original Swin Transformer is also reformulated to process spatiotemporal input.

Abnormal crowd behaviour [15]focused on panic and escape behaviour detection that may appear because of violent events and natural disasters. First, optical flow vectors are computed to generate a motion information image (MII) for each frame, and then MIIs are used to train a convolutional neural network (CNN) for abnormal crowd event detection. The proposed MII is a new formulation that provides a visual appearance of crowd motion. The proposed MIIs make the discrimination between normal and abnormal behaviours easier. The MII is mainly based on the optical flow magnitude, and angle difference computed between the optical flow vectors in consecutive frames. abnormal crowd behaviour detection based on distribution of magnitude of optical flow [16](DMOF),global event influence model (GEIM) he proposed GEI integrates the crowd motions and social psychology attributes to improve the description of crowds. For this, low-level motion features are abstracted as crowd attributes of scale, velocity, and disorder. Then, the detailed definitions and mathematical expressions of GEI are presented through calculating the convolution of rise factor and decay factor[17]. In these existing methods they are detecting anomaly and localization in crowd at parks, malls by using CNN, LSTM and encoder decoder-based models. even though its detecting abnormal events the LSTM model cannot perform multi head attenuation process so the computational time is high compared to our method and in the transformer model it has backbone architecture CNN.

The main contribution of ourproposed VASN is listed:

- Implementation of video Swin transformer for improving the image resolution in each frame in the video where the spatial and temporal computation are self-attenuated.
- To achieve the speed of accuracy trade on image classification for action recognition compared to the other methods.
- To improve the anomaly detection by reducing the false detection and to improve the localization of crowd in surveillance.

2. Literature Survey

Yuan et.al.[2] proposeda prediction-based video anomaly detection approach named TransAnomaly. They combined the U-Net and the Video Vision Transformer (ViViT) to capture richer temporal information and more global contexts. The modified the ViViT to make it capable of video prediction for the anomaly detection performance and also calculated regularity scores with sliding windows then it evaluated the impact of different window sizes and strides. The anomaly localization is by tracking the location of patches with lower regularity scores.

Lee *et.al*[3] the goal is to find abnormalities in the walking pattern of the pedestrians propose a modified Time-Series Vision Transformer (TSViT), a method for anomaly detection in video, specifically for tailing detection with a small datasetoutperforms popular CNN-based architectures, as the CNN architectures tend to overfit with a small dataset of time-series images. The performance of CNN-based architecture gradually drops, as the network depth is increased, to increase its capacity. On the other hand, a decreasing number of heads in Vision Transformer architecture shows good performance on time-series images, and the performance is further increased as the input resolution of the images is increased

Wang *et. al*. [4] The new abnormality indicator is derived from the hidden Markov model which learns the histograms of optical flow orientations of the observed video frames. This indicator measures the similarity between the observed video frame and existing normal frames. The proposed method is evaluated and validated on several video surveillance datasets.

Wan *et. al.*[21] proposed a long video event retrieval algorithm based on super frame segmentation. By detecting the motion amplitude of the long video, a large number of redundant frames can be effectively removed from the long video, thereby reducing the number of frames that need to be calculated subsequently. Then, by using a super frame segmentation algorithm based on feature fusion, the remaining long video is divided into several Segments of Interest (SOIs) which include the video events.

Huang et. al. [5] proposed a temporal-aware contrastive network (TAC-Net) to address the above problems of anomaly detection for intelligence video surveillance. TAC-Net is an unsupervised method that utilizes deep contrastive self-supervised learning to capture the high-level semantic features and tackles anomaly detection with multiple self-supervised tasks. During inference phase, the multiple task losses and contrastive similarity are utilized to calculate the anomaly score. Experimental results show that this method superior to state-of-the-art approaches on three benchmarks, which demonstrates the validity and advancement of TAC-Net

Gao *et. al*. [6]proposed a Dilated Convolutional Swin Transformer (DCST) for congested crowd scenes. Specifically, a window-based vision transformer is introduced into the crowd localization task, which effectively improves the capacity of representation learning. Then, the well-designed dilated convolutional module is inserted into some different stages of the transformer to enhance the large-range contextual information. Extensive experiments evidence the effectiveness of the proposed methods and achieve the state-of-the-art performance on five popular datasets.

Doshi and Yilmaz [7] proposed an online anomaly detection method in surveillance videos with asymptotic bounds on the false alarm rate, which in turn provides a clear procedure for selecting a proper decision threshold that satisfies the desired false alarm rate. The algorithm consists of a multi-objective deep learning module along with a statistical anomaly detection module, and its effectiveness is demonstrated on several publicly available data sets where we outperform the state-of-the-art algorithms.

Nawaratne*et. al.*[9] proposed the Incremental Spatio-Temporal Learner (ISTL) to address challenges and limitations of anomaly detection and localisation for real-time video surveillance. ISTL is demonstrated and evaluated on accuracy, robustness, computational overhead as well as contextual indicators, using three benchmark datasets. Results of those experiments validate and confirmed its suitability for real-time video surveillance.

Nasaruddin*et. al*. [10] described a method for learning anomaly behaviour in the video by finding an attention region from spatiotemporal information, in contrast to the full-frame learning and a robust background subtraction (BG) for extracting motion, indicated the location of attention regions is employed. The resulted regions are finally fed into a three-dimensional Convolutional Neural Network (3D CNN). Specifically, by taking advantage of C3D (Convolution 3-dimensional), to completely exploit spatiotemporal relation, a deep convolution network is developed to distinguish normal and anomalous events.

