A Survey on the Accuracy and Performance of Video Anomaly Detection Models

Mounish Juvvadi, Lohith Kumar Reddy Kondreddy, Santosh Sai Gowtham Pasala, Lokesh Reddy Gangasani
Computer Engineering Department
San José State University (SJSU)
San José, CA, USA
Email: {mounish.juvvadi, lohithkumarreddy.kondreddy, santoshsaigowtham.pasala, lokeshreddy.gangasani}@sjsu.edu

***Abstract*—This survey evaluates the effectiveness of various anomaly detection models. Leveraging publicly available datasets of different sizes, we assess the performance of multiple machine learning algorithms in detecting anomalies in video streams. We compare a range of algorithms, each with distinct methodologies for anomaly detection. Our analysis focuses on a) the accuracy of these algorithms in identifying anomalous events and b) computational efficiency, measured in time required for both training and testing phases. Detailed analyses of results are presented, particularly concerning the reduction of data and frame size to enhance the performance of selected VAD models. Through rigorous experimentation and evaluation, this study provides insights into the strengths and limitations of different anomaly detection approaches across various video surveillance scenarios.**

***Index Terms*—Anomaly Detection; Deep learning; Spatial Autoencoder; Transformer; Unsupervised learning; Video processing**

Fig. 1. Illustration of Video Anomaly Detection [4]

I. Introduction

Anomaly detection in computers involves identifying unusual patterns or deviations in data that may signal potential threats. It is an important problem that has been researched within diverse research areas and application domains [1]. Recently, there has been a growing trend in the development and adoption of video surveillance systems, driven by heightened security needs, better affordability of hardware, and the rise of machine learning-based approaches. Video Anomaly detection holds significant potential for a variety of applications in modern society, such as surveillance systems, assisted living, healthcare, robotics, and sports analysis [2][3].

Recent advancements in deep learning have introduced a variety of innovative approaches using traditional convolutional neural networks and YOLO, aimed at enhancing the accuracy and efficiency of anomaly detection systems. These methods leverage diverse aspects of video data, including spatial features and temporal dynamics. To this end, we aim to conduct a survey of four leading VAD models, each employing unique methodologies and frameworks to address the challenges inherent in anomaly detection. We seek to provide insights into the latest advancements and possible future directions in the field of video anomaly detection.

First, we study Attention-based residual autoencoder for video anomaly detection (ASTNET) which utilizes a unified network that integrates spatial and temporal information effectively within a residual autoencoder framework, demonstrating high performance across several benchmark datasets. This model is distinguished by its use of channel attention modules and a temporal shift method, enhancing its capability to extract relevant features for anomaly detection [5].

Second, Attribute-based Representations for Accurate and Interpretable Video Anomaly Detection (AI-VAD) introduces a novel method that computes anomaly scores through a density-based analysis of object attributes such as velocity and pose. This method blends deep, implicit representations with interpretable attributes, offering a mix of high accuracy, especially on complex datasets like ShanghaiTech [6].

Third, Fast Anomaly Detection via Spatio-temporal Patch Transformation (FASTANO) which addresses the computational demands and tendency for over-generalization in prediction-based VAD systems enhances the learning process by generating irregular patch cuboids within normal frame cuboids. This method focuses on improving the model's ability to discern normal activities accurately, thereby mitigating the impact of anomalous data during training [7].

Fourth, The Diversity-Measurable Anomaly Detection (DMAD) framework emphasizes the importance of reconstruction diversity in anomaly detection. By employing a Pyramid

Deformation Module, this model assesses the severity of anomalies by analyzing multi-scale deformation fields, aiming to delicately balance detecting anomalies and reconstructing diverse normal patterns [8].

To understand the differences among the four forementioned approaches, we design and conduct a list of experiments. From the experimental results, we analyze and compare these models by meticulously examining their methodologies, assessing their accuracy using key metrics such as AUC, PSNR, and CPU time, and measuring their performance on training and inference across different dataset sizes and frame sizes.

Our contributions are as follows:

- We conduct a survey of four models and provide detailed information about each.
- We perform experiments and compare the models under diverse conditions.

We found that in our comparative analysis, AI-VAD consistently outperformed the other models. This model exhibited superior AUC scores even with reduced data, proving its effectiveness in complex environments. It achieved the highest AUC scores on average, particularly excelling in the Avenue and Shanghai datasets. DMAD, while performing best in simpler datasets like Ped2. FASTANO demonstrated the shortest CPU time, making it suitable for real-time applications. DMAD, despite its high PSNR values indicating better image reconstruction, had longer CPU time.

The rest of the paper is organized as follows. Chapter II begins with a survey of related works in the field, setting the stage by contextualizing our research within existing literature. Chapter III discusses the methodology, detailing the criteria and approaches used to compare various VAD models. Chapter IV presents the evaluation results and analysis of the compared models. Chapter V explores related works, and Chapter VI provides the conclusion of our survey.

II. Background on Anomaly detection

Anomaly detection, also known as outlier detection, is a critical component of intelligent surveillance systems, focusing on identifying abnormal objects or unusual human behaviors in video sequences[2]. The goal is to distinguish rare and unusual occurrences, and anomalies from the regular patterns in the data. Anomalies can take various forms, such as unexpected events, errors, outliers, or fraudulent activities, and detecting them is crucial in numerous domains for maintaining system reliability, security, and efficiency. However, many existing anomaly detection techniques fail to retain sufficient accuracy due to so-called "big data" characterized by high-volume, and high-velocity data generated by a variety of sources [9].

