Nexo Revista Científica
Vol. 36, No. 03, pp. 404-421/Junio 2023

ISSN-E 1995-9516
Universidad Nacional de Ingeniería

http://revistas.uni.edu.ni/index.php/Nexo
https://doi.org/10.5377/nexo.v36i03.16462

Generative and non-parametric model for real-time event detection in social networks based on textual analysis

Modelo generativo y no paramétrico para detección de eventos en tiempo real en redes sociales basado en análisis textual

Masoumeh Aziziansiadar

Tehran University of Applied Sciences (affiliated to Sharif Academic Jahad), Tehran, Iran

Corresponding author E-mail: masume.azee@gmail.com

(recibido/received: 05-febrero-2023; aceptado/accepted: 08-mayo-2023)

ABSTRACT

One of the things that is followed in monitoring systems is the detection of rare events in real time among the multitude of common events in social networks. Considering the lack of recognition and unavailability of rare events, their detection is considered a challenge. In this research, a new architecture and approach based on generative adversarial network infrastructure was presented to detect common and rare events in real time. In this research, the attempt is to provide a new approach to the performance of architectures based on deep generative adversarial networks, a way to solve various problems without supervision with a semi-supervisory approach and adversarial generative infrastructure. This architecture is based on the automatic extraction and use of video input data features. The results of the equal error rate in the UCSDped1 and UCSDped2 datasets were 2.0 and 17.0, respectively, in the performance characteristic curve.

Keywords: Event detection; social networks; generator; text analysis; non-parametric; real time.

RESUMEN

Una de las cosas que se sigue en los sistemas de monitorización es la detección de eventos raros en tiempo real entre la multitud de eventos comunes en las redes sociales. Teniendo en cuenta la falta de reconocimiento y la falta de disponibilidad de eventos raros, su detección se considera un desafío. En esta investigación, se presentó una nueva arquitectura y un enfoque basado en una infraestructura de red antagónica generativa para detectar eventos comunes y raros en tiempo real. En esta investigación se intenta brindar un nuevo enfoque al desempeño de arquitecturas basadas en redes generativas antagónicas profundas, una forma de resolver diversos problemas sin supervisión con un enfoque semi-supervisor e infraestructura generativa antagónica. Esta arquitectura se basa en la extracción automática y el uso de funciones de datos de entrada de video. Los resultados de la misma tasa de error en los conjuntos de datos UCSDped1 y UCSDped2 fueron 2,0 y 17,0, respectivamente, en la curva característica de rendimiento.

Palabras claves: Detección de eventos; redes sociales; generador; análisis de texto; no paramétrico; tiempo real.

1. INTRODUCTION

Providing security is one of the basic needs of societies. Nowadays, providing security through video surveillance has become common for control in social networks. Also, the cost and number of human resources to control and monitor the content of the videos received on social networks is high. Therefore, it is necessary to take advantage of modern science, and all the necessary control and evaluation measures in these monitoring systems should be done automatically. One of the things that is followed in monitoring systems is the detection of rare events in real time among the multitude of common events in social networks. Various methods have been used in video surveillance systems to detect common and rare events. In this research, an integrated structure based on deep learning is presented for detection and localization common and rare events by using generative adversarial network. Rare events often lead to accidents and damage in social networks, and these systems warn with timely detection in order to provide an appropriate response (Krizhevsky et al., 2017).

The aim of the current research is to train an intelligent system instead of using and displaying common events to the guards and to announce the real-time detection processes automatically. The proposed algorithm is inspired and implemented by generative adversarial neural network. In fact, this algorithm is like a watchman who conjures up all unusual situations in his mind, and detection them without the need for examples of rare events. For this purpose, an architecture based on the model of convolutional neural networks (Krizhevsky et al., 2017) for text analysis and generative adversarial network (Goodfellow et al., 2020) is presented for detection and localization common and rare events in real time in video.
This research is organized in six sections. In the second part, the related researches will be reviewed and in the third part, the definition of generative adversarial networks will be discussed as the infrastructure of this research. Proposed architecture and experimental tests are presented in the fourth and fifth sections, respectively, and conclusions and future suggestions are presented in the sixth section.

2. LITERATURE REVIEW

In the research of Zhou et al. (2016) and Mousavi et al. (2015), different meanings and viewpoints about rare events have been proposed. The problem space of detecting common and rare events faces a lot of uncertainty in real conditions. Due to the fact that all the variations in rare events are not always occurring, it is not possible to have an appropriate training data set that contains all examples of rare events. As a result, solving the problem using supervised approaches will not be successful.

One of the important challenges in detecting rare events is the lack of a specific definition of rare events. The definition that most researches have referred to include events that have a low probability of occurrence. The events that usually occur in any environment are considered as common events and the rest as rare events. One of the other challenges in identifying rare events is the place where the rare event occurs (Mousavi et al., 2015; Xiang and Gong, 2005).

Researchers have addressed the issue of detecting common and rare events in video from different perspectives (Xiang and Gong; 2008; Anjum and Cavallaro, 2009; Zaharescu and Wildes, 2010; Sodemann et al., 2013; Tran et al., 2015; Medel and Savakis, 2016; Chong and Tay, 2017). Every condition may be rare or common according to the filmed environment. For example, on the sidewalk, the passage of people is a natural and common thing, but the traffic of vehicles is considered a rare event in that environment. The problem of rare event detection is a multi-step problem and, in most cases, reaching the desired accuracy requires pre-processing such as background separation, object classification and environment segmentation (Zhou et al., 2016). Each of these operations is considered a problem and challenge by itself.

The necessary parameters correspond to the desired problem and also the installation location of the cameras in different environments such as the corridor, parking lot and room are different. The videos produced by

surveillance cameras (ASUN) are considered as third-person videos. There is another category called first-person video, in which the camera is mounted on a person's clothes, body, or hat. This category of video is not analyzed in this research (Mousavi et al., 2015).

There are different views and classifications in the review and analysis of the input of the problem of detecting common and rare events. Categorizing videos into videos of quiet and crowded environments is one of the usual categories in this field. Also, other cases such as videos that were filmed for one environment, but from several points (Li et al., 2013) and videos that are related to closed or open environments (Zaharescu and Wildes, 2010) are also considered as different categories. The current research related to the detection of rare events is placed in the first category (quiet and crowded environments).