Feng *et. al.*[8] introduced a two-stream approach that offers an autoencoder-based structure for fast and efficient detection to facilitate anomaly detection from surveillance video without labeled abnormal events and presented post hoc interpretability of feature map visualization to show the process of feature learning, revealing uncertain and ambiguous decision boundaries in the video sequence. Experimental results on Avenue, UCSD Ped2, and Subway datasets showed that this method detected abnormal events well and explain the internal logic of the model at the object level.

Duman andErdem[11] proposed a framework (OF-ConvAE-LSTM) to detect anomalies using Convolutional Autoencoder and Convolutional Long Short-Term Memory in an unsupervised manner. Besides the deep learning model, the feature extraction stage based on dense optical flow is applied in the framework to obtain the velocity and direction information of foreground objects. The experiments were carried out on three popular public datasets consisting of Avenue, UCSD Ped1, and UCSD Peds2. The experimental results hadshown that the complex distribution of the pattern of regular motion changed with high accuracy

Chen*et. al.*[12]proposed a framework based on bidirectional prediction, which predicted the same target frame by the forward and the backward prediction subnetworks, respectively. Then the loss function is constructed based on the real target frame and its bidirectional prediction frame and also proposed an anomaly score

estimation method based on the sliding window scheme which focuses on the foregrounds of the prediction error map. The comparison with the state-of-the-art shows that the proposed model outperforms most competing models on different video surveillance datasets.

3. Methodology

This section describes the proposed VASN work in detail. First, the overview of the proposed method is presented and then how the anomaly in video frames is detected and explained in detail.

3.1. Overview

The proposedVASN consists of video Swin Transformer and convolution network. The transformer encodes the input images with high resolution to predict the abnormal images in the video frames. The resulting features are then fed into an anomaly scoring network to detect the anomaly score value using deep anomaly detector and it is illustrated in Figure 1(Gao *et. al. (*2021))[6].

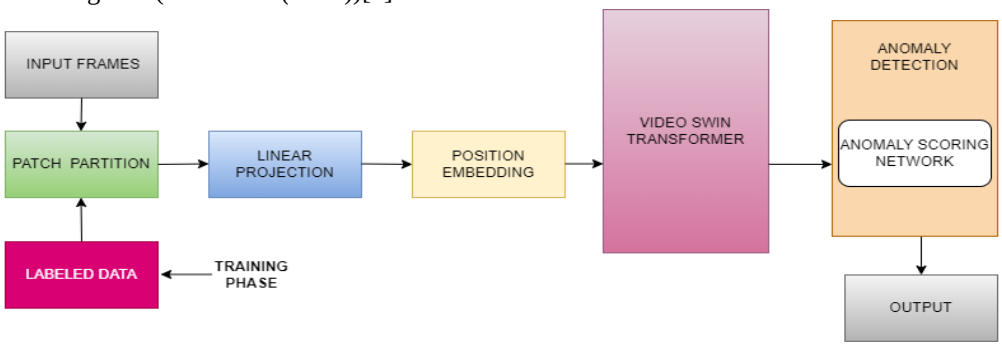

Figure 1: Overview of VASN.

3.2. Dataset

To detect the anomaly in the video surveillance in the method, we are selectingUMN which is publicly available dataset and it is composed of three scenesnamely, a lawn, interior, and plazawith a resolution of 240 × 320.The videos used in this work are recorded using static camera. The UMN dataset was used to evaluate the model's capacity with the frame-level criterion. All scenes are related to the escaping action of crowds.The footage starts with a medium density crowd of people acting out "normal" movement. After some time, the crowd members suddenly disperse in different directions as if panicked. In this dataset, the evacuation behaviours of crowds are assigned as abnormal.Using this small dataset in proposedVASN there will be increase in accuracy due to pretraining.Each frame's anomaly is detected and used for image matching, image categorization, and action recognition as mentioned in Figure 2 Wang*et. al* [22].

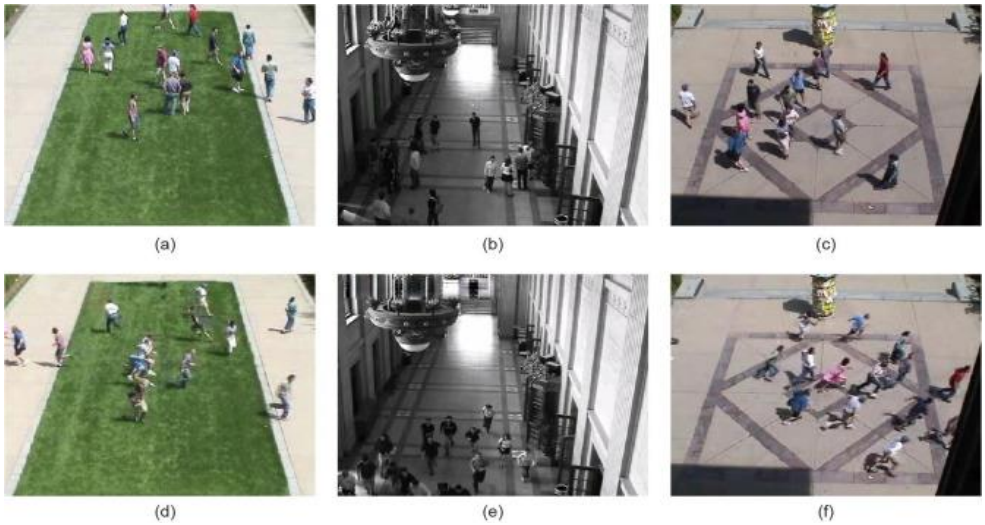

Figure.2. UMN dataset for three scenes. (a, d) show a scene on a lawn, (b, e) show an indoor scene, and (c, f) show a scene in a plaza. The evacuation behaviour of crowds (d–f) are assigned as abnormal.