Video anomaly detection stands at the intersection of several advanced technologies, involving video processing, machine learning, and pattern recognition. It is a challenging task because most anomalies are scarce and non-deterministic [11]. Its effectiveness depends on the ability to train models that can generalize well from training data to real-world scenarios. As technology advances, the integration of more sophisticated AI models and increased computational power is likely to enhance the accuracy and efficiency of video anomaly detection systems, making them more prevalent in ensuring public

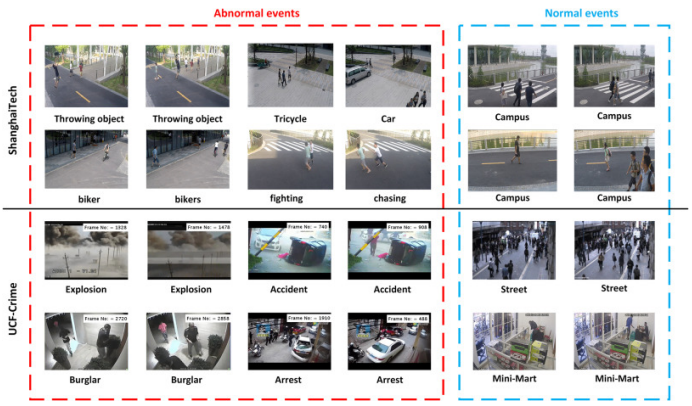

Fig. 2. Anomaly in different scenarios [10]

safety and operational integrity. Video anomaly detection has advanced through the integration of unsupervised learning techniques like autoencoders, which learn normal patterns without labeled anomalies. Many approaches investigate the reconstruction difference between normal and abnormal patterns but neglect that anomalies do not necessarily correspond to large reconstruction errors [11].

Furthermore, hybrid models that combine CNNs for spatial analysis and LSTMs for temporal insights create robust systems capable of understanding complex behaviors over time, enhancing the detection accuracy in dynamic scenarios [12]. Particularly in crowded environments to understand how individuals behave when they are part of a large group such as marketplaces or public events [13], advanced models utilize techniques like crowd density estimation to discern abnormal behavior effectively, even amidst dense populations. Additionally, the integration of real-time alerting systems is crucial, as it enables immediate processing and response to detected anomalies, ensuring timely interventions in critical situations such as public safety and emergency responses. These innovations collectively improve the capability, accuracy, and responsiveness of anomaly detection systems, making them indispensable in modern surveillance and monitoring applications.

A. Background on Deep Learning Based Anomaly Detection

1) Neural Networks: Neural networks are computational models with interconnected nodes organized into layers, that utilize weighted connections. In a neural network, the connection between one neuron to another exists with some strength known as weight or synaptic weight. The on and off state of a neuron is decided by the threshold function [14]. They use activation functions to learn mappings from input to output data, making them versatile for tasks such as pattern recognition, decision-making, image classification, natural language processing, and reinforcement learning. In the context of machine learning, deep learning employs neural networks with multiple layers to autonomously extract hierarchical representations from data. This intrinsic capacity enables deep learning models to discern detailed features, leading to success in applications like image and speech recognition.

2) Auto-Encoder: In the domain of unsupervised learning, autoencoders offer a specialized neural network architecture. They consist of an encoder for data compression and a decoder for reconstruction. Proficient in tasks such as data compression, feature learning, and anomaly detection, autoencoders quantify the instances of reconstruction, aiding in the identification of irregularities within the data [15].

3) Attention Mechanism: To enhance the capability of models to focus on specific parts of input data, attention mechanisms are incorporated. Whether in image captioning, machine translation, or sequential data processing, attention mechanisms enable the model to assign varying degrees of importance to different elements, improving overall performance and interpretability. Even in the case of limited computing power, it can process more important information with fewer computing resources [16].

B. Diversity-Measurable Anomaly Detection (DMAD)

DMAD includes an Information Compression Module and a Pyramid Deformation Module to detect anomalies in various scenarios. The framework focuses on enhancing reconstruction diversity to improve anomaly detection performance on video surveillance and industrial defect detection tasks. Experimental results demonstrate the effectiveness of the DMAD framework in detecting anomalies, even in the presence of contaminated data and anomaly-like normal samples.

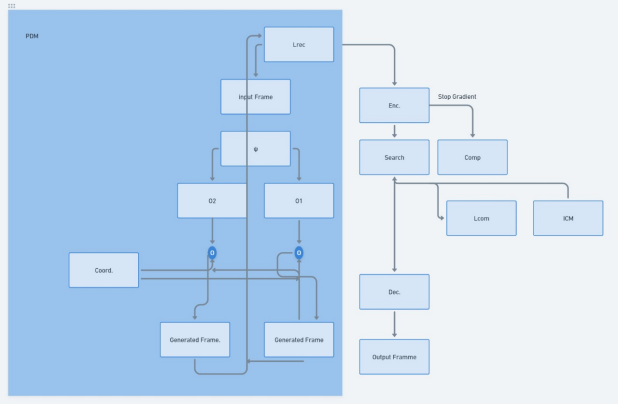

Fig. 3. Architecture of DMAD [8]

DMAD framework consists of two main components: the Information Compression Module (ICM) and the Pyramid Deformation Module (PDM). The Information Compression Module is responsible for compressing information and creating prototypical embeddings of normal patterns. It helps in capturing the typical characteristics of normal data points. On the other hand, the Pyramid Deformation Module is designed to model diverse normal patterns and measure the severity of anomalies. By decoupling deformation from prototypical embedding, the PDM ensures that the anomaly score is more reliable and accurate. This module enhances anomaly discrimination and reconstruction diversity, which is crucial for effective anomaly detection. The integration of the Information Compression Module and the Pyramid Deformation Module in the DMAD framework aims to improve anomaly detection performance by enhancing reconstruction diversity while avoiding undesired generalization on anomalies. The framework has been shown to be effective in detecting anomalies in various scenarios, including surveillance videos and industrial images. It also demonstrates robustness to contaminated data and anomaly-like normal samples [8].