In order to identify rare events, the accuracy and performance of algorithms in quiet environments is better than in crowded environments. Most of the simple and traditional methods provide the desired detection with relative accuracies (Xia et al., 2015). One of the most important challenges of detecting rare events in crowded environments is the problem of overlapping people and objects in the environment. The computational cost of manual methods and the non-uniformity of video traffic are among the other problems of crowded environments. Activities and researches such as detection of the movement and passage of boats in prohibited areas, parks, sidewalks, and city trains are clear examples of researches conducted in quiet environments. To identify common and rare events, spatial, temporal and spatio-temporal features can be defined. The spatial feature includes the representation of the shape and texture of the objects in a frame of the video. In this type of feature, the relationship between successive frames of a video is not considered, and the main focus is on the feature of adjacent areas in the input data (Xia et al., 2015). The temporal characteristic is related to the relationship between the areas of the scenes over time (consecutive frames) (Sabokrou et al., 2017; Biswas and Babu, 2017; Chong and Tay, 2017; Tran et al., 2015; Saligrama, 2013; Mahadevan, 2010).

The temporal characteristic is divided into two general categories: long-term temporal characteristic and short-term temporal characteristic (Xia et al., 2015). The spatio-temporal characteristic is a combination of two described characteristics. In the space of researches related to the field of video analysis, the use of this feature in the direction of identifying the process of changes in any area will help to solve the problems related to behavior detection (Sabokrou et al., 2017; Chong and Tay, 2017; Zhou et al., 2016; Tran et al., 2015; Saligrama et al., 2013; Gorzałczany and Rudziński, 2017).

Considering the limitations in the detection of rare events, the focus is on learning common events. In fact, after learning the common events by the algorithm, according to the environmental conditions, the events that fall outside the appropriate threshold are recognized as rare events (Marsden et al., 2016). Threshold means the degree of concordance of the features extracted by the network compared to the features learned during training (Morris and Hogg, 2000). Some researches, such as Dong et al (2009), have put the main focus on identifying the optimal threshold. Albusac et al. (2009), using the generation of three laws for each camera, identify the unusual behavior of objects with the help of three components: object class, object position, and object speed. To deal with the indeterminacy of each camera, fuzzy logic has been used in decision-making.

The primary methods of detecting common and rare events are often based on modeling the movement path of objects. In the mentioned approach, rare event detection is done by identifying outlier data in common events. In fact, if the desired object has not followed a common movement path, it is considered as an unusual movement of that object. Jiang et al. (2011) presented the idea of using three levels of spatio-temporal and path concepts in order to detect rare events. Shandong Wo et al. (2010) identified existing anomalies with K-means clustering. Kratz and Nishino (2009) identified the position of movement patterns through the three-dimensional Gaussian spatio-temporal gradient distribution. In this approach, rare events are detected by Hidden Markov Model.

Most of the mentioned researches, based on complex manual characteristics, are considered to represent the scene and the movements within it. But recently, the trend of research activities in this field is based on deep learning (Sabokrou et al. 2018; Anjum and Cavallaro, 2009; Tran et al. 2015; Medel and Savakis, 2016; Saligrama et al., 2013). Also, the approach of extracting cubic chunks has been used by Cascade classifier to identify and check the type of extracted chunk by Sabokrou et al. (2017) for automatic feature learning. The use of the fully connected structure in presenting the representation of the feature related to the rare event is also referred to as deep-anomaly in the research of Sabokrou et al (2017). Considering the change of approach that is also evident in other vision and machine learning issues; This research also uses deep learning approaches for textual analysis and generative adversarial network to solve the problem of detecting and locating common and rare events in the video, which will be explained later.

3. DESCRIPTION AND MODELING OF THE PROBLEM

The generative adversarial network, which is called GAN for short, has been addressed as a spark for deep learning (LeCun and Bennett, 2016). More details of this network will be provided below.

3-1. Introducing the generative adversarial network

This network was presented by Goodfellow et al. in 2020. This research has created a generative adversarial network in the field of machine learning and deep networks by implementing the idea of learning two neural networks in the space of adversarial learning. GAN consists of two sub-networks known as generator and discriminator. Each of these networks has a special architectural structure, but the output of the generator is used as the input of the discriminator.

The generative network is used to learn to produce examples similar to the original data set of the problem. In its initial version, this network has a vector sampled from a specific (for example, normal) distribution as its input. This vector reaches a structure corresponding to the data structure of the original data set in the output layer after passing through the layers. In fact, in the simple generative network, it is common to map the input vector to the data structure of the original data set. In fact, data is produced, which is the main issue in the data dimensions of the data set. It should be noted that the meaning of real data is the set of training data of the problem and the meaning of fake data is the data generated by the generator network.
In the research of Goodfellow and his colleagues, the generative network has been likened to "a counterfeit group effort to produce fake currency" and the discriminator network to "the police trying to detect counterfeit currency". The counterfeiter (generative network) tries to deceive the police (discriminator network), but in return, the police try not to be deceived. The generator network is known by the generator function G and the discriminator network is known by the discriminator function D. These two networks are competing with each other with the min-max rule and the zero-sum rule. The overall cost function V (G, D) is defined as follows (Goodfellow et al. 2020):

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}\left[\log\left(1 - D(G(z))\right)\right] \quad (1)$$

The discriminator network is fed from the set of training data and tries to learn the distribution of the original data. This network is trying to drive the real probability of the training data (real data) to one and the real probability of all the outputs of the generating network (fake data) to zero. In the above relation, the probability of the discriminator network for the input data x is shown by D(x). In the process of training the network, when x is sampled from the training data distribution $p_{data}(x)$, the probability of this sample tends to one. The z vector is generated by sampling the specific distribution of p. The generative network produces G(z) by receiving samples of z in its output (this output is the same as the original data structure of the

problem). The discriminator network should produce the zero probability as the value of D(G(z)) in the output by receiving the generated artificial (fake) data G(z). This is while the generator network is trying to make the production data G(z) similar to the training data of the problem and finally D(G(z)) will be equal to one. This process of competition between the three mentioned cost functions continues until the relative learning of the two networks.

D and G networks have their own parameters and architectures. If G network parameters are considered with $\Theta^{(g)}$ and D network parameters are considered with $\Theta^{(D)}$; The cost function is denoted by $V(\Theta^{(g)}. \Theta^{(d)})$, the output of the generating network by $x^{\wedge} = G(; \theta^{(g)})$ and the discriminator network by D (x; $\Theta^{(d)}$).