3.3. Video Swin Transformer

The major component of the architecture is the Video Swin transformer block, which is built by replacing the multi-head self-attention (MSA) module in the standard Transformer layer with the 3D shifted window based multi-headself-attention module and keeping the other components unchanged. Specifically, a video transformer block consists of a 3D shifted window based MSA module followed by a feed-forward network, specifically a 2-layer MLP, with GELU non-linearity in between. Layer Normalization (LN) is applied before each MSA module and FFN, and a residual connection is applied after each modulefrom that of the preceding layer's self-attention module.

The input video in this method is considered as I which consists of sampled image frames $f = \{f_1, f_2 \ldots\ldots\}$from I,the pre-processed labelled data {LD} are given in the trainingphase and the video frames within the input video is denoted as {UD}.

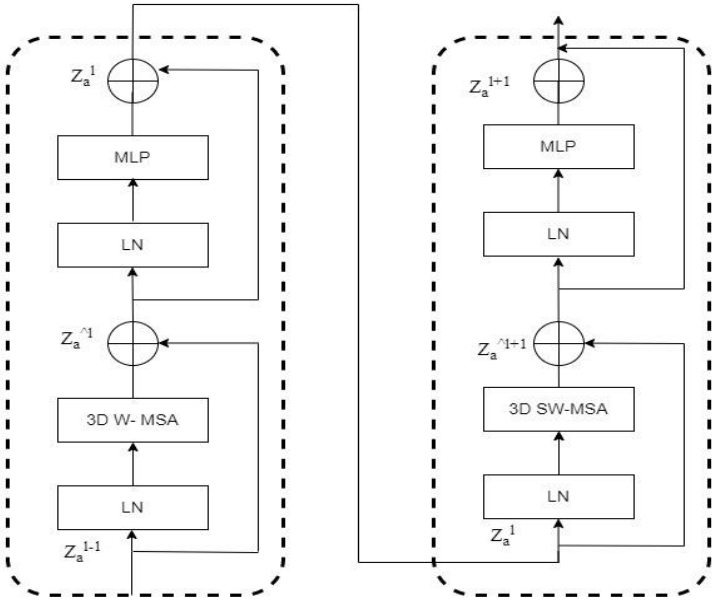

Figure 3:Video Swin Transformer blocks

i) Patch Embedding

In the video Swin transformer the 3D patch partitioning layer obtains frames=T2×H4×W4 3D tokens, with each patch/token consisting of a 96-dimensional. Givenan image $i_f \in R^{H\times W^2\times d^2}$withthe size of height H, width W and dimension d, it is reshaped into a sequence, with the linear projection. In our model the spatial and temporal are frames are self-attenuated with the featureschannel where the video is partitioned into frames and it is reshaped into a sequence, for l layer it has regular partition$\left(\frac{T}{P}, \frac{H}{P}, \frac{W}{M}\right)and(l+1)$ has shifted windows which are shifted by $\left(\frac{P}{2}, \frac{P}{2}, \frac{M}{2}\right)$ layer. Finally, thesequence of embedding vectors is fed into the TransformerEncoder.

Specifically, the operation of patch embeddings is formulatedas:

$$[Z_n = i_f class;\ i\,{}^1_{fp}EB; i^2_{fp}EB; i^3_{fp}EB; \ldots\ldots i^n_{fp}EB] + E_{PO}; \quad (1)$$

Where $i_f class$ denoted as the embedded patches Z_n and EB denotes the process of learnable embeddings $EB \in R^{PM^2\times d}$. $PM^2$is the no. of 3d token in relative position bias.

The multi-layer 3D-shifted windowthen considers different levels of spatial-temporal attention over these video patch embeddings. We add learnable positional embeddingE_{PO}.Finally, thesequence of embedding vectors is fed into the Transformer module.

ii) **3-D Shifted Based Msa Modules**

The computational formulas of the modules are formulated as(Liu*et. al.* (2021) [17]):

$\widehat{Z_a} = 3DW - MSA\left(LN(Z_a^{l-1})\right) + Z_a^{l-1} \qquad l = 1,2,3 \ldots n \qquad (2)$

$\widehat{Z_a}^{l+1} = 3DSW - MSA\left(LN(Z^l{}_a)\right) + Z_a^l \qquad l = 1,2,3 \ldots . n \qquad (3)$

$\hat{Z}_a$ and $\widehat{Z_a}^{l+1}$ are the output of the modules of 3DW-MSA and 3DSW-MSA for the l block respectively. The final encoded patches are fed into the anomaly detection network to find the abnormal movement in the video.

3.3. Anomaly Detection

The anomaly detection phase get input from the Swin transformer. In this phase the anomalywas detected with the help of the labeled anomaly set$\left(LD = \left\{\left(a_{n+1,}b_{n+1}\right), \ldots \left(a_{n+m,}b_{n+m}\right)\right\}\right)$.This labeled anomaly data was fed to the model at the time of training to make the detection process effective and easy. From the anomaly setLD, the anomaly cluster centers are identified using the K-means clustering algorithm. This clustering algorithm effectively identify the clusters in which the similar anomaly actions are grouped. The result of the clustering will be many clusters and noises. Here the noises are also treated as cluster centers because these noises are also labeled anomalies (Gao *et al*. (2021)). Now calculate the center of mass of these clusters as the cluster centers $C = \left\{c_{1,}c_{2,} \ldots c_{q,}\right\}$ via K-means clustering.

i. Distance Score Value:

