Inteligencia Artificial 28(75) (2025), 281-297
doi: 10.4114/intartif.vol28iss75pp281-297

INTELIGENCIA ARTIFICIAL

http://journal.iberamia.org/

ABC-WSVAD: Swarm Optimization for Weakly-Supervised Video Anomaly Detection

Iman Mostafa [1, *], Marwa Gamal [1], Rehab F. Abdel-Kader [2], Khaled Abd El Salam [1,3]

[1] Electrical Engineering Department, Suez Canal University, Ismailia, Egypt.
[2] Electrical Engineering Department, Port said University, Port said, Egypt.
[3] Computer Science Department, College of Information Technology, Misr University for Science and Technology (MUST), 6th of October City 12566, Egypt.
* Corresponding author: Iman Mostafa (e-mail: iman.mostafa@eng.suez.edu.eg).

Abstract Nowadays, Video Anomaly Detection (VAD) has undergone a significant transformation due to advancements in Deep Learning (DL) and Computer Vision (CV). VAD holds substantial importance in various applications, particularly security, given the increasing spectrum of criminal activities. Conventional supervised anomaly detection techniques heavily rely on meticulously labeled data, which is time-consuming to annotate and is the basis for training anomaly classifiers. However, assembling extensive annotated datasets for VAD poses challenges due to the hard and time-consuming of the task. Weakly Supervised (WS) approaches reduce reliance on fully labeled data by utilizing alternative supervision sources or weak labels for training anomaly detection models. Despite their promise, many WS methods experience prolonged processing times, critical for timely crime detection. The main objective of this research is to outperform previous work in terms of performance and training time. In this paper, we propose the utilization of Artificial Bee Colony (ABC) optimization for training a fully connected neural network. This method efficiently searches for optimal weight configurations to minimize the error or loss function. Our approach is rooted in Multiple Instance Learning (MIL), tailored specifically for Weakly Supervised Video Anomaly Detection (WSVAD). The presented method was tested on a set of 1900 videos sourced from the comprehensive UCF-Crime dataset, utilizing the Inflated 3D ConvNet (I3D) feature extractor. The experimental outcomes highlight the superiority of the proposed ABC-WSVAD algorithm compared to other methods. Notably, the methodology outperforms baseline approaches by a 4.36% increase in the Area Under the Curve (AUC), underscoring its superior effectiveness in anomaly detection.

Keywords: Video Anomaly Detection, Weakly-Supervised anomaly detection, UCF-Crime, swarm optimization, Artificial Bee Colony Optimization.

1 Introduction

Video Anomaly Detection (VAD) is a research field that focuses on detecting anomalous events or activities in surveillance videos. VAD is intensively studied due to its potential applications in many areas such as healthcare [1], [2], [3], IT security [4], [5], [6], and video surveillance. The anomalies that occur in the real world are diverse and complicated. It is challenging to compile a list of all possible anomalous events. As a result, it is preferable if the algorithm for detecting anomalies does not rely on any prior knowledge of the events. To put it another way, anomaly detection should be carried out under minimal supervision.

ISSN: 1137-3601 (print), 1988-3064 (on-line)

The general process of VAD is shown in Figure 1, where data are loaded and pre-processed. Then, features are extracted by different techniques such as 3D Convolutional Network (C3D) [7], Inflated 3D ConvNet (I3D) [8], or VideoSwin [9]. The extracted features are then trained by the chosen deep learning model depending on the project objective which generates a score to classify data as normal or abnormal. In the context of the going beyond approaches, weakly supervised techniques have gained great attention in recent years due to their great achievements in the VAD field.

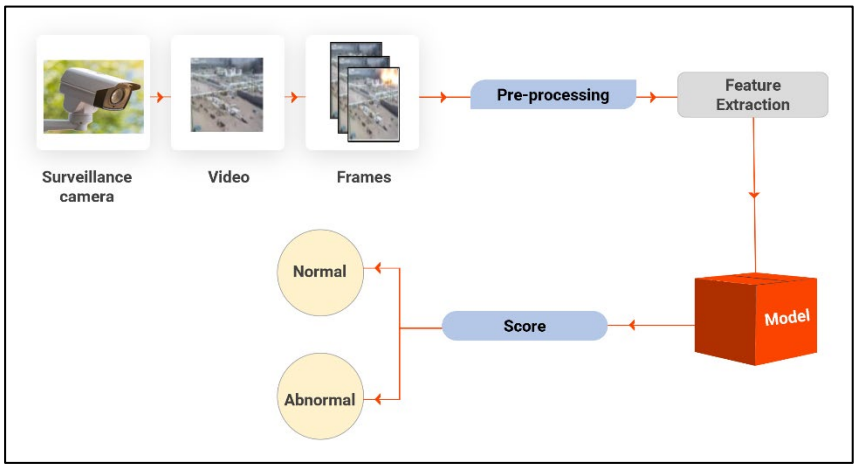

Figure 1. General Process of Video Anomaly Detection

Weakly Supervised Anomaly Detection (WSAD) in video surveillance faces challenges due to the limited availability of labeled training data. This can prompt overfitting on the labeled data and inadequate generalization to new data [10], [11]. Another difficulty is determining what constitutes an anomaly in video data. Rare events or behaviors that are difficult to objectively define are examples of anomalies [12]. This can complicate the process of generating accurate labels for training data.

The abovementioned challenges are partially solved with the Multiple Instance Learning (MIL) method which was first proposed by Sultani et al. [13]. In MIL, the training data are stored in bags, with multiple instances in each bag. Normal and anomaly videos are considered as bags designed for a network that processes video clips independently from each other. Bags that have at least one abnormal snippet are considered positive bags, while the bag that has only normal snippets are negative. Learning a classifier that can classify bags rather than individual instances is the objective of MIL. The main obstacle in MIL is that the labels of the instances within each bag are unknown, and only the bag label is provided. This means the classifier must learn to recognize the relevant instances within each bag that contribute to the bag's label while ignoring irrelevant or noisy instances.

Although Sultani et al. [13] has made a significant contribution to weakly supervised learning in VAD, it also has some limitations that are worth critiquing. Depending on the use of C3D for feature extraction, it has notable limitations compared to more advanced models like I3D. C3D does not leverage the advantages of pre-training on large, diverse video datasets like I3D, which allows the I3D to generalize more effectively to various video content, including rare and subtle anomalies. This makes C3D less adaptable and potentially less accurate in real-world surveillance scenarios.

Optimization methods have been applied widely before in many fields such as machine learning, data science, engineering, and many other fields. These methods aim to determine the best values for parameters, weights, or configurations that result in the best possible solution to a given problem. They play an important role in decision-making and problem-solving processes by automating the search for the best solution in complex scenarios where an exhaustive search is not possible. One of these methods is swarm optimization algorithms which are inspired by natural swarms' collective behavior, such as bird flocks, fish schools, and insect colonies such as Particle Swarm

Optimization (PSO) [14], Artificial Bee Colony (ABC) [15], Ant Colony Optimization (ACO) [16], Firefly Algorithm (FA) [17] and Bat Algorithm (BA) [18].

Our contribution is introducing ABC optimization during the training phase of a weakly supervised anomaly detection algorithm for VAD, particularly on a sizable dataset. The ABC algorithm optimizes the initial weights of a straightforward five-layer neural network, enabling faster convergence and improved generalization, which are crucial when working with limited or weakly labelled data. This approach demonstrates that complex models are unnecessary to achieve superior accuracy compared to the previous state-of-the-art. By leveraging optimized initial weights, the training process becomes more efficient in terms of computational resources and performance, even with challenges posed by scarce labeled data.