C. Attribute-based Representations for Accurate and Interpretable Video Anomaly Detection (AI-VAD)

AI-VAD introduces a video anomaly detection that combines explicit attribute-based representations with implicit deep representations to achieve state-of-the-art performance on commonly used datasets. The methodology involves three stages: pre-processing, feature extraction, and density estimation. In the pre-processing stage, optical flow maps and bounding boxes are extracted to represent objects and motions. Velocity, pose, and deep features are then extracted from these representations to create object-level feature descriptors. The deep features, pre-trained on external datasets, capture residual attributes not described by velocity and pose features. Density estimation is used to score samples as normal or anomalous, with low estimated density indicating anomalies. The model's interpretability allows for making decisions based on semantic attribute-based representations, making it applicable in critical environments.

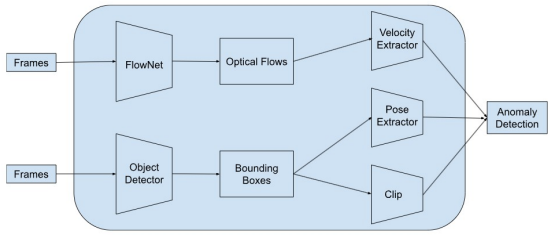

Fig. 4. Architecture of AI-VAD [6]

The model's effectiveness is demonstrated by its high correlation with ground-truth anomalous events, showcasing its ability to accurately detect anomalies in videos. The use of semantic attributes for representation allows for automatic interpretation of anomalies based on unusual values in velocity and pose features. Object-based methods, enabled by accurate object detection, outperform frame-level methods and demonstrate the feasibility of crafting accurate and interpretable semantic features for anomaly detection [6].

D. Fast Anomaly Detection via Spatio-temporal Patch Transformation (FASTANO

FASTANO introduces a fast anomaly detection method for surveillance videos using spatial rotation transformation (SRT) and temporal mixing transformation (TMT) to enhance the learning of normal features. The model is trained on normal frames with artificially generated abnormal patches and achieves competitive accuracy while surpassing previous works in terms of speed. The proposed method does not rely on pre-trained networks and is suitable for real-world

anomaly detection. The method involves a patch anomaly generation phase followed by an autoencoder architecture, applying spatial random transformation (SRT) and temporal motion transformation (TMT) to focus on moving objects in video frames. The model is trained to minimize prediction loss and generate frames resembling the ground truth, achieving competitive results on three datasets.

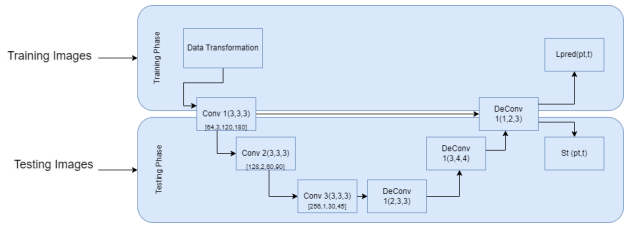

Fig. 5. Architecture of FASTANO [7]

The SRT involves rotating patch cuboids to capture foregrounds and encourage the model to focus on moving objects, while TMT shuffles patch sequences in the temporal axis to generate abnormal movement. This approach does not rely on pre-trained networks and is designed to be computationally efficient and suitable for real-world anomaly detection. The model demonstrates frame-level detecting performance by rapidly decreasing the anomaly score when anomalies appear and increasing it when abnormal objects disappear. Additionally, the model is capable of localizing anomalies and achieves high detection speed, making it suitable for real-time applications [7].

E. Attention-based residual autoencoder for video anomaly detection (ASTNET)

ASTNET utilizes a residual autoencoder architecture with a deep convolutional neural network-based encoder and a multi-stage channel attention-based decoder. By effectively combining spatial and temporal features, the network outperforms existing methods on three standard benchmark datasets. The proposed approach includes the use of channel attention modules to extract contextual dependency and the temporal shift method to model temporal information, showcasing promise for real-world surveillance applications.

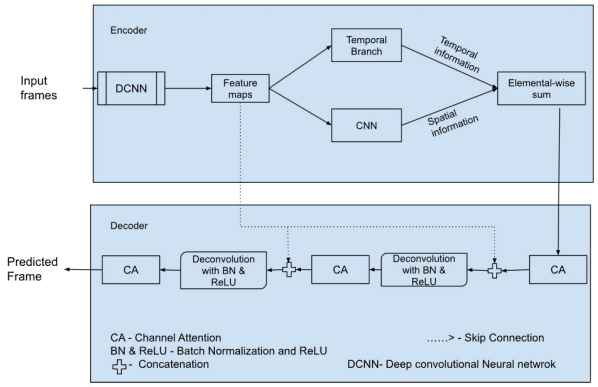

Fig. 6. Architecture of ASTNET [5]

The architecture features an encoder-decoder structure, with the encoder capturing appearance and motion details through a deep and wide convolutional neural network, such as WiderResnet. This network extracts high-level features which are then channeled into temporal and spatial branches. The temporal branch employs a novel shifting technique to model temporal features across frames, while the spatial branch aggregates feature from multiple frames using a 1×1 convolution and integrates channel attention to enhance performance.

During training, the network learns to predict future frames, evaluating its performance by computing an anomaly score (S(t)) for each frame. This score is derived from comparing the Peak Signal-to-Noise Ratio (PSNR) of predicted and ground-truth frames, with higher anomaly scores signaling potential deviations from normal behavior. To ensure consistency, PSNR scores are normalized to the [0,1] range, facilitating standardized anomaly detection across varied input scenarios [5].