After clustering the labeled data, then find out the similarity between the samples a_{ud}in the unlabeled data set $UD = \{a_1, a_2, \ldots a_n\}$ and cluster centers. For finding the similarity between the cluster centers and unlabeled data here Euclidean distance is utilized. The Euclidean distance between the sample a_{ud} and the cluster center C_P is computed according to the equation (3):

$$\lambda(a_{ud}, C_P) = \sum_{j=1}^{d}\left(a_{udj} - C_{Pj}\right)^2 \qquad (4)$$

Here d is the dimension of a_{ud}. Now the minimum distance between a_{ud} and cluster center was obtained by the following equation (4):

$$dt_{ud} = \min_{C_P} \lambda(a_{ud}, C_P) \qquad (5)$$

After calculating the Euclidean distance among the sample a_{ud} and its nearest cluster center C_P, the Distance Score Value (DSV) is computed for each sample a_{ud} in the unlabeled data setUD. The value of DSV was calculated using the equation (5):

$$DSV(a_{ud}) = \frac{\max_{a_{ud}} dt_{ud} - dt_{ud}}{\max_{a_{ud}} dt_{ud} - \min_{a_{ud}} dt_{ud}} \qquad (6)$$

Now the unlabeled data or input are match with some similar or nearest anomaly actions with the help of clustering the labeled anomaly data and measuring the distance among them. Finally, the Distance Score Value was calculated for every sample from the unlabeled data set. Based on the distance score value of a_{ud}, the similarity among a_{ud} and its nearest anomaly cluster center can be estimated. When the value of $DSV(a_{ud})$is closer to 1, the more likely a_{ud} is to be potentially anomaly. If the value of $DSV(a_{ud})$is closer to 0, the more likely a_{ud} is to be normal data.

ii. Isolation Score Value:

The quantity of labeled anomalies in the training set is substantially lower than the amount of unlabeled data in the training set in this method application situation, hence a few labeled anomalies cannot cover all kinds of anomalies. As a result, mining the labeled anomalies information can only ensure the identification of anomalies of the same kind as the labelled anomalies. To fully use the information included in most normal data inUD, apply iForest to compute the isolation score value $ISV(a_{ud})$ of each sample a_{ud} to anticipate the distribution of samples in the unlabeled data setUD.

iii. Initial Anomaly Score:

The initial anomaly score (IA) was calculated using the isolation score value $ISV(a_{ud})$ and distance score value $DSV(a_{ud})$which is shown in the equation 6 (Gao *et al*. (2021) [6]):

$$IA = (1 - \tau)DSV(a_{ud}) + \tau\, ISV(a_{ud}) \qquad (7)$$

Here $\tau \in (0,1)$ and $IA \in (0,1)$. When the value of IA is closer to 1, the more likely a_{ud} is to be potentially anomaly. If the value of IA is closer to 0, the more likely a_{ud} is to be normal data.

To fully extract the information from the unlabeled data UD, after obtaining the initial anomaly scores $IA = \{ia_1, ia_2, \dots ia_n\}$ of all samples inUD, compute the expectationμ_r and standard deviation σ_r of IA:

$$\mu_r = \frac{1}{n}\sum_{m=1}^{n} ia_m \quad (8)$$

$$\sigma_r = \sqrt{\frac{1}{n-1}\sum_{m=1}^{n}(ia_m - \mu_r)^2} \quad (9)$$

4. Anomaly Scoring Network

The initial anomaly score, labeled data and unlabeled data are given as input to the anomaly scoring network. In the anomaly scoring network $\psi(a;\theta)$is the scoring function which is to be learned in this network. Here a is the input and θ is the parameter to be learned for the respective input data. The anomaly scoring network produce the output $\psi(a;\theta) = AS$.

Let the input space of $\psi(\cdot;\theta)$be $D \subseteq S^d$, the output space be S, and the original data after feature extraction be L. Here D denoted the whole dataset that is labeled and unlabeled data ($D = \{a_1, a_2, \dots a_n, (a_{n+1,}b_{n+1}), \dots (a_{n+m,}b_{n+m})\}$). The objective function $\psi(\cdot;\theta): D \mapsto S$ of the network is divided into two parts. One is feature extraction learner $\phi(\cdot;\theta_f): D \mapsto L$ and another one is anomaly score learner $\varphi(\cdot;\theta_s): L \mapsto S$ with $k = \phi(\cdot;\theta_f)$ and $\theta = \{\theta_f, \theta_s\}$.

After the feature extraction the new data representation is $k \in L$. Particularly, $\varphi(\cdot;\theta_s)$is a fully connected neural network with $T \in N$ hidden layer with the weight metrics $\theta_f = \{W^1, W^2, \dots W^T\}$, where $W^t$is the parameter of the t-th hidden layer. Let $a_t \in S^\varsigma$ is the output of the t-th hidden layer and $W^t = \{w_1, w_2, \dots w_\varsigma\}$; then, calculate the a_t by using the following equation 9:

$$a_t = \{\eta(w_1^G a_{t-1}), \eta(w_2^G a_{t-1}), \dots, \eta(w_\varsigma^G a_{t-1})\} \quad (10)$$

Here $\eta(\cdot)$ is the ReLu activation function. $\varphi(\cdot;\theta_s)$is the anomaly scoring learner that uses a single linear neural unit in the output layer, whose input is the new data representation k after feature extraction, and the output is the anomaly score AS. Let $\theta_s$be the parameters of the anomaly score learner and $\theta_s = \{\omega_1, \omega_2, \dots, \omega_q, c\}$. Now $\varphi(\cdot;\theta_s)$ is computed as follows

$$\varphi(k;\theta_s) = \sum_{i=1}^{q} \omega_i k_i + c \quad (11)$$

Now the anomaly score for each sample and weights are obtained in this phase. The anomaly scoring function should satisfy the condition$\psi(a_1) > \psi(a_2)$. Here $a_1$is anomaly and $a_2$is the normal data object. In the testing phase, apply the anomaly scoring function $\psi(a;\theta)$ to assign anomaly scores to the incoming data objects and identify anomalies by observing their anomaly score.

i. Concentration Loss:

Here few labeled anomaly data and high unlabeled data are used for anomaly detection. This extreme imbalance will lead to the poor accuracy. Because the noises in the unlabeled data are given repercussions to the anomaly detection process (Gao *et al*. (2021)). To overcome this problem, here concentration loss function is utilized.