The proposed method has practical applications in real-world surveillance systems, particularly in public security and crime prevention. It can be deployed for monitoring urban environments, transportation hubs, and critical infrastructure where detecting anomalies in video feeds is crucial for ensuring public safety. Additionally, the model could be applied in healthcare for monitoring patient activities and detecting abnormal behaviors, as well as in industrial settings for detecting equipment failures or safety violations through video data. The method may struggle with ambiguous or context-dependent anomalies that require a deeper semantic understanding beyond the current feature extraction and MIL framework. The model's performance depends on the quality and generalization ability of the pre-trained I3D feature extractors. Poorly pre-trained extractors can limit effectiveness. Moreover, Although ABC optimization reduces training iterations, applying the method to very large-scale datasets may still require significant computational resources.

The rest of the paper is organized as follows: Section 2 presents the state of the presents state-of-the-art concerning different methods of detecting anomalies in videos. The proposed method is described in Section 3. Section 4 shows the experimental results. Finally, Section 5 provides the conclusion of the paper and highlights directions for future work.

2 Related work

Recently many contributions have been introduced to the VAD field. In this section, we will show diverse techniques applied in this field such as traditional methods, weakly supervised methods, transformers, memory methods, language models, and swarm optimization methods.

Before the evolution of deep learning algorithms, traditional methods were taking a great role in anomaly detection using unsupervised methods. These methods depend on the availability of normal videos only, and the anomalous ones are considered as an outlier detection problem. Hand-crafted features were used to utilize the problem of one-class classification [19], [20], [21]. Lu et al. [22] proposed a dictionary-based approach to learn normal behaviors and employed reconstruction errors to identify anomalies. These methods work well in controlled settings but struggle with scalability and robustness when applied to large-scale video data. The large volume and high variability of such datasets create significant obstacles, typically resulting in reduced performance and reliability in real-world surveillance applications. In addition, these approaches depend on complex preprocessing steps and sometimes lead to overfitting.

Most of the recent work in VAD used WSAD with MIL following Sultani et al. [13]. Zhong et al. [23] introduced a novel method for using an action classifier and Graph Convolutional Networks (GCN) as noisy label cleaners. They signified that during training MIL methods suffered from error propagation. This approach managed to overcome the issue of having video-level labeling in the UCF-Crime dataset. It converted the problem into a direct classification task based on a cross-entropy function and a temporal-ensembling strategy. Although this method gives better performance, it is computationally expensive and the model’s ability to generalize remains unsatisfactory and is restricted to simple scenes. Zhang et al. [24] adopted the approach in [13] as their baseline and introduced a new inner bag loss (IBL) to reduce the gap between the lowest and highest scores in the negative bag while increasing it in the positive one. They replaced the first fully connected layer (FCN) of [13] with a temporal convolution network (TCN) [25] to connect between the preceding and the current information of the instance

followed by two fully connected layers. Zhu et al. [26] adopt the model of Sultani et al. by adding an attention block [27]. They made use of the PWC-Net [28] to extract the motion-aware features. Tian et al. [29] introduced Robust Temporal Feature Magnitude learning (RTFM) which learns the temporal features to capture long- and short-term temporal dependencies to understand the feature magnitude more accurately. Neglecting other video attributes, RTFM results in an amplification of abnormal feature magnitudes and a reduction in the magnitudes of normal features. To address this issue, Chen et al. [30] introduced a Magnitude-Contrastive Glance-and-Focus Network (MGFN) for anomaly detection to address the issue in Tian et al. [29]. This method pushes the magnitude of abnormal features to be larger and the normal ones to be smaller without considering other video attributes. Although it showed promising results on benchmark datasets, it is computationally costly and needs high processing. Yang et al. [31] presented a novel pseudo-label generation and self-training framework called Text Prompt with Normality Guidance (TPWNG) for WSVAD. This method leverages the pre-trained CLIP [32] model's language-visual knowledge to align video event descriptions with frames, generating pseudo-labels. However, the quality of the pseudo-labels produced determines the model's accuracy. Performance may suffer if pseudo-label mistakes spread throughout the self-training process.

Some papers used transformers to address the VAD problems inspired by its great achievements in Natural Language Processing (NLP). Yuan et al. [33] make a combination between the U-Net [34] and the Video Vision Transformer (ViViT) [35] to capture broader global contexts and more detailed temporal information. They named their model TransAnomaly, which is a prediction-based VAD method. In addition, the model can execute anomaly localization. Feng et al. [36] proposed a model based on a Convolution Transformer (CT) with dual discriminator GAN (D2GAN). They developed a new self-attention module that is focused on spatio-temporal modeling in video sequences. The CT is capable of encoding temporal information efficiently in a sequence of feature maps and the D2GAN was developed to enhance the prediction of future frames using the Wasserstein GAN with gradient penalty (WGAN-GP) [37]. Li et al. [38] created a Transformer-based Multi-Sequence Learning (MSL) network to learn video-level anomaly probability as well as snippet-level anomaly scores.

Transformers excel at capturing long-range temporal dependencies in video sequences, which is critical for identifying anomalies that occur over time. However, they require significant computational resources, especially for long video sequences, making them impractical for real-time applications when resources are limited.

Guo et al. [39] combined discriminative and generative models with dual memory modules to address data imbalance in VAD. The dual memory module obtains sparse feature representations in both normality and abnormality spaces, improving detection performance. Zhou et al. [40] introduced dual memory units with separate normal and abnormal memory banks to improve the distinction between them. This approach addresses the limitation of previous methods, which focused solely on extracting anomaly data representations without accounting for the influence of normal data. Wang et al. [41] proposed a Dual-Stream Memory Network (DSM-Net) for anomaly detection that takes advantage of historical data to extract spatial-temporal correlations between video events.

Some recent papers comprised language models in VAD. Lv et al. [42] proposed integrating video-based large language models (VLLMs) to eliminate thresholding and enhance explainability. They introduced a Long-Term Context (LTC) module and a three-phase training method to improve VLLM fine-tuning efficiency and reduce data annotation costs. Zanella et al. [43] introduced Language-based VAD (LAVAD), a novel, training-free method for VAD that leverages pre-trained large language models (LLMs) and vision-language models (VLMs). The scalability and high computational cost associated with training large language models can present significant challenges, especially when it comes to real-time anomaly detection in large-scale systems.

Few swarm optimization techniques were previously used in the VAD field. It was first introduced in this field by Vagia et al. [44]. They combined swarm intelligence and histograms of oriented gradient descriptors to form a new feature capable of determining normal regions using a Support Vector Machine (SVM) [45] framework. Raghavendra et al. [46] proposed a method for detecting global anomalies in crowded scenes that use PSO to optimize the interaction force computed by the Social Force Model (SFM). Radha et al. [47] proposed an efficient anomaly detection system for video surveillance applications that use Wireless Visual Sensor Networks (WVSNs).