3-2. Development and application of generative adversarial network

According to the idea presented in the research of Mirza and Osindero (2014), by injecting separate information, it becomes a conditionally generative adversarial network. In fact, conditional information is injected into the set of networks in the form of additional information. In the example of handwritten digits MNIST[1], the mentioned condition is considered on the class label. This idea has also been used to express the approaches of learning multi-state models in the dataset MIRFlicker-25000 (Huiskes and Lew, 2008). The use of convolutional layers in the generative adversarial network infrastructure has been used in the research of Radford et al. (2015). Normally, convolutional networks use supervised learning. In this way, according to the general type of unsupervised learning in GAN, an attempt was made to use the advantages of supervised learning and unsupervised learning together by using convolutional layers (Radford et al., 2015).

The goal of semantic segmentation is to assign a label to each pixel of the image. For attribution, a certain amount of labeled data is required at pixel-level and is often not available. To show this lack of information, in the research of Souly et al. (2017), an attempt was made to firstly focus on a large amount of unlabeled or semi-labeled information and secondly on the unrealistic images produced by the generative adversarial network. The basic architecture of the mentioned research is presented as a semi-supervised structure based on generative adversarial network. Schlegl et al. (2017) have used unsupervised methods to identify abnormalities in filmed data along with candidate signs. The AnoGAN method is designed to learn manifold of normal anatomical variability, along with the anomaly scoring overview based on the mapping of the image space to the latent space. In fact, the goal is to apply the anomaly label to the new data and score the patches of images in order to fit the learned distribution (Schlegl et al., 2017).

4. PROPOSED ARCHITECTURE

In this section, various aspects of the proposed architecture, including the data used, architecture and algorithm are explained. This architecture detects and locates rare events in real time in a video, and naturally, other common events are considered.

[1] http://yann.lecun.com/exdb/mnist/

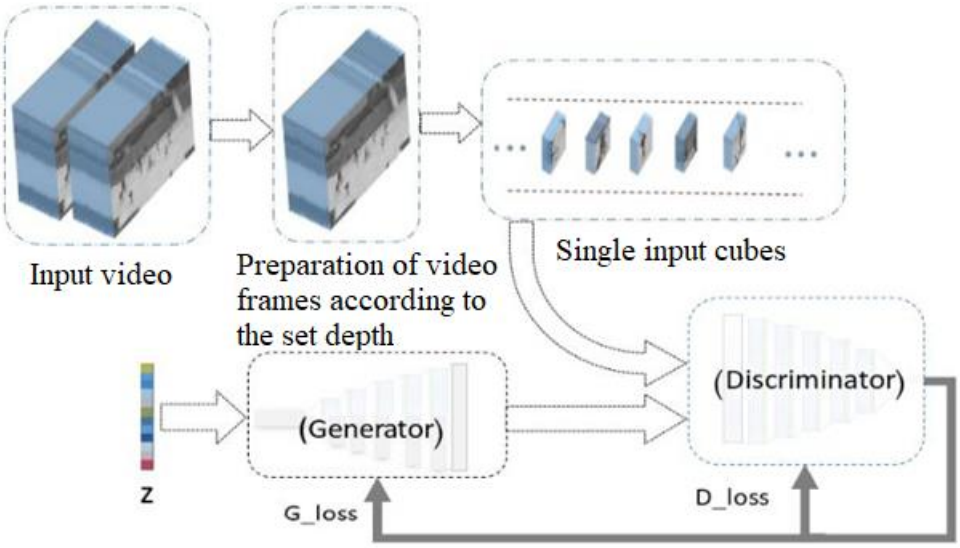

Figure 1. General overview of the proposed architecture

The working procedure of this architecture is that at first, frames are extracted from the received video according to the defined number of frames. In the next step, all the frames in a batch are fragmented according to the region size setting parameter. At the end of this step, frame cubes are created. If the set number of frames is considered to be one, this cube will be converted to one frame. The process of data preparation in the stages of training and testing is similar.

The architecture of the proposed algorithm is similar to the generative adversarial network, consisting of two generative and discriminator networks based on two phases of training and testing. In the training phase, the schematic structure of Figure 1 is used. In the test phase, only the discriminator network is used. In fact, in the test phase, the input data are prepared according to the mentioned stages, and when these data are entered into the discriminator network, a probability is given that the desired data is common. Then, in the post-processing stage, according to the environmental conditions and the threshold parameter, common and rare events are determined. In the following, different parts of architecture are described.

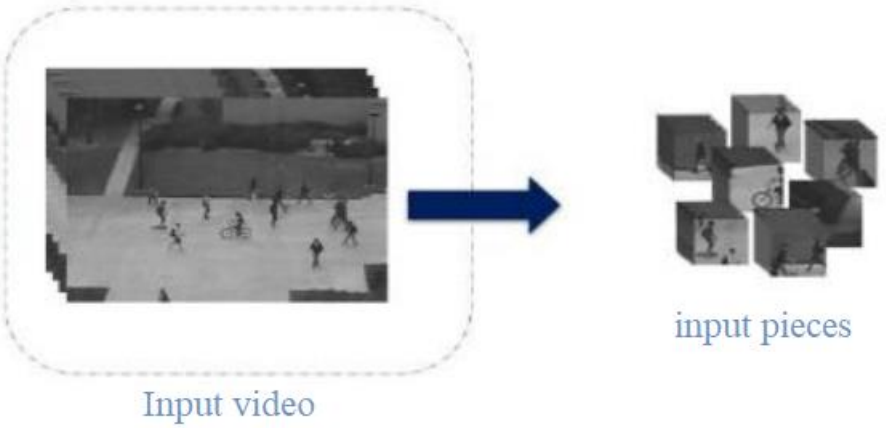

Figure 2: Preparation method for sending to the generative adversarial network

4-1. Data preparation

The input of the proposed architecture will be a number of video frames. Each frame is zoned according to the spatial and temporal points, according to the required accuracy. Then, in parallel, patches are created and prepared for pre-processing (Figure 2). Next, the data is fed into the training or testing structure of the generative adversarial network (Figure 1).

An important point in the structure of most researches based on convolutional networks is the lack of professional pre-processing (Bertini et al., 2012; Sabokrou et al., 2018; Chong and Tay, 2017; Xia et al., 2015; Dong et al., 2009, Schlegl et al., 2017; Ioffe et al., 2015; Jodoin et al., 2008; Jiang et al., 2011).