A new loss function named concentration loss that can be dynamically scaled according to the sample weight. The proposed mechanism applies concentration loss to make sure that the anomaly score of normal data could have a significant deviation from that of anomalies even on a circumstance of extreme class imbalance of positive (anomalies) and negative samples(unlabeled data) in training set. The concentration loss function is calculated using the following equation 11:

$$L(\psi(a;\theta), \mu_r, \sigma_r, \alpha, \beta) = (1-\alpha)(1-b)(1-IA)^\beta |gap(a)| + \alpha b\, max\big(0, m - gap(a)\big) \quad (12)$$

$$gap(a) = \frac{\psi(a) - \mu_r}{\sigma_r} \quad (13)$$

Here $gap(a)$is calculated with the help of the above equation after obtaining the standard deviation and expectation value. Here α is the class balance parameter $\alpha \in [0,1]$ and β is the variant of the proposed VASN model. The purpose of the concentration loss is to reduce the loss and improve the accuracy in the imbalanced data situation.

Algorithm

I**NPUT**: video with n frames (T H W) with images i_f

OUTPUT: anomaly scoring function $\psi(a;\theta)$

Embedding video frame patchesZ_n

Encoding images with position embedding EP

Extract low feature in each frame.

Converting Low resolution to HighResolution images $\hat{Z}_a$ and $\widehat{Z_a}^{l+1}$

Let Cluster samples be LD Labeled data and UD be Unlabeled Data in anomaly scoring network

Calculate Euclidean distance and cluster centre value$dt_{ud} = \min_{C_P} \lambda(a_{ud}, C_P)$

Calculate distance score value $DSV(a_{ud}) = \frac{\max_{a_{ud}} dt_{ud} - dt_{ud}}{\max_{a_{ud}} dt_{ud} - \min_{a_{ud}} dt_{ud}}$

Calculate initial anomaly score value$IA = (1-\tau)DSV(a_{ud}) + \tau\, ISV(a_{ud})$

Compute expectation and standard deviation$\mu_r = \frac{1}{n}\sum_{m=1}^{n} ia_m, \sigma_r \sqrt{\frac{1}{n-1}\sum_{m=1}^{n}(ia_m - \mu_r)^2}$

Initialize Θ value

Extract features by fully connected neural network $\varphi(\cdot;\theta_s)$

Loss function $gap(a) = \frac{\psi(a)-\mu_r}{\sigma_r}$

Calculate output anomaly score value in terms $L(\psi(a;\theta), \mu_r, \sigma_r, \alpha, \beta)$

If Condition $\psi(a_1) > \psi(a_2)$

End

5. Result And Discussion

i. Crowd Localization:

In crowd location, we calculate instance-level Precision, Recall, and F1-measure are calculated under the adaptive scale for each head.

$$percision = \frac{TR.POS}{TR.POS + FL.POS}$$

$$recall = \frac{TR.POS}{TR.POS + FL.NEG}$$

$$F1 = \frac{precision.recall}{precision + recall}$$

where TR.POS, FL.POS, FL.NEG denote the number of True Positive, False Positive, and False Negative, respectively. The TR.POS, FL.POS, and FL.NEG are calculated.Here the F1 score values shows that the proposed model achieves better results compared to the other methods.

a) ST+base decoder: A crowd counting and localization model. Swin Transformer (ST)is used as a backbone to extract features. The decoder is added to the top of Stage 4 in Swin Transformer, which consists of one convolutional layer and two de-convolutional layer output high-resolution score maps.

b) ST+FPN: Different from ST+base decoder, (feature pyramidal map) FPN decoder is employed to fuse the outputs of four stages in Swin Transformer and to output high-quality score maps.

c) DCST+FPN: Compared with ST+FPN, the backbone is replaced by DCST(Dilated convolutional Swin transformer).

d) Proposed VASN method: video Swin transformer is implemented for high image resolution in the video.

In the table 1 shows that the some of the previous method for localization of crowd with the different data set bycomparing the results shows that the score value is increased with the better resolution of images in pixel wise in the video clips

TABLE 1: F1 score values

METHOD	LOCALIZATION		
	F1 score	precision	recall
ST+base decoder	74.5	81.0	69.7
ST+FPN	79.8	80.5	73.0
DSCT+FPN [6]	81.6	84.8	79.5
Proposed VASN	**83.2**	**86.3**	**82.2**

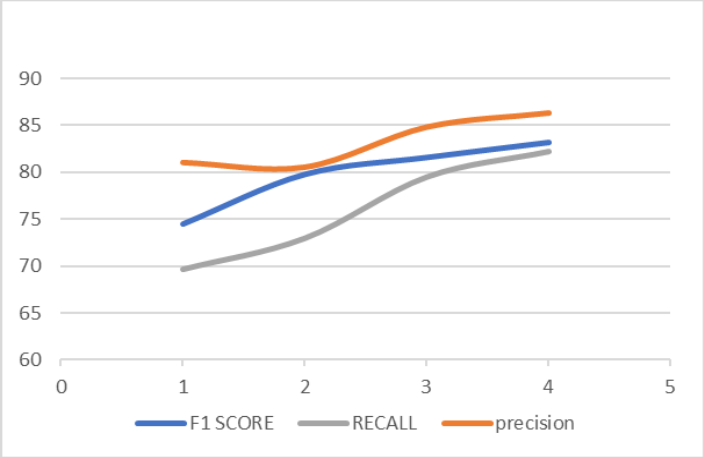

Figure 4: F1 score evaluation

The graphvalue shows the f1 score value , precision , recall value of different method in comparsion with the proposed VASN method.