III. Methodology

Our research primarily centers on conducting a comparative analysis of four models dedicated to video anomaly detection. This analysis is conducted across various phases, encompassing the execution of these models with different frame sizes, dataset sizes, and diverse datasets. Additionally, we assess CPU times, followed by a comparison of AUC scores and PSNR values. By varying the size of the training data from 100% down to 12%, the experiment aims to assess the models' adaptability under different data availability scenarios. The reduction in training data size simulates scenarios where limited labeled data may be available for model training, mimicking real-world constraints such as resource limitations or data scarcity. Experimenting with different frame sizes, such as 128x128 and 256x256, allows for the examination of how model performance varies with image resolution. Higher frame resolutions provide more detailed information but may also increase computational complexity, thus affecting CPU times. Calculating training times provides insights into the computational efficiency of each model, indicating how quickly they can be trained on a given dataset. Models with shorter training times are preferable, especially in time-sensitive applications or scenarios with limited computational resources. Testing times reflect the efficiency of models in processing new data and making predictions, which is crucial for real-time or near-real-time applications. The experiment anticipates identifying models that strike a balance between accuracy and computational efficiency, aiming for optimal performance in practical deployment scenarios.

The initial section of our study provides an in-depth description of the experimental setup. Here, we outline the configuration of the models, hardware, and libraries utilized. Following this, the subsequent section delves into the datasets employed, providing detailed insights into their characteristics and relevance to our research. Finally, in the concluding section, we describe the evaluation metrics utilized to compare the models' performances.

A. Experimental Setup

The configuration for this computational setup comprises Python and its necessary libraries, along with the PyTorch framework, leveraging GPU hardware for optimized computational performance. Prior to implementation, a thorough installation of all required Python libraries is conducted. Subsequently, hardware compatibility with software is verified, particularly ensuring alignment between the CUDA version and PyTorch version for seamless operation. The experiment is conducted on system with Ubuntu - 20.04, AMD Ryzen 7 7745HX with Radeon Graphics, 3601 Mhz, 8 Core(s), 16 Logical Processor(s), CPU - 16GB RAM 5.15GHz clock cycle, Graphics - NVIDIA® GeForce RTX™ 4060 8GB GDDR6 8.00GHz, Memory number of channels - 2*32 bit, Peak Bandwidth - 2800 MHz, Bus Type Pci-express 5.0 (32.0) gt/s 99.79 MHz. Datasets, including PED 2, Avenue, and Shanghai Tech, are downloaded for experimentation. Data preprocessing is imperative, showcased by the conversion of videos in the Shanghai Tech dataset into frames to align with model requirements. This step involves meticulous formatting to prepare data for subsequent processing. Directory paths are then established, encompassing datasets, training, and testing directories for streamlined access. Training commences, accompanied by timer implementation to monitor performance metrics. Following training, the model undergoes testing, similarly timed to evaluate its efficiency. This systematic approach facilitates the acquisition of training and testing durations, crucial for assessing model effectiveness.

B. Data Sets

1) Ped2: The Ped2 anomaly detection dataset was captured using a stationary camera positioned at an elevated vantage point, providing a bird's-eye view of pedestrian walkways. It consists of 16 training video samples and 12 testing video samples, providing a comprehensive dataset for the study and development of anomaly detection algorithms in surveillance scenarios.

Throughout the footage, the density of crowds on these walkways is varied, ranging from sparse to densely packed. Anomalies in the dataset arise from two main sources: the presence of non-pedestrian entities traversing the walkways and irregular pedestrian motion patterns. Examples of commonly occurring anomalies include cyclists, skateboarders, individuals pushing small carts, and pedestrians deviating from designated paths to walk on the surrounding grass. Additionally, a few instances of individuals using wheelchairs were also captured. To facilitate experimentation and evaluation, the videos were segmented into clips, with each clip comprising approximately 180 frames.

2) CHUCK Avenue: The Avenue Dataset comprises 16 training and 21 testing video clips, totaling 30652 frames captured within the CUHK campus avenue. Of these frames, 15328 belong to the training set, while the remaining 15324 frames constitute the testing set.

The dataset encompasses 15 sequences, each lasting approximately 2 minutes. In the training videos, normal situations are predominantly observed, providing a baseline for model training. Conversely, the testing videos feature a mix of normal and abnormal events, serving as a comprehensive evaluation set for anomaly detection algorithms.

3) ShanghaiTech: The ShanghaiTech Campus dataset comprises 13 scenes characterized by diverse lighting conditions and camera angles. With over 270,000 training frames, it offers a substantial dataset for anomaly detection research.

Featuring a multi-scene campus environment, this dataset encapsulates various scenarios, including both human-related anomalies such as jumping and cycling, and non-human-related anomalies like the presence of large vehicles. Notably, the dataset's distinctiveness lies in its diverse scenes, camera angles, and lighting conditions, providing a comprehensive testbed for anomaly detection algorithms.

C. Evaluation Metrics

In this section, we delve into the various metrics and scenarios used to compare the Video Anomaly Detection (VAD) models. Key considerations include the size of the datasets used, the frame size of the video segments, and the measurement of CPU times. The frame size in video segments is particularly important as it affects the temporal resolution at which the model identifies anomalies, impacting both the precision and reliability of detection. Larger frame sizes might integrate more contextual information but could also lead to delays in the detection of sudden anomalies. Additionally, the size of the datasets is crucial, as larger datasets typically provide a more robust basis for training, potentially enhancing the model's ability to generalize across different scenarios. By evaluating processing times, we assess both efficiency and the demand for computational resources. These factors are critically examined as they directly influence the VAD models' effectiveness and their applicability in varied real-world settings.

Following this, we examine three key metrics: AUC (Area Under the Curve), PSNR (Peak Signal-to-Noise Ratio) and CPU (Central Processing Unit) time. AUC provides a comprehensive measure of a model's ability to distinguish between classes, specifically the presence or absence of anomalous activity. We assess the model's accuracy in detecting anomalies by comparing the labels assigned to generated frames with ground truth annotations. Each frame is associated with a binary label, indicating the presence (1) or absence (0) of an anomaly. By evaluating the models based on the Area Under the Curve (AUC) score, we quantify their ability to distinguish between normal and anomalous instances. A higher AUC score signifies superior performance, indicating the model's effectiveness in accurately identifying anomalies within the video data. PSNR is a metric commonly used to measure the quality of reconstructed images or videos by comparing them to the original, uncompressed version. PSNR calculates the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the accuracy of its representation. We compute the PSNR value between each generated frame and its corresponding ground truth frame. A lower PSNR value indicates reduced accuracy in predicting the frame, suggesting the presence of an anomaly.