4-2. Entering information into the network

When feeding the discriminator network with the samples of the original data set (Figure 3), this network tries to bring the probability of the samples of the original data set or $P_{real\_data}(x)$ closer to the maximum real value (i.e. one). In the generative network, the conditions are slightly different due to its generative nature. In fact, the generative network, starting from a vector sampled from a specific distribution, manages the values in such a way that in the end it reaches the construction of an array of the same size and similar to the input data set of the discriminator network. Then the generated data is audited to discriminator delivery.

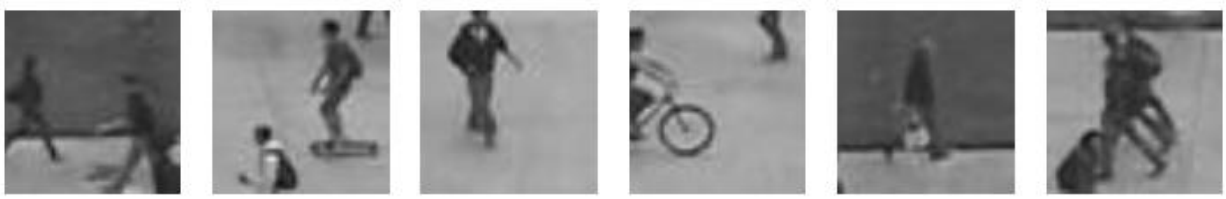

Figure 3. An example of the data area prepared for sending to the competitive network

The current research is based on the basic research of Radford et al. (2015). In the structure of the current research, the idea of increasing the speed of batch normalization expressed in the research of Ioffe et al. (2015) has also been used. Also, various operator functions, including ReLU and PLReLU have been used in architecture (He et al., 2015). The framework for learning generative adversarial networks is such that it provides a good view of the distribution of the original data. Also, due to the unsupervised nature of detecting rare events in real time, using this learning framework, it is possible to study the set of common environmental data and cover the previous limitations. The rare event detection procedure is also in line with the response of the discriminator network. In generative network architecture, it is used as a producer of unusual conditions. This network can also be used in activities such as data augmentation. The important point is that not necessarily all the cases produced by the generator network are among rare conditions, so in this research we call them uncommon cases. In other words, there may be uncommon cases that are not possible to be realized in the real environment as rare. Therefore, these uncommon events are not placed in any category for being unreal (therefore, they are neither rare nor common).

In this approach, the detection of common and rare events is based on the probability value of the discriminator network. Since each video is made up of a number of frames, the input information is entered into the network according to the spatial conditions of each part of the frame. In fact, the regions of common event frames are considered as a set of training data for training. After the training stage, the discriminator network model has the ability to detect common and rare events.

4-3. Architecture and training of generator adversarial network

In this structure, the discriminator network considers a probability $P_{normal\_data}(x)$ for each common video training data. During the training procedure, the discriminator network will drive this probability to one for

the training data and to zero for the production samples of the generative network (fake). At the end of dual network learning, equation (2) is established (Goodfellow, 2020).

$$D^*(x) = \frac{P_{normal\_data}(x)}{P_{normal\_data}(x) + P_{generated\_data}(x)} \qquad (2)$$

According to equation (2), at the end of the optimal training conditions, the data produced by the generative network is very similar to the original data set, so the discriminator network assigns it a probability of one (100% real). Now, according to the mentioned equation, the discriminator network model has to announce the realness of each data with a probability of 50% from now on. In this case, the generating network has performed so well that the discriminator network is not able to separate the original data from the generated fake data. During the training process, the generative network is trying to produce output in such a way that the probability assigned to it by the discriminator network is close to one. This procedure is repeated a certain number of times during training. Then the generative network based on its goal updates the weights. One of the challenges of learning a generative adversarial network is the process of stopping training. The structure of training is always developed in such a way that the learned distribution is in line with the distribution of the training data of the problem (frames of common videos). Also, the output values of the discriminator network should be provided for uncommon data with values close to zero. In this study, the output of the discriminator network is considered as a detective factor and a scoring criterion for each area of the input data. In equation (3), the problem of optimizing the learning process of the proposed inspired framework is expressed in the form of a min-max problem (Goodfellow, 202):

$$\min_G \max_D V(D|G) = \mathbb{E}_{ic\sim P_{normal\_dc}}[log D^*(ic)] + \mathbb{E}_{gc\sim P_g}[\log(1 - D^*(gc))] \qquad (3)$$

ic means the cube of the input data, *normal_ic* is the common data in the given problem, P_{normal_ic} is the distribution of the common data set, P_g is the learned distribution of the generative network, *gc* is the cube produced by the generative network. The symbol is used to sample the original dataset as well as the output dataset of the corresponding generative network with the associated subscript. In this equation, the discriminator network (D) is trying to show the probability value of one in the output for the original inputs x. It is also trying to show the value of zero for the data produced by the generating network in the output. The purpose of the generating network is the opposite of the objective of the discriminator network. In fact, the generating network is trying to produce the generated data in such a way that the probability of $D^*(gc)$ approaches one (the degree of realness of the data).

5. IMPLEMENTATION

The architecture presented in (Radford et al., 2015) has been used as the basic architecture of this research. The proposed architecture is a non-parametric method, which does not depend on a specific distribution and extracts the desired functions from the data. In the single-frame approach, all convolutional layers are performed with $1 \times 5 \times 5$ kernels. These kernels pass with a movement step of 2 in each direction. In fact, a 5×5 kernel with a depth of one is applied to the input data of the convolutional layer. On the left side of Figure 4, the generative network architecture is drawn and on the right side, the discriminator network architecture is drawn. The discriminator and generative networks of 5 convolutional layers are designed according to the architecture inspired by the research of Radford et al. (2015).