TABLE 2: comparison of Imageclassification Accuracy with TOP1 and TOP5 accuracy

METHOD	Dataset	ACCURACY		VIEWS	FLOPS	PARAMS
		TOP 1	TOP 5			
TimeSformer -HR	ImageNet 21k	83.5	96.8	1×3	1725	122.5
ViViT-L	ImageNet 21k	84.2	96.2	4 × 3	3999	318.2
Swin-B	ImageNet21k	85.3	97	4×3	290	89.7
Proposed VASN	UMN dataset	87.2	98.5	10×5	2130	205.2

In the Table 2, the comparison of the proposed VASN method with the previous method with TOP1and TOP 5 accuracy in the dataset for abnormal behaviourin the video surveillance for detecting anomaly of the images in each frame. The previous models of various transformer and our proposed VASN method is compared for showing the image resolution accuracy shows better results because this model uses high spatial resolution frame. The view indicates temporal clip ×spatial crop.

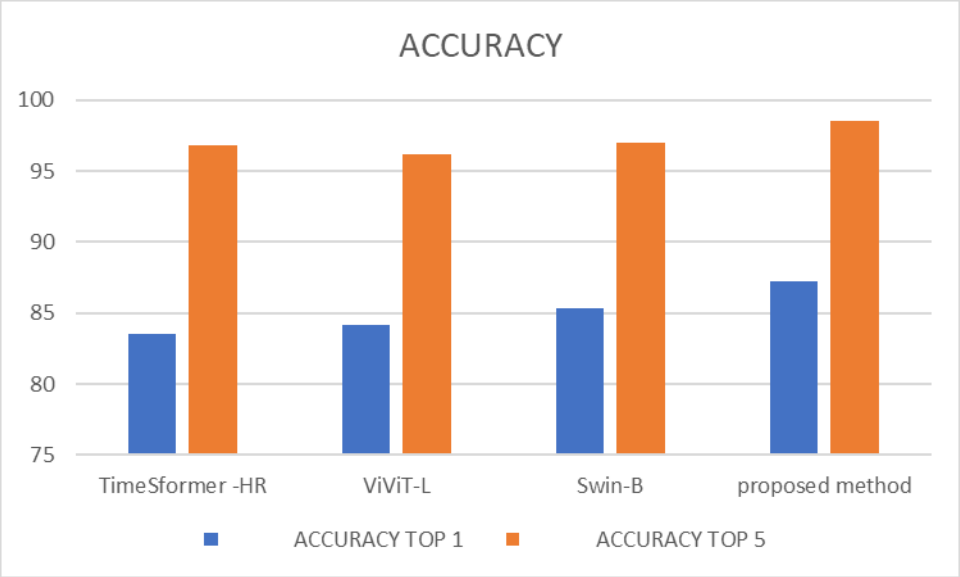

Figure 5: comparsion of Top1 and Top5 accuracy values

Evaluation of Anomaly detection:

To simulate real-world anomaly detection scenarios, we first randomly sample several anomalies from the original data set and randomly exchange 5% of their features in pairs to generate noises. Then we randomly divide the original data set into two subsets. The first subset accounting for 80% of the original data set is the original training set, and the remaining 20% of the original data set accounts for the test set.

We further randomly sample several anomalies from the original training set to form a labeled anomaly set A. Next, we mix the noises with the normal data in the original training set to simulate an unlabeled data set UD with contamination and control the contamination rate by adjust the quantity of noises. Last, the final training set consist of unlabeled data set (UD) and labeled anomalies (LA).

Our method of generating training set and test set can guarantee: (i) in the training set, theunlabeled dataset UD contains different anomaly types from the labeled anomaly dataset A; (ii) the test set contains different anomaly types from the labeled anomaly dataset.

Here we are usingConNet for feature extraction in the anomaly scoring network.compared to the other methods con Net achieves good results in detection of anomaly There are three reasons account for the best performance achieved by ConNet.

First, comparedwith unsupervised methods, our method effectively utilizes a few labeled anomalies to improve the detection accuracy. Second, our method avoids the problem that negative samples (unlabeled data) dominate the model training process dueto the imbalance of positive and negative samples by increasing the influence of positive samples (labeled anomalies)on the model training process. Third, we employ a priorestimation module to estimate the data distribution of the dataset and integrate the estimation results into the model trainingprocess, which effectively alleviates the negative impact ofnoises on the accuracy of model detection.