It detects anomalies in video data using compressed sensing and PSO and transmits the necessary measurements to the network operator. Qasim et al. [48] used ACO clustering algorithm for abnormal event detection in crowded environments in surveillance videos. Priyadharsini et al. [49] built a hybrid DL system based on a pre-trained CNN and a One-class SVM where improved PSO isolates the most salient regions in the video frames. Alsolai et al. [50] proposed a vision-based anomaly system based on the EfficientNet [51] with an Improved Chicken Swarm Optimizer (ICSO) [52] to detect and classify anomalies to assist visually impaired people. Kumar and Rani [53] applied Multi-Feature Tensor Subspace Learning and Robust Principal Component Analysis for feature extraction while PSO-based CNN for anomaly detection. Other research applied swarms for anomaly detection but on different occasions not in videos. Qureshi et al. [54] introduced a new intrusion detection system based on a random neural network and an artificial bee colony algorithm (RNN-ABC). Bekri et al. [55] built an unsupervised classifier for anomaly detection in an intelligent irrigation control system by combining PSO and clustering methods. This paper concentrated on applying ABC optimization using a weakly supervised anomaly detection method.

3 The Proposed Method

Our model aims to achieve better performance than previous work in the least amount of time. Our proposed framework is divided into 3 phases as shown in Figure 2. The First phase is the feature extraction where features are extracted from a pre-trained Inflated 3D ConvNet (I3D) model. The second phase is applying ABC optimization on the weights through the training of the neural network. The last one is performing the Loss function for anomaly detection. Following Sultani [13], where normal and abnormal videos are considered as bags (positive and negative), videos are segmented into snippets and considered instances in MIL.

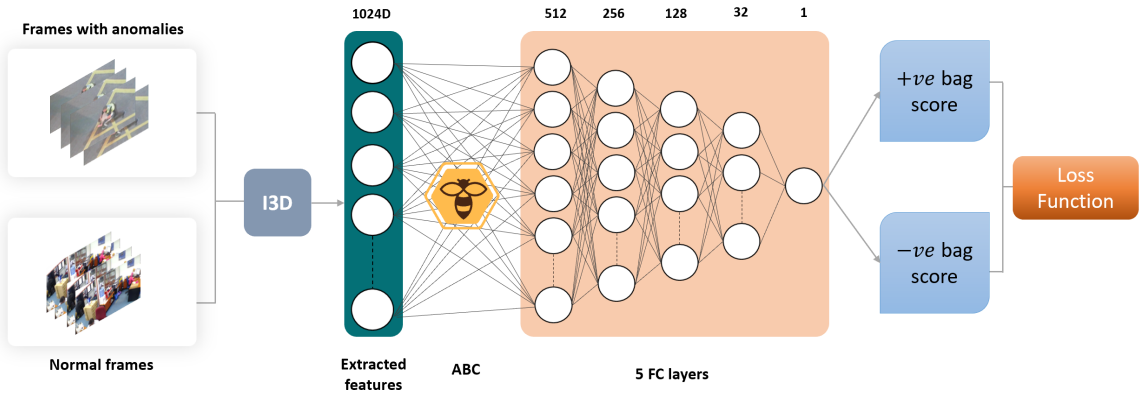

Figure 2. The overall framework of the proposed method. For feature extraction, I3D extractor is used which generates 1024D features. After extracting the features, artificial bee colony optimization is applied to train a fully connected neural network of five layers. Finally, the ranking loss function is computed between the highest score instances in the negative (-ve) and positive (+ve) bag scores

3.1 Feature Extraction

For feature extraction, we employ an I3D feature extractor. It is typically pre-trained on large-scale video datasets known as Kinetic dataset which offer excellent generalization capabilities. I3D is an extension of 2D ConvNets which enables it to learn features from both spatial and temporal video dimensions. Moreover, it achieves better results than the C3D on several state-of-the-art performances as it is deeper which enables it to learn more complex features from videos. In the feature extraction process, the videos are divided into a fixed number of frames T each of size (H × W). As a result, the input dimension in the training process for a batch size B is (B × T × F) where F is the number of features extracted from each frame (feature_size).

3.2 Artificial Bee Colony

The Artificial Bee Colony (ABC) algorithm simulates the intelligent behavior of honeybee foraging. It comprises employed bees, onlooker bees, and scout bees, each playing distinct roles in searching for the best solution. Employed bees visit existing food sources, onlooker bees observe the dance ceremony to select the next food source based on the performance of employed bees, and scout bees randomly pick new food sources.

In the context of optimizing the neural network weights, each potential solution corresponds to a set of weights in the neural network. These weights need to be encoded in a way that can be manipulated by the ABC algorithm. The ABC algorithm starts by initializing a population of food sources (solutions) representing different sets of weights.

$$Population = \{W_1, W_2, \dots\dots\dots, W_{population\_size}\} \tag{1}$$

After that, the fitness function would evaluate the neural network performance with the given set of weights. This fitness function guides the search for optimal solutions of weights. In our algorithm, the AUC is applied as the objective metric for optimization. The objective of the fitness function is to find the optimal weights that maximize the AUC during the optimization process. By maximizing the AUC, the ABC algorithm aligns the optimization process with the evaluation criteria, which is particularly advantageous in imbalanced datasets like UCF-Crime. This approach ensures that the trained neural network can distinguish between true positives and false positives for better anomaly detection.

$$Fitness(W_i) = \max\,(AUC) \tag{2}$$

The employed bees are responsible for exploring neighboring weights (Wi') and adjusting them based on their fitness values.

$$Wi' = W_i + \phi_i \cdot (W_i - W_j) \tag{3}$$

Where i ≠ j and ϕi is a random value between [−1,1] and Wi' the new solution. The onlooker bees select sets of weights with better performance $F(\theta_i)$ and explore modifications to those weights based on the following probabilistic function for choosing the food sources.

$$P(W_i) = \frac{F(\theta_i)}{\sum_{k=1}^{S} F(\theta_k)} \tag{4}$$

Where S is the number of food sources, θi is the ith food source or solution, and $\sum_{k=1}^{S} F(\theta_k)$ is the summation of the fitness values for all possible solutions or weights.

The scout bee identifies weight sets that have not improved after a certain number of iterations and replaces them with new randomly generated weight sets with the predefined limits specified by the search space limits [Wmin, Wmax]. Throughout the iterations, the ABC algorithm keeps track of the best solution (set of weights) found during optimization. Then, the algorithm continuously assesses the neural network's fitness using various weight sets, updating it if a particular set of weights enhances performance. Finally, the optimization process continues until a stopping criterion is met, which could be reaching a satisfactory level of performance or a maximum number of iterations.

ABC, like other optimization algorithms, aids in increasing the rate of convergence and preventing the neural network from becoming stuck in local minima. It offers a global search capability, allowing the neural network to explore a larger solution space, improving performance and generalization on complex tasks. Algorithm 1 shows the pseudocode for applying the ABC algorithm on a neural network.

Algorithm 1 ABC Pseudocode

1. **Inputs**= Training dataset
2. **Output**= best weights values
3. **Initialize**:
    - population size =200, maximum number of iterations =50, search space for neural network weights [W_{min}=-1, W_{max}=1]
4. **For each** record:
5. Generate the initial population by scout bees
6. Compute the fitness function **(2)**
    - Train the neural network with the current weights.
    - Calculate AUC as the fitness value for the current solution.
7. Set iteration to 1
8. # **Employee bees Phase**
9. **For each** employee bee :
10. Select a new solution from neighbor data
11. Modify the weights of the current bee **(3).**
12. Evaluate the fitness of the modified weights
13. **If** (the modified weights lead to better fitness) **then**
14. update the bee's position
15. **# Onlooker bees phase:**
16. **For each** onlooker bee:
17. Select a solution depending on the probability values $p(w_i)$ for the solution **(4).**
18. Modify the weights of the onlooker bee.
19. Evaluate the fitness of the modified weights
20. **If** the modified weights lead to better fitness **then**
21. update the bee's position
22. **# Scout bees phase:**
23. If (food source has exceeded the trial limit) **then**
24. Abandon the food source
25. Replace these bees with new random solutions within the weight_range [-1,1]
26. Memorize the best solution from all the solution
27. **If** the fitness of the new weights is better **then**
28. update the population.
29. iteration = iteration + 1
30. **Return** the best solution found.