Particularly, when models trained solely on normal behavior encounter anomalous frames, their predictive accuracy diminishes, resulting in lower PSNR scores. To gauge overall performance, we aggregate the PSNR values for all frames and calculate their average, representing the final PSNR score for each model. In AI-VAD, we can't calculate PSNR since no frames are generated. Instead, we evaluate the velocity and pose representations by comparing them to the ground truth labels. CPU time measures the duration a CPU spends actively processing a program's instructions, crucial for assessing a video anomaly detection system's efficiency. It indicates the system's capability to analyze and detect anomalies in real time, balancing accuracy and speed. Minimizing CPU time is vital for optimizing these systems, especially in environments requiring immediate response to unusual activities.

IV. Evaluation Results

In our comparative analysis, we thoroughly evaluated the performance and efficiency of four VAD models—DMAD, FASTANO, ASTNET, and AI-VAD across various frame sizes, data sizes, and multiple datasets including Ped2, Avenue, and ShanghaiTech. This evaluation consists of metrics such as AUC, PSNR, and CPU times to know the model's effectiveness and applicability in real-world surveillance situations.

A. AUC Performance Results and Analysis

1) AUC Values of Various Models by Dataset and Frame Size for 100% Dataset: We observe that in Figure 7, the DMAD model has the highest AUC score of 99.14 in the Ped2 dataset for both frame sizes (128x128 and 256x256), while ASTNET has the lowest with an AUC of 93.20. However, when the datasets are changed to Avenue and Shanghai, the performance of these models is affected. In the Avenue dataset, the AI-VAD model surpasses DMAD, achieving the best performance with AUC scores of 93.32 at the 256x256 frame size and 83.36 at the 128x128 frame size. In the Shanghai dataset, AI-VAD again performs best, while ASTNET has the least AUC in both frame sizes. On average, AI-VAD has the highest AUC and ASTNET has the lowest.

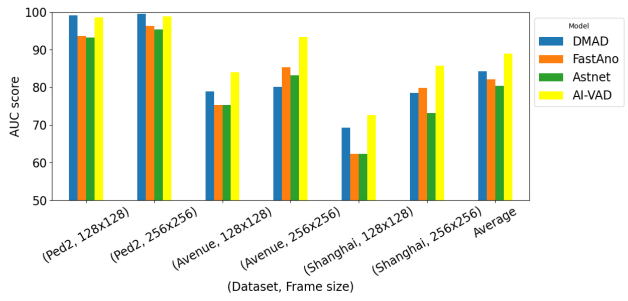

Fig. 7. AUC Values of Various Models by Dataset and Frame Size for 100% Dataset

2) AUC Values of Various Models by Dataset and Frame Size for 50% Dataset: We observe that in Figure 8, the AUC score in the 50% dataset size is slightly declined compared to the 100% dataset, but the relative ranking of the models mostly remains consistent. In the Ped2 dataset, the DMAD model still has the highest AUC of 99.38 at the 256x256 frame size. ASTNET, while improving slightly over its performance at the 100% dataset size, still has the least AUC in 128x128 frame size. Moving to the Avenue dataset, AI-VAD has performed the best with an AUC of 93.11 at 256x256, a slight decrease but still maintaining a substantial lead over the other models. DMAD shows more significant drops in performance, which underscores its sensitivity to reduced data volumes in more complex datasets. In Shanghai, AI-VAD again tops the charts demonstrating its robustness across various datasets and ASTNET remains with the lowest AUC score in this dataset. On average, AI-VAD has the highest AUC and ASTNET has the lowest.

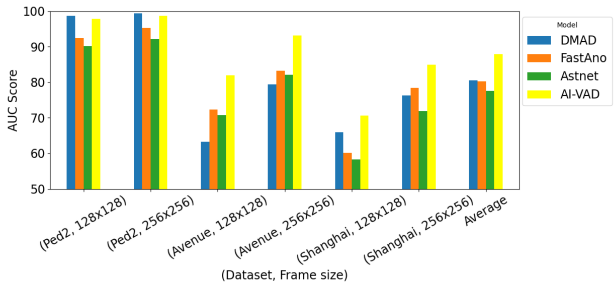

Fig. 8. AUC Values of Various Models by Dataset and Frame Size for 50% Dataset

3) AUC Values of Various Models by Dataset and Frame Size for 25% Dataset: We observe that in Figure 9, the patterns of model performance begin to show more pronounced differences due to the substantial reduction in data. Despite this, the DMAD model continues to perform well in the Ped2 dataset with an AUC of 98.52 at 256x256 frame size and the ASTNET model performance remains the lowest among the models, scoring 92.5 at 256x256 frame size. In the Avenue dataset, AI-VAD's performance continues to outperform other models with an AUC of 91.23 at 256x256, indicating its superior ability to handle complex datasets with limited data and the DMAD has the lowest AUC in this dataset. On average, AI-VAD has the highest AUC and ASTNET has the lowest.