Table 3: Accuracy evaluation (%)of anomaly score value between VASNand the existing methods

Methods	**Anomaly Score Value**
Online +false alarm [13]	70.9
Trans Anomaly [2]	86.7
Deep anomaly [10]	95.74
Proposed VASN	**98.92**

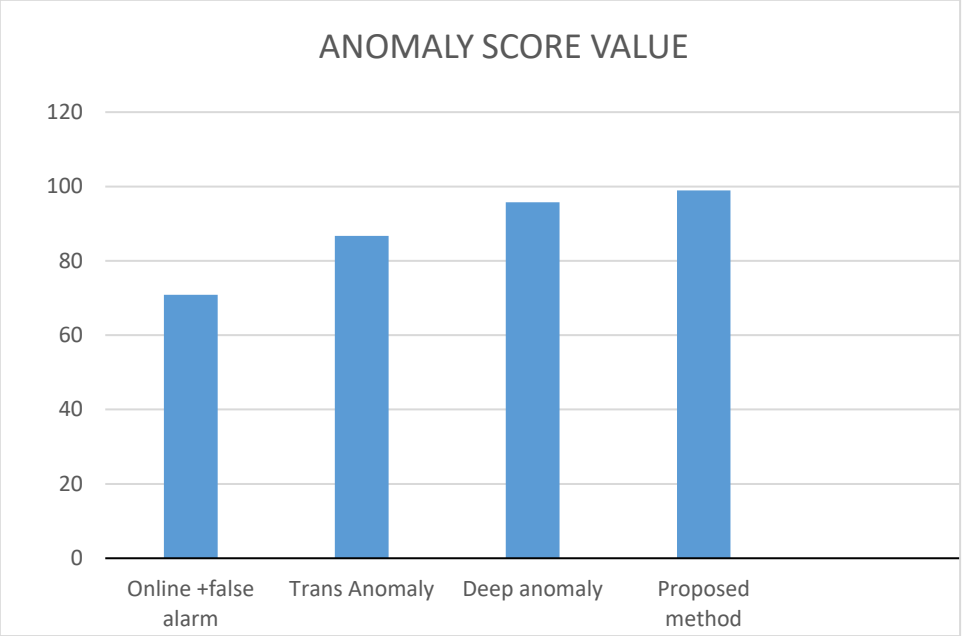

Figure 6: Accuracy evaluation (%)**of anomaly score value betweenVASNand the existing method**

The output the anomaly score values are compared with previous method in this tabulation with different methods for detecting the anomaly in the video surveillance. The proposed method VASN shows that anomaly score values achieves more accuracy and the computational time is reduced. The proposed method VASN finds the anomalous segment in each frame while the previous work failed to detect in the congested crowd behaviour.

6 Conclusion

In this paper, anomaly detection is predicted using proposed VASN in the crowd behaviour by integrating video Swin transformer and anomaly scoringnetwork.Experimental results showed that the anomalies are predicted with high spatial resolution of images in each frame and also showed that accurate prediction of abnormal action of the crowd like theft, accidents, robbery and other crime scenes. The F1 score value is 83.2, the image recognition resolution in top1 and top 5 accuracy 87.2, 98.5. The anomaly output score value can obtain 98.92 of accuracy compared with previous methods.

References

1. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., & Hu, H. (2021). Video SwinTransformer. In *arXiv [cs.CV]*. http://arxiv.org/abs/2106.13230
2. H. Yuan, Z. Cai, H. Zhou, Y. Wang and X. Chen, "TransAnomaly: Video Anomaly Detection Using Video Vision Transformer," in IEEE Access, vol. 9, pp. 123977-123986, 2021, doi: 10.1109/ACCESS.2021.3109102.
3. Lee, J., Lee, S., Cho, W., Siddiqui, Z. A., & Park, U. (2021). Vision transformer-based tailing detection in videos. *Applied Sciences*, *11*(24), 11591. https://doi.org/10.3390/app112411591
4. Wang, T., Qiao, M., Deng, Y., Zhou, Y., Wang, H., Lyu, Q., &Snoussi, H. (2018). Abnormal event detection based on analysis of movement information of video sequence. *Optik*, *152*, 50-60. https://doi.org/10.1016/j.ijleo.2017.07.064
5. C. Huang, Z. Wu, J. Wen, Y. Xu, Q. Jiang and Y. Wang, "Abnormal Event Detection Using Deep Contrastive Learning for Intelligent Video Surveillance System," in IEEE Transactions on Industrial Informatics, doi: 10.1109/TII.2021.3122801.
6. Gao, J., Gong, M., & Li, X. (2021). Congested crowd instance localization with Dilated Convolutional Swin Transformer. In arXiv [cs.CV]. http://arxiv.org/abs/2108.00584
7. Doshi, K., & Yilmaz, Y. (2021). Online anomaly detection in surveillance videos with asymptotic bound on false alarm rate. *Pattern Recognition*, *114*, 107865. https://doi.org/10.1016/j.patcog.2021.107865