3.3 Loss Function

The loss function (MIL ranking loss) applied is of Sultani et al. [10] model for the anomaly detection process using MIL annotations. The MIL ranking loss function has demonstrated its efficacy in real-world applications for VAD, yielding cutting-edge outcomes across multiple benchmark datasets. It can handle the weakly supervised tasks that gain knowledge from the distribution of anomaly scores of individual video segments. In addition, it is robust to add noise, which is crucial for VAD because anomalous events are frequently uncommon and surrounded by regular occurrences. Since the video segment level annotations are unknown, the most straightforward approach would be to use a MIL ranking loss (see (5)). It is based on separating positive (anomalous segments) and negative (normal segments) instances in terms of anomaly score. By using (6) and (7), two regularization terms are added: the smoothness term and the sparsity term respectively. The smoothness constraint ensures that anomaly scores vary smoothly across video segments while the sparsity constraint ensures that only a few segments in the anomalous bag have high anomaly scores.

The smoothness constraint ensures that anomaly scores vary smoothly across video segments. On the other hand, the sparsity constraint ensures that only a few segments in the anomalous bag have high anomaly scores.

$$L(X_a, X_n) = \max(0,1 - \max f(U_a^i) + \max f(U_n^i)) \quad (5)$$

Where X_a represents a positive bag of a positive video and X_n is a negative bag of a negative video, U_a and U_n represent the abnormal and the normal video segments respectively, $f(U_a^i)$ and $f(U_n^i)$ expresses the associated predicted scores, respectively.

$$Smoothness = \sum_i^{n-1}(f(U_a^i) - f(U_a^{i+1}))^2 \quad (6)$$

$$Sparsity = \sum_i^n f(U_a^i) \quad (7)$$

Then two penalty terms are added to the smoothness (λ_1) and the sparsity constraints (λ_2) respectively. λ_1 enforces sparsity in anomaly scores, emphasizing fewer but stronger signals to better identify clear anomalies. λ_2 stabilizes the model and reduces sensitivity to noise. After adding these terms, the final loss (L_f) will be as the following equation [13] :

$$L_f = max(0,1 - max f(U_a^i) + max f(U_n^i)) + \lambda_1 \sum_i^{n-1}(f(U_a^i) - f(U_a^{i+1}))^2 + \lambda_2 \sum_i^n f(U_a^i) \quad (8)$$

In conclusion, After the ABC algorithm finds the best weights by maximizing AUC, the MIL ranking loss is used to improve the model further. The MIL loss separates positive and negative bags based on their anomaly scores, adding more detailed refinement. In the final evaluation phase, AUC is used as a performance metric to compare our method with other state-of-the-art approaches. This distinction is crucial because the first use of AUC is internal to the optimization process, while the second use is external, providing a standardized measure of model performance.

4 Experimental Results

4.1 System Properties

The proposed method has been trained and tested using Windows 10 with 64-bit operating system, Intel(R) Core(TM) i7-10610U CPU @ 1.80GHz, 2304 Mhz, 4 Cores, 8 Logical Processor(s). The installed RAM is 16.0 GB. Jupyter Lab Notebook is used for coding.

4.2 Dataset

In this paper, our method is conducted on one of the most popular and public datasets which is UCF-Crime dataset [13][1]. It is a large-scale dataset with 1900 long real-world surveillance videos with more than 47000 incidents of a total of 128 recorded hours. Thirteen anomalies are included, which are explosion, fighting, robbery, shooting, abuse, arrest, arson, assault, road accident, stealing, shoplifting, and vandalism as shown in Figure 3. It can be used for two tasks: 1) general anomaly detection by considering normal videos in a group and abnormal videos in another group, and 2) recognizing only the type of the 13 different anomalies mentioned above.

[1] Available at: https://www.dropbox.com/sh/75v5ehq4cdg5g5g/AABvnJSwZI7zXb8_myBA0CLHa?dl=0

UCF-Crime is split into 1610 videos for training, where 800 are normal and 810 are abnormal videos, and 290 for testing where 150 normal and 140 abnormal with a resolution of 240 × 320. For this benchmark, video-level labels are only provided for the training videos, and the temporal annotation is for the testing set. Many researchers prefer to make use of this dataset because the videos are from real-world scenes and the variety of anomalies it has [13], [29], [56], [24], [57]. The problem with this dataset is the varying duration of the clips and some clips have been repeated. Moreover, high-processing hardware is needed for extracting the features.

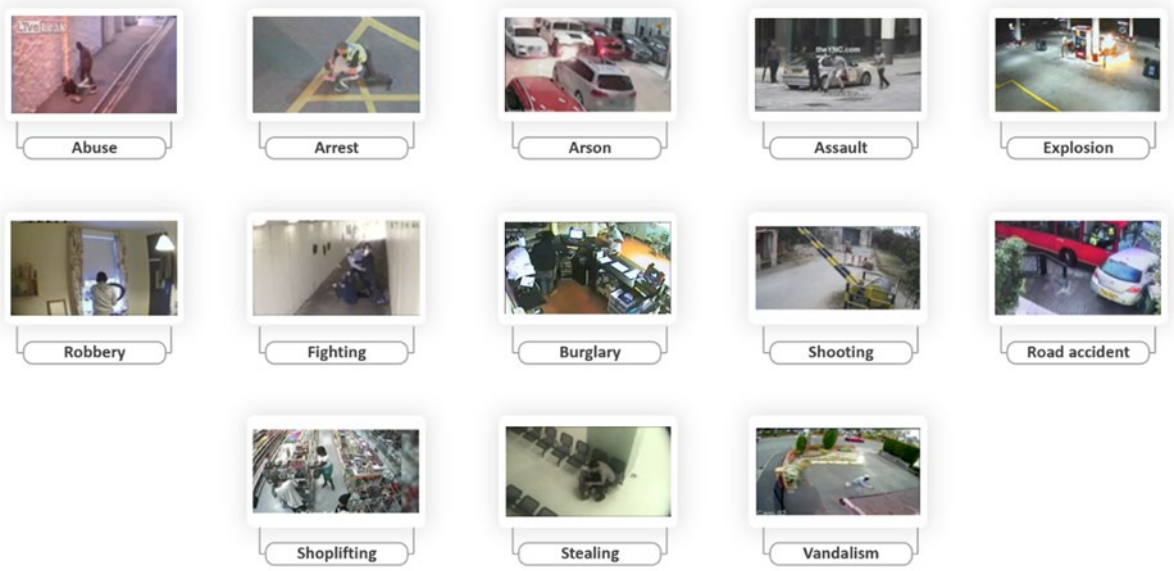

Figure 3. Samples of 13 abnormal snippets in the UCF-Crime dataset

Table 1: Comparison Between Various Anomaly Datasets With the UCF-CRIME

Dataset		Year	# Videos	# Frames	Scenes	# Of Anomaly types	Clip duration
UMN [58]		2006	11	7700	3	1	-
Subway [59]	**Entrance**	2008	1	72,401	1	5	2 hr.
	Exit		1	136,524	1	3	
UCSD [60], [61]	**Ped 1**	2010	80	14,000	1	5	-
	Ped 2		26	4,560	1	5	
Avenue [62]		2013	37	30652	1	3	1-2 min
ShanghaiTech [63]		2017	437	317,398	13	130	-
UCF-Crime [13]		2018	1900	13M	20	13	128 hrs. (total)
UCFCrime2Local [64]		2019	300	-	-	6	>1 hour
Street scene [65]		2020	81	203,257	1	17	-
TIMO [66]		2021	1588	612,000	2	-	-
UBI-Fights [67]		2021	1000	-	Multiple scenes	1	80 hrs. (total)
UBnormal [68]		2022	543	236,902	29	22	2.2 hrs (total)