4) AUC Values of Various Models by Dataset and Frame Size for 12% Dataset: figure 10 shows that at the 12% dataset size, all models face their most challenging test due to the severe limitation in data volume. In Ped2, DMAD achieves an AUC of 96.34 at 256x256, which, while lower than its performance at higher data volumes, shows its capability in simpler dataset conditions. ASTNET's performance, scoring 90.4 at 256x256, remains the lowest. In the Avenue and Shanghai dataset, AI-VAD's resilience shines through with an AUC of 89.12 at 256x256, still leading by a significant margin despite the overall decline in scores across all models. DMAD's struggle becomes more apparent here, indicating its

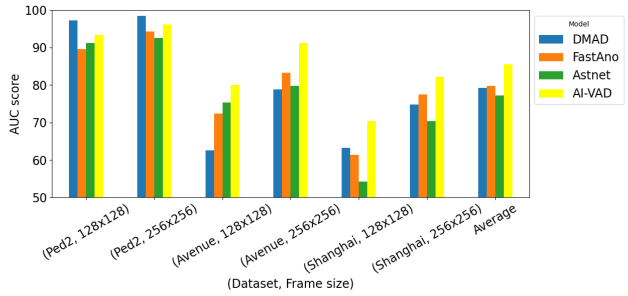

Fig. 9. AUC Values of Various Models by Dataset and Frame Size for 25% Dataset

lesser adaptability to complex and limited data. Overall the AI-VAD model exhibits the highest average AUC and the ASTNET model has the lowest.

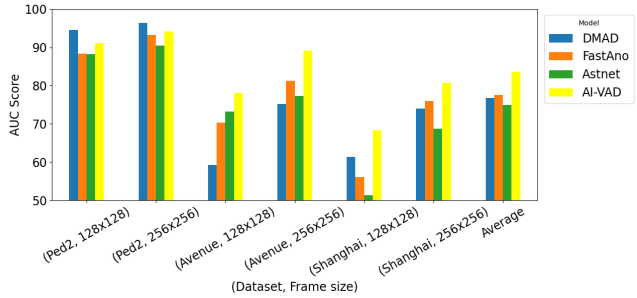

Fig. 10. AUC Values of Various Models by Dataset and Frame Size for 12% Dataset

The analysis of the four anomaly detection models—DMAD, FASTANO, ASTNET, and AI-VAD—across varying dataset sizes (100% to 12%) and frame sizes (128x128 and 256x256) reveals distinct performance dynamics under different data constraints and resolution settings. AI-VAD consistently demonstrates robustness and superior performance, particularly excelling in complex datasets like Avenue and Shanghai, even under severe data reductions. This model's effectiveness is notably more pronounced at the higher frame size of 256x256, emphasizing the advantage of higher-resolution data in enhancing anomaly detection accuracy. In contrast, DMAD shows excellent results in simpler scenarios such as the Ped2 dataset at full data capacity, especially at 256x256, but its performance declines more noticeably than AI-VAD in more complex environments and smaller datasets, illustrating potential limitations in adaptability. ASTNET has the lowest performance struggling across both frame sizes in all dataset reductions, showing its challenges in extracting useful patterns from sparse and complex data. This comprehensive analysis not only underscores the significance of choosing models based on their resolution handling capabilities and robustness to data sparsity but also highlights the importance of optimizing anomaly detection systems to leverage higher resolutions, thereby maximizing their efficacy in real-world scenarios.

B. PSNR metric results

The Peak Signal-to-Noise Ratio (PSNR) values for different anomaly detection models—DMAD, FASTANO, and ASTNET across various dataset sizes (100%, 50%, 25%, 12%) and datasets (Ped2, Avenue, Shanghai) provide a quantitative measure of the image quality produced by these models. Below is a detailed analysis of how each model performs in terms of PSNR across different data availability and frame sizes. Note that In AI-VAD, we can't calculate PSNR since no frames are generated.

1) PSNR Values of Various Models by Dataset and Frame Size for 100% Dataset: Figure 11 displays that at full dataset capacity, DMAD consistently achieves the highest PSNR values across all datasets, notably reaching 51.54 at 256x256 and 50.32 at 128x128 in the Ped2 dataset, indicating superior image reconstruction quality at both frame sizes. This trend of high performance continues in the Avenue and Shanghai datasets with PSNR values of 46.40 and 45.53 respectively at 256x256, and slightly lower scores at 128x128. FASTANO, while trailing DMAD, shows competitive PSNR values, notably achieving 47.64 at 256x256 and 42.72 at 128x128 in Ped2, indicating robustness in detail preservation. ASTNET records lower PSNR values, with its best at 44.32 at 256x256 and 38.12 at 128x128 in Ped2, suggesting some limitations in maintaining image quality compared to the other models. Overall DMAD has the highest PSNR score and ASTNET has the lowest.

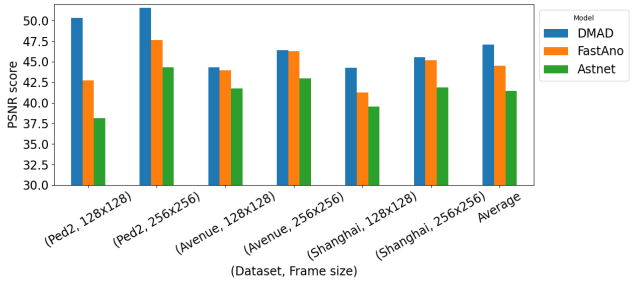

Fig. 11. PSNR Values of Various Models by Dataset and Frame Size for 100% Dataset

2) PSNR Values of Various Models by Dataset and Frame Size for 50% Dataset: In Figure 12, With the dataset size reduced to half, DMAD continues to lead in PSNR values, maintaining a high of 51.43 at 256x256 and 50.25 at 128x128 in Ped2. This minor drop highlights DMAD's resilience in producing high-quality image outputs even with less data. FASTANO and ASTNET show reductions in PSNR; however, FASTANO manages to reach 45.51 at 256x256 and 41.24 at 128x128 in Ped2, underscoring its capability to adapt reasonably well to reduced data. ASTNET sees a decrease across all datasets but remains particularly challenged in Shanghai, where its highest PSNR is 41.11 at 256x256 and 38.11 at 128x128. Overall DMAD has the highest PSNR score and ASTNET has the lowest.