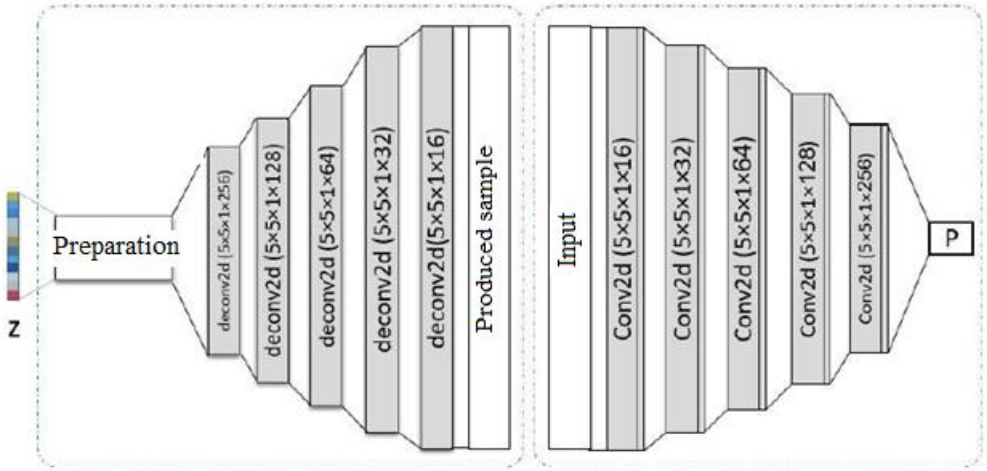

Figure 4. Generative and discriminator networks in adversarial generative convolution network

In the implemented configuration, the amount of feature map associated with each convolutional layer is adjusted according to the processing equipment. Since the architectural basis of this research is the research of Radford et al. (2015) the growth of feature map in architectural layering is considered as a factor of powers of 2. It should be noted that the use of the proposed architecture is inspired by the architecture presented by Radford et al. (2015) and Ullah and Petrosino (2015).

In the research of Ullah and Petrosino (2015), the characteristics of the pyramid have been examined in detail. Also, in different experiments, the input vector of the generative network is considered separately in the size of 100, 200, 400 and 500, so that the effect of the input size of the z vector on the output of the generative network and also the training of the discriminator network is investigated.

In the presented architecture of the current research, the output of each convolutional layer is given to the batch normalization layer according to the research of Ioffe et al. (2015), with the values of the parameters in Table 1. It should be mentioned that in this research, according to the different implementations of the batch normalization function, to apply this layer, the implementation of the research of Ioffe et al. (2015) in the special function of the TensorFlow library named tf.contrib.layers.batch_norm[2] has been used. The output of the batch normalization step has also passed through the detector (activity function).

Table 1. Batch normalization layer arguments

Parameter name	Values
Epsilon	1E-5
Momentum	0.9

It should be mentioned that in the non-parametric convolutional layer, for the initialization of the weights, a truncated normal distribution with a standard deviation of 02.0 and a zero bias value is used. Also, the initialization of Glorot and Bengio (2010) has also been used in some proposed architecture experiments for initialization. Also, the amount of information leakage coefficient in LReLU is considered to be 2.0. Tensorflow version 2.1 was used in this research. In order to implement the deconvolutional operation, the convolutional transpose implemented in the TensorFlow library named tf.nn.conv2d_transpose[3] is used. In case lower versions are used for implementation, there are limitations that have been taken into account.

[2] https://www.tensorflow.org/api_docs/python/tf/contrib/layers/batc h_norm
[3] https://www.tensorflow.org/api_docs/python/tf/nn/conv2d_transp ose

For example, in TensorFlow versions lower than 7.0, tf.nn.deconv2d is used to implement deconvolutional operations.

6. EXPERIMENTAL EXPERIMENTS

This section describes the evaluation criteria and data sets and then analyzes the various tests on the proposed method.

6-1. Evaluation criteria

According to the main goal of detecting common and rare events, this problem is modeled as a two-class problem. One of the usual methods in examining and evaluating two-class problems is to show the performance by the Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC). These criteria are measured based on characteristics such as the true positive rate and the false positive rate that are given below:

6-1-1. True positive rate (TPR)

The correctness rate of the suggested data from the total correct data is referred to as the true positive rate or sensitivity, whose equation is as equation 4. TP and FP mean true positive and false positive, respectively.

$$TPR = \frac{TP}{TP + FP} \qquad (4)$$

6-1-2. False positive rate (FPR)

The rate of the functional error rate of the proposed architecture in detect compared to the total amount of available detection is known as the false positive rate, whose equation is according to 5.

$$FPR = \frac{FP}{TN + FP} \qquad (5)$$

6-1-3. Characteristics of FPS

Frames per second (FPS) is used to evaluate the time efficiency of video analysis systems. This feature shows the number of frames calculated per unit of time (seconds).

6-2. Data used

In this study, two data sets of UCSD (University of California San Diego) with the names UCSDped1 and UCSDped2 have been used for the purpose of teaching and testing proposed architecture[4]. These two data sets include video images of people's traffic surveillance cameras on the sidewalk and street in real time. In this dataset, pedestrians, skateboarders, cyclists and vehicles pass. The UCSDped1 dataset includes 34 training video samples and 36 test video samples. The UCSDped2 dataset consists of 16 training video samples along with 12 test video samples. In this data set, the movement of people is horizontal. Most training examples include 120 frames. Examples such as the movement of a skateboarder, the passage of a vehicle, carrying a backpack are examples of rare pieces of data, and the rest of the patches (such as pedestrians walking, the nature of the environment) are examples of common patches. In this test, the

[4] http://www.svcl.ucsd.edu/projects/anomaly/dataset.html

network architecture has been examined on different sizes of the input frame from 35×35 to 60×60; And 45×45 was among the best situations that were chosen according to the position of the camera and common expected objects and the criteria of examining the problem to advance the research.

6-3. Hardware and software details used in the research

The processing platform of the research has a graphics card with NVIDIA Geforce GTX TITAN specifications, with a space of 32 gigabytes of read only memory. Of course, according to the conditions, several other processing platforms have also been evaluated, the details of which have also been presented. Also, in the software discussion, Linux operating system and Ubuntu distribution and version 04.16 are used. It should be noted that Python version 5.3 and TensorFlow 2.1 were used for implementation (Table 2).

Table 2. hardware/software details of the current research

hardware/software	**details**
Processor	Intel Core i7-3770 @3.40GHz ×8
RAM	32.0 GB
System type	64-bit
GPU	NVIDIA Geforce GTX TITAN
os49	Ubuntu 16.04

6-4. Results of examination on UCSDped1 and UCSDped2

The UCSDped1 collection is considered one of the hard collections due to the depth movement that people have (they move away from the camera and the existing objects change position from large to small due to the camera position). In the following, the results of the receiver operating characteristic on these sets of data are given. The false positive rate in this batch of patients is high compared to other similar patients due to the difficult conditions of detection. As a result, the equal error rate was higher than UCSDped2.

The results obtained from the UCSDped1 and UCSDped2 datasets are shown in Tables 3 and 4, respectively. Also, their comparative chart is given in Figures 5 and 6. The results of this research have a good accuracy compared to the average of recent researches, and this amount is in the appropriate ranking compared to other researches in terms of its processing speed.