Table 1 presents a comprehensive comparison of various anomaly detection datasets, highlighting key attributes such as the year of release, the number of videos, the total number of frames, the number of scenes, anomaly types, and average clip duration. This table aims to illustrate the diversity and scale of datasets available for anomaly detection, emphasizing the UCF-Crime. This comparison highlights the differences in dataset characteristics, which are crucial for understanding the challenges of different datasets in this domain. Moreover, it shows that UCF-Crime surpasses the frame numbers, the types of scenes, and the anomaly types dataset making it a robust benchmark for evaluating VAD methods. Furthermore, it is distinguished by the presence of realistic world scenarios, not synthetic data.

4.3 Evaluation Metrics

The proposed method was evaluated using the Area Under the Receiver Operating Characteristic curve (AUC-ROC) and four other measures (precision, recall, accuracy, and F1 score) on the UCF-Crime dataset.

The AUC-ROC curve is widely used to assess the performance of anomaly detection models because of their effectiveness in capturing various aspects of model performance. It is used particularly in imbalanced datasets and anomaly detection tasks. This metric is computed using the test's frame-level ground truth annotation. The larger the AUC, the better the model's performance.

In addition, the abovementioned four metrics (9), (10), (11) ,and (12) are empolyed to compare the performance of the proposed optimized model with ABC and without optimization where the equations are as follows:

$$\text{Precision} = \frac{TP}{TP+FP} \tag{9}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{10}$$

$$\text{F1-Score} = \frac{2\,x\,TP}{2\,x\,TP+FN+FP} \tag{11}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{12}$$

Where TP (True Positive) is the number of correctly classified anomaly events, FP (False Positive) is the number of events where the model incorrectly identifies a normal event as an anomaly, TN (True Negative) is the number of normal events correctly classified, and FN (False Negative) is the number of events where the model incorrectly identifies an anomalous event as normal.

Some research suggests using the Region-Based Detection Criterion (RBDC) and the Track-based Detection Criterion (TBDC) to supplement the AUC measure. However, these two measures are ineffective in the weakly-supervised detection.

As a result, in our ABC-WSVAD method, the existing approaches are followed by employing the AUC-ROC metric for evaluation on the UCF-Crime dataset.

4.4 Implementation Details

Pytorch framework [69] is used in our implementation. Following [13], each video is segmented into 32 video segments (T=32). The pre-trained I3D extractor is used for extracting features (1024D). These features are then fed to five Fully Connected Network (FCN) layers with 512, 256, 128, 16, and 1 nodes respectively. The dropout between layers is 0.4 and the activation function used after each layer is Relu except for the last layer, the sigmoid

function is applied. For training the network, the Adam optimizer [70] is applied with a learning rate of 0.0005 and weight decay of 0.0005. The training epochs are set to 50 with batch size 32 for each randomly selected normal and abnormal video.

For ABC hyperparameters, the colony size is adjusted to 200 particles, scout to 40% of the colony size, and max iterations are set to 50. The weight bounds were adjusted between -1 and 1. Figure 4 illustrates the movement of ABC particles in the optimization process, where each particle (black dots) represents a potential weight set. The particles explore and exploit the search space to find the mejores weights. The figure visually tracks these movements (blue lines), highlighting how particles iteratively converge toward an optimal set of weights that minimize the error, thus improving the model's performance.

Moreover, we examined applying 4 layers, 6 layers and 7 layers NN with different dropouts and different optimizers such as RMSprop and Adadelta but the accuracy was not satisfactory. In addition, PSO is examined on the same model and it gives less accuracy by 1%.

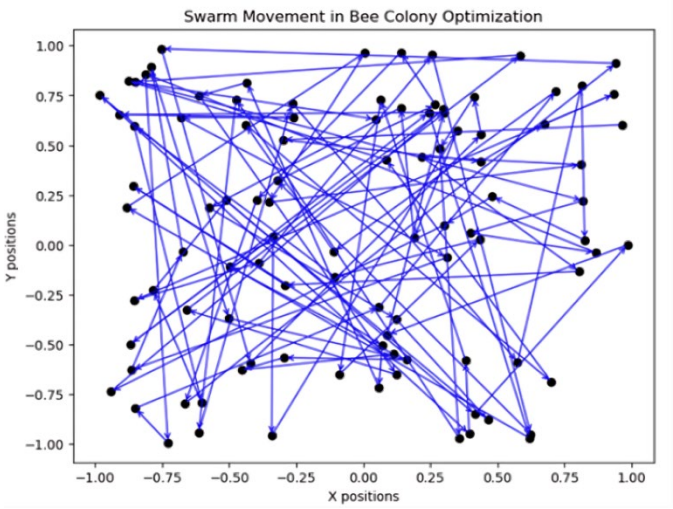

Figure 4. Movement of ABC particles during searching for the optimal weights. Black dots are the swarm particles, while the blue arrows show the direction of these particles.

4.5 UCF-Crime Results

To ensure a fair and transparent evaluation, the performance of the proposed ABC-WSVAD model was compared with several state-of-the-art methods on the UCF-Crime dataset, as shown in Table 2. The comparison included two approaches: results directly reported in published papers and custom implementations of methods. The custom implementations, namely “Proposed ABC-WSVAD” and “WSVAD (without optimization),” were benchmarked under identical conditions, while the results for other methods were drawn directly from their respective published papers. Table 2 provides a comparative analysis of the Area Under the Curve (AUC) performance of various methods applied to the UCF-Crime dataset, categorized by their supervision level: unsupervised and weakly-supervised approaches. It highlights key attributes such as the features extracted, AUC percentage, and the number of iterations, offering a comprehensive overview of the methods' effectiveness and efficiency.

As shown in Table 2, our result achieves better performance compared with all the previous un-supervised approaches such as Basic One-class Discriminative Subspaces (BODS) [20], Generalized One-class Discriminative Subspaces (GODs) [20] Generative Cooperative Learning (GCL) [71].