3) PSNR Values of Various Models by Dataset and Frame Size for 25% Dataset: The bar graph in Figure 13 indicates

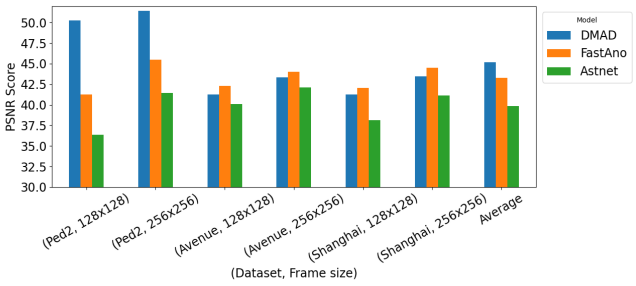

Fig. 12. PSNR Values of Various Models by Dataset and Frame Size for 50% Dataset

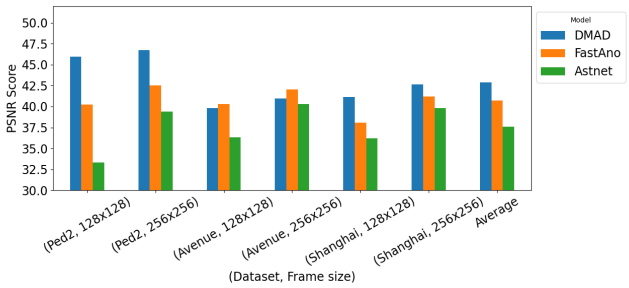

Fig. 14. PSNR Values of Various Models by Dataset and Frame Size for 12% Dataset

that all models exhibit further reductions in PSNR values at a quarter dataset size. DMAD's performance decreases but still tops with a PSNR of 49.43 at 256x256 and 47.87 at 128x128 in Ped2, demonstrating its effectiveness in handling lower data volumes. FASTANO maintains a relatively stable output with a best of 45.64 at 256x256 and 41.72 at 128x128 in Ped2, while ASTNET struggles more significantly, peaking at 41.22 at 256x256 and 38.86 at 128x128 in Avenue. The trend indicates that all models are affected by reduced data but DMAD remains comparatively robust by exhibiting the highest PSNR score and ASTNET has the lowest PSNR.

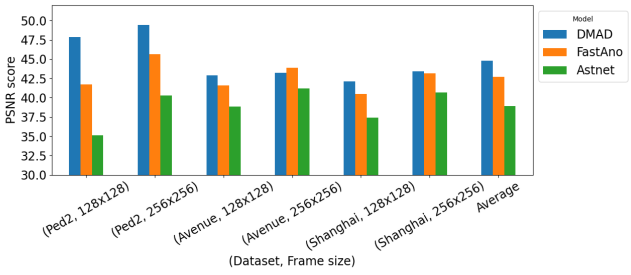

Fig. 13. PSNR Values of Various Models by Dataset and Frame Size for 25% Dataset

4) PSNR Values of Various Models by Dataset and Frame Size for 12% Dataset: Figure 14 shows the bar graph of PSNR values of the 12% dataset. Operating with only 12% of the data presents the most severe challenge for all models. DMAD's highest PSNR at this data level is 46.76 at 256x256 and 45.96 at 128x128 in Ped2, illustrating a notable decline but still outperforming other models. FASTANO and ASTNET experience further drops, with FASTANO reaching 42.51 at 256x256 and 40.24 at 128x128 in Ped2, and ASTNET peaking at 40.31 at 256x256 and 36.34 at 128x128 in Avenue. This level marks the lowest performance for ASTNET, particularly in maintaining image quality, which is critical in anomaly detection. Overall DMAD has the highest PSNR score and ASTNET has the lowest.

Throughout the dataset sizes from 100% to 12%, DMAD consistently displays superior image quality, achieving the highest PSNR values, which highlights its efficiency in image reconstruction under varied data conditions. This model's ability to maintain high PSNR values even at reduced dataset sizes demonstrates its strong underlying algorithms and their effectiveness in handling data sparsity. FASTANO, though slightly behind DMAD, shows resilience, particularly at higher resolutions, suggesting that while it may not always match DMAD in image fidelity, it remains a viable option for applications where slightly lower image quality is acceptable. ASTNET, however, struggles as dataset sizes decrease, showing the largest drops in PSNR values. This trend might indicate limitations in ASTNET's ability to reconstruct high-quality images from sparse data, suggesting a potential area for further model optimization or the need for more robust training methods. This pattern across dataset sizes underscores the importance of choosing the right model based on the specific requirements of an application, considering both the availability of data and the necessity for high image quality. In settings where high fidelity in image reconstruction is critical, DMAD proves to be the most reliable, whereas in scenarios where data is more limited, the selection between FASTANO and ASTNET would depend on the acceptable trade-off between performance and image quality. This also points to the potential benefits of enhancing model training and algorithm refinement to boost the performance of FASTANO and especially ASTNET under conditions of data scarcity.

C. Training time results

1) Training Time of Various Models by Dataset and Frame Size for 100% Dataset: In Figure 15, for the 100% dataset size frame size plays a crucial role in training time across all models. In the Ped2 dataset, the training times are notably shorter at smaller frame sizes, with AI-VAD being particularly efficient (0.15 hours at 128x128) and DMAD being least efficient (0.55 hours at 128x128). As frame size changes to 256x256, DMAD's training time nearly doubles to 1.1 hours, highlighting its sensitivity to increased data complexity. In contrast, FASTANO's training time remains very low in the Avenue and Shanghai datasets. Overall, on average, the FASTANO model has the lowest training time and the DMAD model has the highest training time.