Also, the ease of training and the short stages of training are also among the other advantages of this research. In fact, using the generative adversarial approach proposed by this research to solve the problem, compared to the traditional and recent methods used in this field has a desirable growth of 18 percent compared to the basic approach. Of course, compared to some researches such as Sabokrou et al. (2017), the equal error rate is much higher, but the advantage of this research compared to the previous methods is the implementation of an integrated infrastructure (end to end) and great ease in the training and testing procedure. Also, the benefit of the side features of the generative network in the production and data augmentation can be mentioned besides its main purpose. It should be noted that positioning is also done as a single response by the discriminator network detection and does not require any other special steps.

Table 3. The results obtained from the UCSDped1 collection

Method	EER (%)	AUC (%)
Adam et al. (2008)	38	65
MPCCA+SF (Mahadevan et al. 2010)	32	59
Xu et al. (2015)	25	82
Li et al. (2013)	21	9087

Sabokro et al. (2017)	08	2.93
Proposed architecture	20	64

In Figure 5, the results of the equal error rate of different methods for the UCSDped1 data set are depicted. The approach of the current research has been able to play a very good role compared to the basic and appropriate methods of adversarial learning. The advantage of integrated training (end-to-end) research compared to the cross-sectional and complex training approach of Sabokrou et al. is one of the optimal advantages of this method to advance the goal during training and testing.

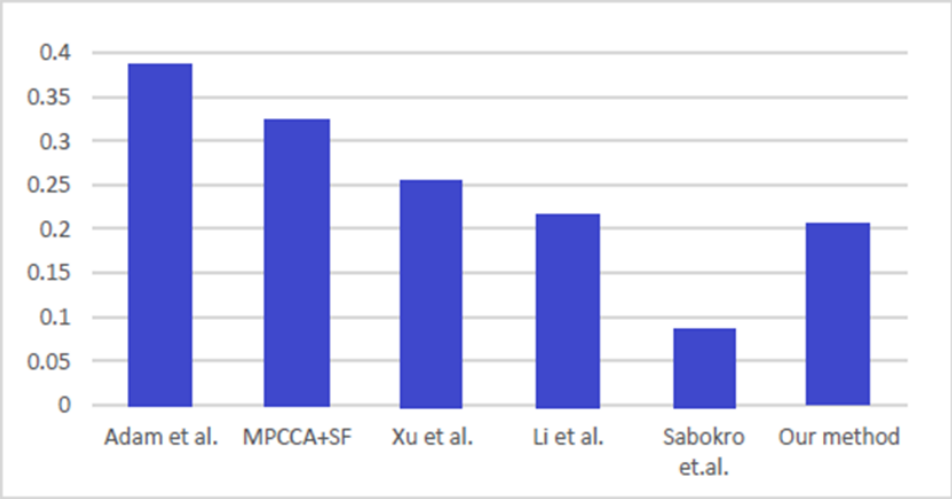

Figure 5. Equal error rate of UCSDped1 data set in different methods

The results of the examination of the proposed model in the detection of rare conditions of the UCSDped2 data set are also given in Figure 6. As it is clear, the proposed approach can have better results compared to the basic methods. Also, due to the simplicity of the UCSDped2 data set compared to UCSDped1, the research Deep-anomaly is also limited to stating the results of UCSDped2. But the results of the current research were on both datasets, and the results indicated its better performance on the UCSDped1 dataset. As mentioned in the upcoming suggestions section, the generative adversarial approach will be a new path in the field of detecting and locating common and rare events.

Table 4. Results obtained from the UCSD ped2 collection

Method	EER (%)	AUC (%)
Adam et al. [46]	42	63
MPCCA+SF [20]	36	2.61
Xu et al. [47]	21	2.88
Li et al. [8]	5.18	-
Sabokro et al. [3]	5.7	9.93
Deep-anomaly [5]	11	-
Proposed architecture	17	80

Figure 6. Equal error rate of UCSD ped2 dataset in different methods

In Figure 7, the score of the occurrence of rare events in the fragments of one of the UCSDped2 video frames is given in the form of a statistical chart with the approach of visualizing the results.

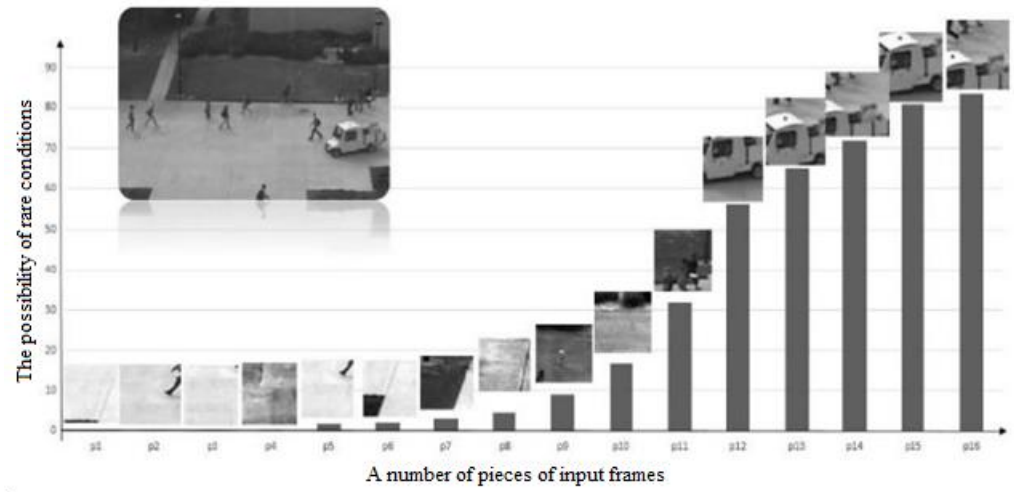

Figure 7. Visualization of the area score diagram in a video frame by the UCSDped2 per-data discriminator network.

6-5. The effect of the input size of the generating network (z)

In this evaluation, the dimensions of the input vector z in the dimensions of 100, 200, 400 and 500 have been considered in order to assess the improvement of training in the generative network. Changing such parameters requires changing various other parameters for the convergence of training. The results of this experiment have shown that it is obvious that the generator network becomes more complicated by choosing large values compared to choosing small values. The complexity of the generative network makes the sensitivity and the need to optimally choose other parameters more difficult in training. The evaluations have shown the value of 200 as a better convergence rate in the generative network for generating data according to Figure 8 in the case of batch normalization.