Moreover, for weakly supervised supervision, our technique surpasses the results of all previous methods using the I3D-RGB features that is used in our method such as Multiple Instance Learning (MIL) [13], Graph Convolutional Label (GCN) [23], Inner Bag Loss (IBL) [24], Motion Aware (MA) [26]. Our model (ABC

optimization with I3D extractor) achieved 79.45% which is higher than the original paper of Sultani et al. [13] which achieved 75.41% when using the C3D extractor. Moreover, Tian et al. [29] experimented the same model of Sultani et al. [13] but by applying the I3D extractor instead of C3D and achieved only 77.92%. Although the AUC of the MA [26] method and our proposed method are very similar, MA requires significantly more iterations to achieve its results. This difference is critical for real-time applications, where faster processing and efficiency are essential. Figure 5 shows the ROC performance of the proposed method on the UCF-Crime dataset using the ABC-WSVAD approach compared with other methods such as Lu et al.[62], MIL[13], and MA [26].

Our goal differs from other models that prioritize accuracy alone. We aim to achieve high accuracy while minimizing training time. In Table 2, the proposed ABC-WSVAD method achieves a competitive AUC of 79.77% with only 26 iterations, compared to other methods that require significantly more iterations, such as GCN [23] with 20,000 iterations and MA [26] with 50,000 iterations. While our AUC value is close to the MA method, our model surpasses in other aspects such as the simplicity of our architecture makes it more scalable and easier to deploy in real-world scenarios. The reduced training iterations and computational requirements of our method make it more scalable for large-scale datasets and real-time applications.

Moreover, after testing, we found that using ABC optimization significantly accelerates the training process after finding the mejores' initial weights compared to using the same model without optimization as shown in Table 3. Unlike some models, which take longer to converge to optimal weights, our approach achieves desirable results more swiftly.

Table 2: Comparison of AUC performance on UCF-Crime with various works.
* indicates it is retrained by Tian et al. [29]

Supervision	Method	Features Extracted	AUC %	No. of iterartions
Un Supervised	Lu et al.[62]	-	65.51	
	BODS [20]	C3D-RGB	68.26	-
	GODS [20]	I3D-RGB	70.46	-
	GCL [71]	I3D-RGB	74.20	1587
Weakly-Supervised	MIL[13]	C3D-RGB	75.41	1000
	MIL *[13]	I3D-RGB	77.92	-
	GCN [23]	TSN-Optical Flow	78.08	20,000
	IBL [24]	C3D-RGB	78.66	-
	MA [26]	PWC-flow	79.00	50,000
	Proposed ABC-WSVAD	**I3D-RGB**	**79.77**	**26**

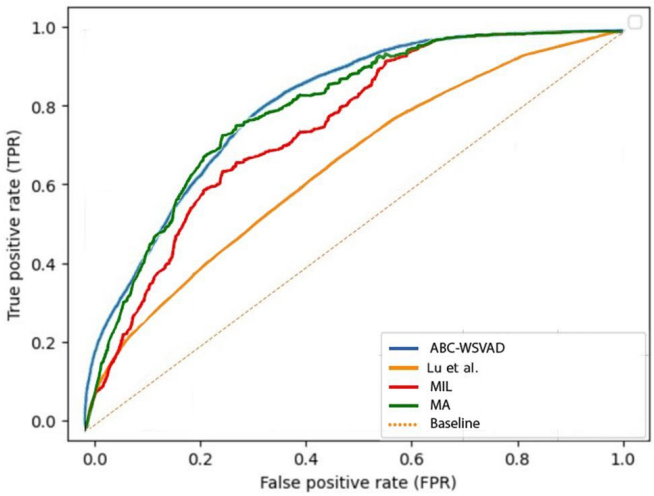

Figure 5. ROC Performance for different methods on UCF-Crime dataset.

Table 3 shows the performance comparison between the WSVAD and ABC-WSVAD models across four evaluation metrics: Sensitivity/Recall, Precision, F1 Score, and Accuracy. The best results in each metric are highlighted in bold. Based on the metrics provided in this table, ABC-WSVAD (model with optimization) is generally better than WSVAD (model without optimization) across most metrics. It has a higher recall of 2%, meaning it is more effective at detecting anomalies. Also, it has a slightly lower precision but maintains a higher overall accuracy of 2%. In addition, the F1 score, which balances both precision and recall, is somewhat better for ABC-WSVAD by 1%, suggesting it offers a better overall performance. Figure 6 visualizes this comparison, showing ABC-WSVAD's higher recall, accuracy, and F1 score, confirming its superior performance.

Table 3: Performance comparison of WSVAD and ABC-WSVAD
Best results are indicated In BOLD

Model	Sensitivity/Recall (%)	Precision (%)	F1 Score (%)	Accuracy (%)	No. of iterations
WSVAD	82	**90**	85	82	49
ABC-WSVAD	**84**	89	**86**	**84**	**26**

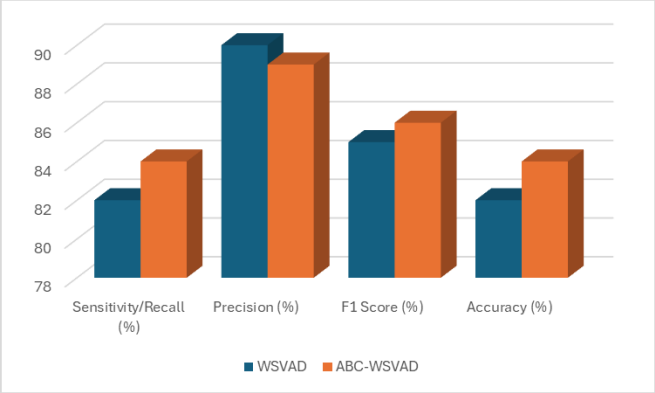

Figure 6. Performance Metrics Comparison of WSVAD and ABC-WSVAD

In our research, the study focused not only on achieving high accuracy but also on experimenting with processing time, as previous state-of-the-art models did not emphasize their processing times in their papers. That's why the training time is examined in our own implementations: one without using optimization and the other with using optimization. The stated accuracy is achieved after only 26 iterations, whereas using the same model without optimization took 49 iterations to reach the same accuracy.

Moreover, unlike methods that require extensive training time, our model achieves an AUC of over 78% in its initial iterations. This is crucial for scenarios requiring quick decision-making or real-time processing. A higher AUC signifies better performance, particularly in classification tasks, indicating our model's effective ability to distinguish between positive and negative instances.

5 Conclusion

Video anomaly detection has become an important field, especially in the last years due to the increased rate of crimes. In this work, for the first time, the ABC swarm optimization is presented for training a fully connected neural network using the weakly supervised technique for video anomaly detection (ABC-WSVAD). This approach leverages the advantages of bee swarm intelligence to optimize Neural Network (NN) weights, thereby accelerating the process of identifying the optimal initial weights and achieving superior results efficiently. Specifically, our methodology involves the extraction of I3D features, followed by the application of bee colony swarm optimization to train a five-layer neural network. Subsequently, the Multiple Instance Learning (MIL) ranking loss is computed to refine the model's performance. Experimental results on a very large-scale dataset UCF-Crime establish better performance than the original paper by nearly 4.36% with the preservation of the same simple model (5 layers NN) which led to a faster training process as time factor is so crucial in VAD. For future work, other types of swarm optimization algorithms will be applied. Moreover, we intend to apply our model to multiple datasets rather than UCF-Crime such as XD-Violence and ShanghaiTech. This ensures the model's effectiveness and robustness across diverse data, making it more generalizable and applicable to real-world scenarios. Furthermore, we aim to investigate techniques to address biases in dataset representation, such as data augmentation and adversarial training. In addition to training efficiency, we aim to investigate prediction speed to evaluate the model's suitability for real-time applications. This will provide a more comprehensive understanding of the model's performance in practical scenarios.