8. Feng, J., Liang, Y., & Li, L. (2021). Anomaly detection in videos using two-stream Autoencoder with post hoc Interpretability. Computational Intelligence and Neuroscience, 2021, 1-15. https://doi.org/10.1155/2021/7367870
9. R. Nawaratne, D. Alahakoon, D. De Silva and X. Yu, "Spatiotemporal Anomaly Detection Using Deep Learning for Real-Time Video Surveillance," in *IEEE Transactions on Industrial Informatics*, vol. 16, no. 1, pp. 393-402, Jan. 2020, doi: 10.1109/TII.2019.2938527.
10. Nasaruddin, N., Muchtar, K., Afdhal, A. et al. Deep anomaly detection through visual attention in surveillance videos. J Big Data 7, 87 (2020). https://doi.org/10.1186/s40537-020-00365-y
11. Duman, E., &Erdem, O. A. (2019). Anomaly detection in videos using optical flow and Convolutional Autoencoder. IEEE Access, 7, 183914-183923. https://doi.org/10.1109/access.2019.2960654
12. Chen, D., Wang, P., Yue, L., Zhang, Y., & Jia, T. (2020). Anomaly detection in surveillance video based on bidirectional prediction. *Image and Vision Computing*, *98*, 103915. https://doi.org/10.1016/j.imavis.2020.103915
13. Doshi, K., & Yilmaz, Y. (2021). Online anomaly detection in surveillance videos with asymptotic bound on false alarm rate. *Pattern Recognition*, *114*, 107865. https://doi.org/10.1016/j.patcog.2021.107865
14. C. Direkoglu, "Abnormal Crowd Behavior Detection Using Motion Information Images and Convolutional Neural Networks," in IEEE Access, vol. 8, pp. 80408-80416, 2020, doi: 10.1109/ACCESS.2020.2990355.
15. M. Gnouma, R. Ejbali and M. Zaied, "Abnormal events' detection in crowded scenes", *Multimedia Tools Appl.*, vol. 77, no. 19, pp. 24843-24864, 2018.
16. Pan, L., Zhou, H., Liu, Y., & Wang, M. (2019). Global event influence model: Integrating crowd motion and social psychology for global anomaly detection in densecrowds. Journal of ElectronicImaging, 28(02),1.
17. Annamalai, Manikandan & Muthiah, Ponni. (2022). An Early Prediction of Tumor in Heart by Cardiac Masses Classification in Echocardiogram Images Using Robust Back Propagation Neural Network Classifier. Brazilian Archives of Biology and Technology. 65. 10.1590/1678-4324-2022210316.
18. Manikandan, Annamalai, M,Ponni Bala. (2022). Intracardiac Mass Detection and Classification Using Double Convolutional Neural Network Classifier. Journal of Engineering Research. 65. https://doi.org/10.36909/jer.12237
19. Sheikdavood K, Surendar P, Manikandan A. Certain Investigation on Latent Fingerprint Improvement through Multi-Scale Patch Based Sparse Representation. Indian Journal of Engineering. 2016; 13(31):59-64.
20. S. Dhanasekaran, Dr. P. Mathiyalagan, Rajeshwaran, A. Manikandan, "Automatic Segmentation of Lung Tumors Using Adaptive Neuron-Fuzzy Inference System ", Annals of RSCB, pp. 17468–17483, Jun. 2021
21. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lucic, M., & Schmid, C. (2021). ViViT: A video vision transformer. 2021IEEE/CVF International Conference on Computer Vision *(ICCV)*. https://doi.org/10.1109/iccv48922.2021.00676
22. S. Wan, X. Xu, T. Wang and Z. Gu, "An Intelligent Video Analysis Method for Abnormal Event Detection in Intelligent Transportation Systems," in IEEE Transactions on Intelligent Transportation Systems, vol. 22, no. 7, pp. 4487-4495, July 2021, doi: 10.1109/TITS.2020.3017505.
23. Wang, T., Miao, Z., Chen, Y., Zhou, Y., Shan, G., &Snoussi, H. (2019). AED-net: An abnormal event detection network. *Engineering*, *5*(5), 930-939. https://doi.org/10.1016/j.eng.2019.02.008
24. Manikandan, A., & Sakthivel, J. (2017a). Recognizable Proof of Biometric System With Even Distorted And Rectification States. *Journal of Advanced Research in Dynamical and Control Systems*, *9*(2), 1393–1398.
25. Manikandan, A., & Jamuna, V. (2017). Single Image Super Resolution via FRI Reconstruction Method. *Journal of Advanced Research in Dynamical and Control Systems*, *9*(2), 23–28
26. Manikandan, A., Suganya, K., Saranya, N., Sudha, V., & Sweetha, S. (2017). Assessment of Intracardiac Masses Classification. *Journal of Chemical and Pharmaceutical Sciences*, *5*,101–103.
27. Ashokkumar, N. & Meera, S. & Anandan, P. & Murthy, Mantripragada & K S, Kalaivani & Alahmadi, Tahani & Alharbi, Sulaiman & Raghavan, S. & Jayadhas, s. (2022). Deep Learning Mechanism for Predicting the Axillary Lymph Node Metastasis in Patients with Primary Breast Cancer. BioMed Research International. 2022. 10.1155/2022/8616535.

28. K. N. R. C. K. D. D. T. “Survey on 2D-DCT Based Image Watermarking With High Implanting Limit and Robustness”. *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 4, no. 10, Oct. 2016, pp. 161-4, doi:10.17762/ijritcc.v4i10.2576.
29. Karpagalakshmi, RC, Tensing, D &Kalpana, A.M. (2016). Image Localization using Deformable Model and its Application in Health Informatics. Journal of Medical Imaging and Health Informatics, vol. 6(8), pp. 1972 - 1976. https://doi.org/10.1166/jmihi.2016.1959.
30. Namrata, K,Karpagalakshmi, R,C, Manikandasaran, S,S. (2017). Implementation of Novel Technique for Image Watermarking Using 2D-DCT. International Journal of Pure and Applied Mathematics, volume 117(16),pp. 221-226.
31. T.Ramalingam, R. Umamaheswari , R. C. Karpagalakshmi , K. Chandramohan , M.S.Sabari. (2021). Location of plant Leaf maladies utilizing picture division. Journal of Image processing and Artificial Intelligence. vol.7(3). http://dx.doi.org/10.46610/JOIPAI.2021.v07i03.002.
32. Gopalan, S. H., Vignesh, V., Mahendran, N., & Dinesh, M. T. P. P. (2021). Dynamic Clinical Trials Management in Anunreliable Environment using Blockchain. *Design Engineering*, 817-822.
33. D. S. S, N. H. A. Rufus, D. Anand, R. S. Rama, A. Kumar and A. S. Vigneshwar, "Evolutionary Optimization with Deep Transfer Learning for Content based Image Retrieval in Cloud Environment," 2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), Trichy, India, 2022, pp. 826-831, doi: 10.1109/ICAISS55157.2022.10011122.