6-6. The effect of batch normalization

In this experiment, the proposed network architecture, with the same parameters and structure, has been implemented and implemented in two different conditions. The use and non-use of batch normalization in the layer arrangement is considered as the conditions of this experiment. The results shown in Figure 8 show the great effect of batch normalization in the learning process of the proposed network as well as networks based on the generative adversarial network. In fact, the application of batch normalization by reducing the internal covariance shift reduces the focus on parameter quantification and controls this procedure at the input of its own layer.

As can be seen in Figure 8, a view of the output of the proposed architecture without using batch normalization (first row) and with batch normalization (second row), the results obtained by using batch normalization are better than not using it.

In the case without applying batch normalization (first row of Figure 8), from iterations 98 onwards, the learning space of the generative network is completely lost and it is necessary to change the various parameters of the network for better convergence. In the case of applying batch normalization (the second line of Figure 8), the network has good convergence in the learning space of the training datasets. In

comparison with the previous state, from the 98 iterations onwards, he has covered the general distribution and focused on the details of the data distribution. Therefore, by using batch normalization, network training is done better.

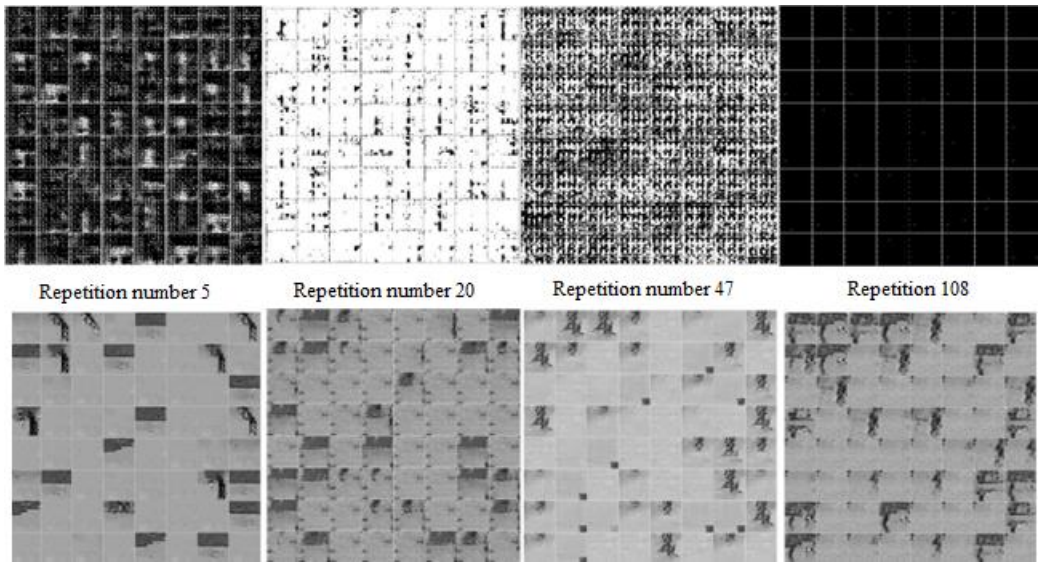

Figure 8. A view of the output of the proposed architecture without applying batch normalization (first row) and with batch normalization (second row).

7. CONCLUSIONS

In this research, a new architecture and approach based on generative adversarial network infrastructure was presented to detect common and rare events in real time. This architecture is based on the automatic extraction and use of video input data features. This research is aimed at detecting common and rare events in real time, focusing on the discriminator network. Also, the method of training proceeds by using the set of training data (common conditions) as training data. The main idea of the research is to use the learned model of the discriminator network in the test stage for the purpose of detection. This detection is in such a way that a score is assigned to each area and based on the obtained score, it is compared with the threshold limit and then its rarity or commonness is determined. The use of batch normalization is also an effective factor in improving the speed and convergence process of the generative adversarial network in architecture, among other results of the current research. The optimal size of patches of the main collection of 45×45 was chosen. Also, the input vector of the generating network was evaluated in different dimensions. The results of the equal error rate in the UCSDped1 and UCSDped2 datasets were 2.0 and 17.0, respectively, in the receiver operating characteristic.

It should be noted that the evaluation in extreme rare conditions (For example, the movement of objects far from the camera and the same color as the environment) with an equal error rate of 2.0 and in medium rare conditions (for example, vehicle traffic on the sidewalk) with an equal error rate of 04.0 in the receiver operating characteristic. Also, detection in 02.0 seconds and equivalent to 300 frames per second for a set of input frames is also one of the other advantages of this proposed architecture in comparison with similar architectures, which shows the ability to be used in the real-time space. Also, the use of parallel implementation to identify each area of the image allows the process of assigning different areas to be done quickly and the score of the area to be determined in a very short period of time. At the end, the obtained results were examined along with other architectures that had good results in this field. According to the application of the integrated approach (end to end) of this research, the ease in the learning stage (training)

is also one of its advantages compared to previous researches with an equal error rate, and comparable results have been presented.
In this research, an attempt was made to provide a new approach to the performance of architectures based on deep generative adversarial networks, a way to solve various problems without supervision with a semi-supervisory approach and generative adversarial infrastructure. In the following, it is possible to use infrastructures similar to the proposed architecture. Also, it will be useful to focus on improving the approach of optimizing the adversarial learning process of this research to solve the problem. Applying the best methods to check all cases of input parameters, such as brute-force, the ability to provide the range of values of each parameter, directions of actions in input parameters are among other things that will help to improve this research.

REFERENCES

Adam, A., Rivlin, E., Shimshoni, I., & Reinitz, D. (2008). Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE transactions on pattern analysis and machine intelligence*, *30*(3), 555-560.

Albusac, J., Castro-Schez, J. J., López-López, L. M., Vallejo, D., & Jimenez-Linares, L. (2009). A supervised learning approach to automate the acquisition of knowledge in surveillance systems. *Signal Processing*, *89*(12), 2400-2414.

Anjum, N., & Cavallaro, A. (2009, September). Trajectory association and fusion across partially overlapping cameras. In *2009 sixth ieee international conference on advanced video and signal based surveillance* (pp. 201-206). IEEE.

Bertini, M., Del Bimbo, A., & Seidenari, L. (2012). Multi-scale and real-time non-parametric approach for anomaly detection and localization. *Computer Vision and Image Understanding*, *116*(3), 320-329.