References

[1] L. Schrader *et al.*, "Advanced Sensing and Human Activity Recognition in Early Intervention and Rehabilitation of Elderly People," *J Popul Ageing*, vol. 13, no. 2, pp. 139–165, Jun. 2020, doi: 10.1007/s12062-020-09260-z.

[2] Z. Wang, Z. Yang, and T. Dong, "A Review of Wearable Technologies for Elderly Care that Can Accurately Track Indoor Position, Recognize Physical Activities and Monitor Vital Signs in Real Time," *Sensors*, vol. 17, no. 2, p. 341, Feb. 2017, doi: 10.3390/s17020341.

[3] Y. M. Galvão, L. Castro, J. Ferreira, F. B. de L. Neto, R. A. de A. Fagundes, and B. J. T. Fernandes, "Anomaly Detection in Smart Houses for Healthcare: Recent Advances, and Future Perspectives," *SN Comput Sci*, vol. 5, no. 1, p. 136, 2024.

[4] V. H. Bezerra, V. G. T. da Costa, S. Barbon Junior, R. S. Miani, and B. B. Zarpelão, "IoTDS: A One-Class Classification Approach to Detect Botnets in Internet of Things Devices," *Sensors (Basel)*, vol. 19, no. 14, p. 3188, Jul. 2019, doi: 10.3390/s19143188.

[5] M. Zalasiński, K. Łapa, and M. Laskowska, "Intelligent Approach to the Prediction of Changes in Biometric Attributes," *IEEE Transactions on Fuzzy Systems*, vol. 28, no. 6, pp. 1073–1083, 2020, doi: 10.1109/TFUZZ.2019.2955043.

[6] K. Demertzis, L. Iliadis, and I. Bougoudis, "Gryphon: a semi-supervised anomaly detection system based on one-class evolving spiking neural network," *Neural Comput Appl*, vol. 32, no. 9, pp. 4303–4314, May 2020, doi: 10.1007/s00521-019-04363-x.

[7] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks," in *2015 IEEE International Conference on Computer Vision (ICCV)*, IEEE, Dec. 2015, pp. 4489–4497. doi: 10.1109/ICCV.2015.510.

[8] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jul. 2017, pp. 4724–4733. doi: 10.1109/CVPR.2017.502.

[9] Z. Liu *et al.*, "Video swin transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3202–3211.

[10] I. Mostafa, K. H El-Safty, M. Gamal, and R. Abdel-Kader, "Abnormal Human Activity Recognition in Video Surveillance: A Survey," *Port-Said Engineering Research Journal*, 2024.

[11] M. Jiang *et al.*, "Weakly supervised anomaly detection: A survey," *arXiv preprint arXiv:2302.04549*, 2023.

[12] B. Steenwinckel, "Adaptive Anomaly Detection and Root Cause Analysis by Fusing Semantics and Machine Learning," 2018, pp. 272–282. doi: 10.1007/978-3-319-98192-5_46.

[13] W. Sultani, C. Chen, and M. Shah, "Real-World Anomaly Detection in Surveillance Videos," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2018, pp. 6479–6488. doi: 10.1109/CVPR.2018.00678.

[14] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95 - International Conference on Neural Networks*, 1995, pp. 1942–1948 vol.4. doi: 10.1109/ICNN.1995.488968.

[15] D. Karaboga, "An idea based on honey bee swarm for numerical optimization," Technical report-tr06, Erciyes university, engineering faculty, computer …, 2005.

[16] M. Dorigo, M. Birattari, and T. Stutzle, "Ant colony optimization," *IEEE Comput Intell Mag*, vol. 1, no. 4, pp. 28–39, 2006.

[17] X.-S. Yang, *Nature-inspired metaheuristic algorithms*. Luniver press, 2010.

[18] X.-S. Yang, "A new metaheuristic bat-inspired algorithm," in *Nature inspired cooperative strategies for optimization (NICSO 2010)*, Springer, 2010, pp. 65–74.

[19] G. Medioni, I. Cohen, F. Brémond, S. Hongeng, and R. Nevatia, "Event detection and analysis from video streams," *IEEE Trans Pattern Anal Mach Intell*, vol. 23, no. 8, pp. 873–889, 2001.

[20] J. Wang and A. Cherian, "Gods: Generalized one-class discriminative subspaces for anomaly detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8201–8211.

[21] A. Chriki, H. Touati, H. Snoussi, and F. Kamoun, "Deep learning and handcrafted features for one-class anomaly detection in UAV video," *Multimed Tools Appl*, vol. 80, pp. 2599–2620, 2021.

[22] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 fps in matlab," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 2720–2727.

[23] J.-X. Zhong, N. Li, W. Kong, S. Liu, T. H. Li, and G. Li, "Graph Convolutional Label Noise Cleaner: Train a Plug-And-Play Action Classifier for Anomaly Detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2019, pp. 1237–1246. doi: 10.1109/CVPR.2019.00133.

[24] J. Zhang, L. Qing, and J. Miao, "Temporal Convolutional Network with Complementary Inner Bag Loss for Weakly Supervised Anomaly Detection," in *2019 IEEE International Conference on Image Processing (ICIP)*, IEEE, Sep. 2019, pp. 4030–4034. doi: 10.1109/ICIP.2019.8803657.

[25] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal Convolutional Networks for Action Segmentation and Detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1003–1012. doi: 10.1109/CVPR.2017.113.

[26] Y. Zhu and S. Newsam, "Motion-Aware Feature for Improved Video Anomaly Detection," in *BMVC*, 2019.

[27] A. Vaswani *et al.*, "Attention is All You Need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, in NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 6000–6010.

[28] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2018, pp. 8934–8943. doi: 10.1109/CVPR.2018.00931.

[29] Y. Tian, G. Pang, Y. Chen, R. Singh, J. W. Verjans, and G. Carneiro, "Weakly-supervised Video Anomaly Detection with Robust Temporal Feature Magnitude Learning," *arXiv preprint arXiv:2101.10030*, 2021.

[30] Y. Chen, Z. Liu, B. Zhang, W. Fok, X. Qi, and Y.-C. Wu, "MGFN: Magnitude-Contrastive Glance-and-Focus Network for Weakly-Supervised Video Anomaly Detection," Nov. 2022, [Online]. Available: http://arxiv.org/abs/2211.15098

[31] Z. Yang, J. Liu, and P. Wu, "Text Prompt with Normality Guidance for Weakly Supervised Video Anomaly Detection," *arXiv preprint arXiv:2404.08531*, 2024.

[32] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, PMLR, 2021, pp. 8748–8763.
[33] H. Yuan, Z. Cai, H. Zhou, Y. Wang, and X. Chen, "TransAnomaly: Video Anomaly Detection Using Video Vision Transformer," *IEEE Access*, vol. 9, pp. 123977–123986, 2021, doi: 10.1109/ACCESS.2021.3109102.
[34] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *MICCAI*, 2015.
[35] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "ViViT: A Video Vision Transformer," *ArXiv*, Mar. 2021, [Online]. Available: http://arxiv.org/abs/2103.15691
[36] X. Feng, D. Song, Y. Chen, Z. Chen, J. Ni, and H. Chen, "Convolutional Transformer based Dual Discriminator Generative Adversarial Networks for Video Anomaly Detection," in *Proceedings of the 29th ACM International Conference on Multimedia*, New York, NY, USA: ACM, Oct. 2021, pp. 5546–5554. doi: 10.1145/3474085.3475693.
[37] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein Generative Adversarial Networks," in *Proceedings of the 34th International Conference on Machine Learning*, D. Precup and Y. W. Teh, Eds., in Proceedings of Machine Learning Research, vol. 70. PMLR, 2017, pp. 214–223. [Online]. Available: https://proceedings.mlr.press/v70/arjovsky17a.html
[38] S. Li, F. Liu, and L. Jiao, "Self-training multi-sequence learning with Transformer for weakly supervised video anomaly detection," *Proceedings of the AAAI, Virtual*, vol. 24, 2022.
[39] X. Guo *et al.*, "Discriminative-generative dual memory video anomaly detection," *arXiv preprint arXiv:2104.14430*, 2021.
[40] H. Zhou, J. Yu, and W. Yang, "Dual Memory Units with Uncertainty Regulation for Weakly Supervised Video Anomaly Detection," Feb. 2023, [Online]. Available: http://arxiv.org/abs/2302.05160
[41] Z. Wang and Y. Chen, "Anomaly detection with dual-stream memory network," *J Vis Commun Image Represent*, vol. 90, p. 103739, 2023.
[42] H. Lv and Q. Sun, "Video Anomaly Detection and Explanation via Large Language Models," *arXiv preprint arXiv:2401.05702*, 2024.
[43] L. Zanella, W. Menapace, M. Mancini, Y. Wang, and E. Ricci, "Harnessing Large Language Models for Training-free Video Anomaly Detection," *arXiv preprint arXiv:2404.01014*, 2024.
[44] V. Kaltsa, A. Briassouli, I. Kompatsiaris, and M. G. Strintzis, "Swarm-based motion features for anomaly detection in crowds," in *2014 IEEE international conference on image processing (ICIP)*, IEEE, 2014, pp. 2353–2357.
[45] D. A. Pisner and D. M. Schnyer, "Support vector machine," in *Machine learning*, Elsevier, 2020, pp. 101–121.
[46] R. Raghavendra, A. Del Bue, M. Cristani, and V. Murino, "Optimizing interaction force for global anomaly detection in crowded scenes," in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, IEEE, 2011, pp. 136–143.
[47] S. Radha, S. Aasha Nandhini, and R. Hemalatha, "Efficient Anomaly Detection System for Video Surveillance Application in WVSN with Particle Swarm Optimization," *Computational Intelligence in Wireless Sensor Networks: Recent Advances and Future Challenges*, pp. 153–177, 2017.
[48] T. Qasim and N. Bhatti, "A hybrid swarm intelligence based approach for abnormal event detection in crowded environments," *Pattern Recognit Lett*, vol. 128, pp. 220–225, 2019.
[49] N. K. PRIYADHARSINI, R. KAVITHA, A. KALIAPPAN, and D. R. D. CHITRA, "Hybrid Deep Learning Technique with One Class Svm for Anomaly Detection in Crowded Environment".
[50] H. Alsolai, F. N. Al-Wesabi, A. Motwakel, and S. Drar, "Improved Chicken Swarm Optimizer with Vision-based Anomaly Detection on Surveillance Videos for Visually Challenged People," *Journal of Disability Research*, vol. 2, no. 2, pp. 71–78, 2023.
[51] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, PMLR, 2019, pp. 6105–6114.
[52] Y. Li *et al.*, "Trajectory optimization of high-speed robotic positioning with suppressed motion jerk via improved chicken swarm algorithm," *Applied Sciences*, vol. 13, no. 7, p. 4439, 2023.
[53] S. N. Kumar and R. S. Rani, "Anomalous Human Action Monitoring in Video Images Using RPCA-MFTSL AND PSO-CNN," *SN Comput Sci*, vol. 5, no. 1, p. 109, 2023.
[54] A.-U.-H. Qureshi, H. Larijani, N. Mtetwa, A. Javed, and J. Ahmad, "RNN-ABC: A new swarm optimization based technique for anomaly detection," *Computers*, vol. 8, no. 3, p. 59, 2019.

[55] M. E. L. Bekri, O. Diouri, and D. Chiadmi, “Dynamic Inertia Weight Particle Swarm Optimization for Anomaly Detection: A Case of Precision Irrigation”.
[56] A. M. Kamoona, A. K. Gosta, A. Bab-Hadiashar, and R. Hoseinnezhad, “Multiple Instance-Based Video Anomaly Detection using Deep Temporal Encoding-Decoding,” *arXiv preprint arXiv:2007.01548*, 2020.
[57] R. Xue, J. Chen, and Y. Fang, “Real-Time Anomaly Detection and Feature Analysis Based on Time Series for Surveillance Video,” in *2020 5th International Conference on Universal Village (UV)*, 2020, pp. 1–7. doi: 10.1109/UV50937.2020.9426191.
[58] “UMN Dataset,” University of Minnesota. [Online]. Available: http://mha.cs.umn.edu/proj_events.shtml
[59] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, “Robust Real-Time Unusual Event Detection using Multiple Fixed-Location Monitors,” *IEEE Trans Pattern Anal Mach Intell*, vol. 30, no. 3, pp. 555–560, Mar. 2008, doi: 10.1109/TPAMI.2007.70825.
[60] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, “Anomaly detection in crowded scenes,” in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2010, pp. 1975–1981. doi: 10.1109/CVPR.2010.5539872.
[61] Mahadevan, Vijay, “UCSD Dataset.” [Online]. Available: http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm
[62] C. Lu, J. Shi, and J. Jia, “Abnormal Event Detection at 150 FPS in MATLAB,” in *2013 IEEE International Conference on Computer Vision*, 2013, pp. 2720–2727. doi: 10.1109/ICCV.2013.338.
[63] W. Luo, W. Liu, and S. Gao, “A Revisit of Sparse Coding Based Anomaly Detection in Stacked RNN Framework,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2017, pp. 341–349. doi: 10.1109/ICCV.2017.45.
[64] F. Landi, C. G. M. Snoek, and R. Cucchiara, “Anomaly Locality in Video Surveillance,” *ArXiv*, vol. abs/1901.1, 2019.
[65] B. Ramachandra and M. J. Jones, “Street Scene: A new dataset and evaluation protocol for video anomaly detection,” in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, Mar. 2020, pp. 2558–2567. doi: 10.1109/WACV45572.2020.9093457.
[66] P. Schneider *et al.*, “TIMo - A Dataset for Indoor Building Monitoring with a Time-of-Flight Camera,” *ArXiv*, vol. abs/2108.1, 2021.
[67] B. Degardin and H. Proença, “Iterative weak/self-supervised classification framework for abnormal events detection,” *Pattern Recognit Lett*, vol. 145, pp. 50–57, May 2021, doi: 10.1016/j.patrec.2021.01.031.
[68] A. Acsintoae *et al.*, “UBnormal: New Benchmark for Supervised Open-Set Video Anomaly Detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20143–20153.
[69] A. Paszke *et al.*, “Pytorch: An imperative style, high-performance deep learning library,” *Adv Neural Inf Process Syst*, vol. 32, 2019.
[70] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
[71] M. Z. Zaheer, A. Mahmood, M. H. Khan, M. Segu, F. Yu, and S.-I. Lee, “Generative cooperative learning for unsupervised video anomaly detection,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 14744–14754.