Biswas, S., & Babu, R. V. (2017). Anomaly detection via short local trajectories. *Neurocomputing*, *242*, 63-72.

Chong, Y. S., & Tay, Y. H. (2017). Abnormal event detection in videos using spatiotemporal autoencoder. In *Advances in Neural Networks-ISNN 2017: 14th International Symposium, ISNN 2017, Sapporo, Hakodate, and Muroran, Hokkaido, Japan, June 21–26, 2017, Proceedings, Part II 14* (pp. 189-196). Springer International Publishing.

Dong, Q., Wu, Y., & Hu, Z. (2009). Pointwise motion image (PMI): A novel motion representation and its applications to abnormality detection and behavior recognition. *IEEE transactions on circuits and systems for video technology*, *19*(3), 407-416.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139-144.

Glorot, X., & Bengio, Y. (2010, March). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 249-256). JMLR Workshop and Conference Proceedings.

Gorzałczany, M. B., & Rudziński, F. (2017). Generalized self-organizing maps for automatic determination of the number of clusters and their multiprototypes in cluster analysis. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(7), 2833-2845.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision* (pp. 1026-1034).

Huiskes, M. J., & Lew, M. S. (2008, October). The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval* (pp. 39-43).

Jiang, F., Yuan, J., Tsaftaris, S. A., & Katsaggelos, A. K. (2011). Anomalous video event detection using spatiotemporal context. *Computer Vision and Image Understanding*, *115*(3), 323-333.

Ioffe, S., & Szegedy, C. (2015, June). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning* (pp. 448-456). pmlr.

Jodoin, P. M., Konrad, J., & Saligrama, V. (2008, September). Modeling background activity for behavior subtraction. In *2008 Second ACM/IEEE International Conference on Distributed Smart Cameras* (pp. 1-10). IEEE.

Kratz, L., & Nishino, K. (2009, June). Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 1446-1453). IEEE.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84-90.

LeCun, Y. (2016). What are some recent and potentially upcoming breakthroughs in deep learning. *Machine learning forums*.. URL https://www. quora. com/What-are-some-recent-and-potentiallyupcoming-breakthroughs-in-deep-learning.

Li, W., Mahadevan, V., & Vasconcelos, N. (2013). Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence*, *36*(1), 18-32.

Marsden, M., McGuinness, K., Little, S., & O'Connor, N. E. (2016, September). Holistic features for real-time crowd behaviour anomaly detection. In *2016 IEEE International Conference on Image Processing (ICIP)* (pp. 918-922). IEEE.

Mahadevan, V., Li, W., Bhalodia, V., & Vasconcelos, N. (2010, June). Anomaly detection in crowded scenes. In *2010 IEEE computer society conference on computer vision and pattern recognition* (pp. 1975-1981). IEEE.

Medel, J. R., & Savakis, A. (2016). Anomaly detection in video using predictive convolutional long short-term memory networks. *arXiv preprint arXiv:1612.00390*.

Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Mousavi, H., Mohammadi, S., Perina, A., Chellali, R., & Murino, V. (2015, January). Analyzing tracklets for the detection of abnormal crowd behavior. In *2015 IEEE Winter Conference on Applications of Computer Vision* (pp. 148-155). IEEE.

Morris, R. J., & Hogg, D. C. (2000). Statistical models of object interaction. *International Journal of Computer Vision*, *37*, 209-215.

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*..

Sabokrou, M., Fayyaz, M., Fathy, M., & Klette, R. (2017). Deep-cascade: Cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes. *IEEE Transactions on Image Processing*, *26*(4), 1992-2004.

Sabokrou, M., Fayyaz, M., Fathy, M., & Klette, R. Fully Convolutional Neural Network for Fast Anomaly Detection in Crowded Scenes. arXiv 2016. *arXiv preprint arXiv:1609.00866*.

Saligrama, V., Arias-Castro, E., Chellappa, R., Hero, A. O., Nowak, R., & Veeravalli, V. V. (2013). Introduction to the issue on anomalous pattern discovery for spatial, temporal, networked, and high-dimensional signals. *IEEE Journal of Selected Topics in Signal Processing*, *7*(1), 1-3.

Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U., & Langs, G. (2017, May). Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *Information Processing in Medical Imaging: 25th International Conference, IPMI 2017, Boone, NC, USA, June 25-30, 2017, Proceedings* (pp. 146-157). Cham: Springer International Publishing

Sodemann, A. A., Ross, M. P., & Borghetti, B. J. (2012). A review of anomaly detection in automated surveillance. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *42*(6), 1257-1272.

Souly, N., Spampinato, C., & Shah, M. (2017). Semi and weakly supervised semantic segmentation using generative adversarial network. *arXiv preprint arXiv:1703.09695*.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 4489-4497).

Wu, S., Moore, B. E., & Shah, M. (2010, June). Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. In *2010 IEEE computer society conference on computer vision and pattern recognition* (pp. 2054-2060). IEEE.

Xia, L., Gori, I., Aggarwal, J. K., & Ryoo, M. S. (2015, January). Robot-centric activity recognition from first-person rgb-d videos. In *2015 IEEE winter conference on applications of computer vision* (pp. 357-364). IEEE

Xiang, T., & Gong, S. (2005, October). Video behaviour profiling and abnormality detection without manual labelling. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1* (Vol. 2, pp. 1238-1245). IEEE.

Xiang, T., & Gong, S. (2008). Incremental and adaptive abnormal behaviour detection. *Computer Vision and Image Understanding*, *111*(1), 59-73.

Xu, D., Ricci, E., Yan, Y., Song, J., & Sebe, N. (2015). Learning deep representations of appearance and motion for anomalous event detection. *arXiv preprint arXiv:1510.01553*.

Ullah, I., & Petrosino, A. (2015). A strict pyramidal deep neural network for action recognition. In *Image Analysis and Processing—ICIAP 2015: 18th International Conference, Genoa, Italy, September 7-11, 2015, Proceedings, Part I 18* (pp. 236-245). Springer International Publishing.

Zaharescu, A., & Wildes, R. (2010). Anomalous behaviour detection using spatiotemporal oriented energies, subset inclusion histogram comparison and event-driven processing. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part I 11* (pp. 563-576). Springer Berlin Heidelberg.

Zhou, S., Shen, W., Zeng, D., Fang, M., Wei, Y., & Zhang, Z. (2016). Spatial–temporal convolutional neural networks for anomaly detection and localization in crowded scenes. *Signal Processing: Image Communication*, *47*, 358-368.