2) Training Time of Various Models by Dataset and Frame Size for 50% Dataset: Figure 16 shows the graph for the training time of the 50% dataset. Reducing the dataset to 50% generally lessens the training times across all models

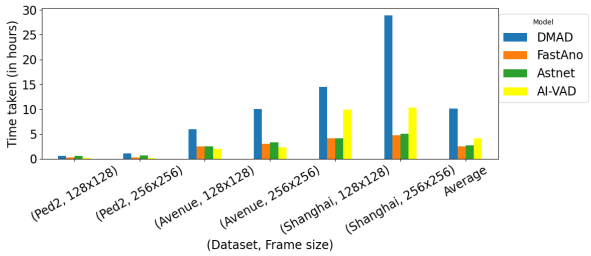

Fig. 15. Training Time of Various Models by Dataset and Frame Size for 100% Dataset

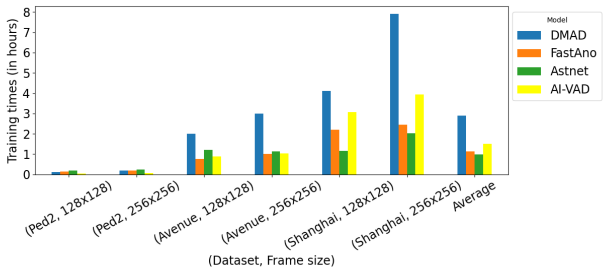

Fig. 17. Training Time of Various Models by Dataset and Frame Size for 25% Dataset

and frame sizes. In the Shanghai dataset at 256x256, DMAD's training time is reduced to 10 hours, and AI-VAD's to 6.05 hours. The 128x128 frame size also shows reduced times, with AI-VAD only requiring 5.80 hours. This reduction indicates a significant relief in computational load. On the other hand, when the average was calculated ASTNET and FASTANO displayed almost the same amount of lowest training time and DMAD had the highest.

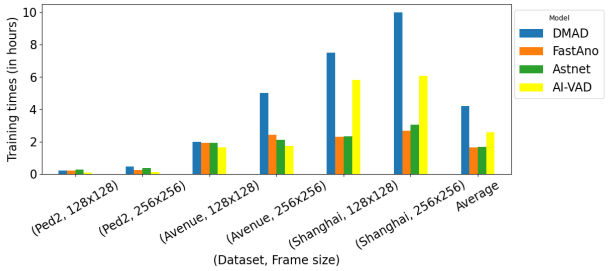

Fig. 16. Training Time of Various Models by Dataset and Frame Size for 50% Dataset

3) Training Time of Various Models by Dataset and Frame Size for 25% Dataset: Figure 17 displays the training time for the 25% dataset. At 25% dataset size, training times continue to decrease. For instance, in the Shanghai dataset at 256x256, DMAD requires 7.9 hours, and AI-VAD only 3.95 hours. At the smaller frame size of 128x128, FASTANO, ASTNET, and AI-VAD models had the same amount of training time in Ped2 and avenue datasets. On average, ASTNET had the lowest training time and DMAD had the highest.

4) Training Time of Various Models by Dataset and Frame Size for 12% Dataset: Figure 18 demonstrates the training time for the 12% dataset. With the dataset reduced to just 12%, all models achieve their lowest training times. In the Shanghai dataset at 256x256, DMAD's training time decreases to 4.3 hours and AI-VAD's to 2.20 hours. but, at 128x128, the FASTANO model took only 1.45 hours of training time, demonstrating its potential for rapid deployment in scenarios with stringent time constraints. On average, the FASTANO model had the lowest training time and DMAD had the highest. This dataset size and frame size reduction collectively offer a clear view of each model's scalability and efficiency under reduced computational demands.

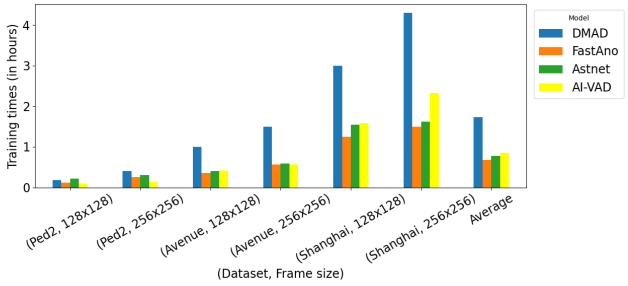

Fig. 18. Training Time of Various Models by Dataset and Frame Size for 12% Dataset

The analysis underscores the importance of considering both dataset size and frame size when evaluating model efficiency. ASTNET and FASTANO models consistently demonstrate superior training time efficiency across all conditions, making them ideal choices for environments where quick model readiness is crucial. DMAD and AI-VAD, while capable in terms of performance but required considerably more time which might limit its practicality in resource-constrained scenarios. On average, FASTANO has the lowest amount of training time when compared to others. Overall, the selection of anomaly detection models should be a balanced decision based on both performance metrics and practical considerations like training time and computational resource requirements. The detailed insights provided here facilitate making informed decisions tailored to specific operational constraints and performance expectations.

D. Inference of Various Models by Dataset

Figure 19 illustrates the time taken per video by four anomaly detection models—DMAD, FASTANO, ASTNET, and AI-VAD—across three datasets: Ped2, Avenue, and Shanghai, with varying frame sizes. In the Ped2 dataset with a frame size of 128x128, FASTANO was the fastest, taking approximately 2 seconds per video, followed by DMAD and ASTNET both taking around 3 seconds, and DMAD at about 5 seconds. With a frame size of 256x256, FASTANO remained

the quickest at roughly 5 seconds, while DMAD and ASTNET were close at around 6 seconds, and ASTNET took the longest at nearly 9 seconds.

For the Avenue dataset with a 128x128 frame size, FASTANO again demonstrated the quickest processing time at about 7 seconds, followed by ASTNET at around 8 seconds, AI-VAD at approximately 9 seconds, and DMAD at 13 seconds. When the frame size was increased to 256x256, AI-VAD led at 12 seconds, with FASTANO, ASTNET, and DMAD taking around 13, 14, and 17 seconds respectively. In the Shanghai dataset with a 128x128 frame size, ASTNET processed the videos in roughly 3 seconds, FASTANO and AI-VAD took about 4 and 5 seconds respectively, and DMAD took 7 seconds. For the 256x256 frame size, ASTNET remained the fastest at about 5 seconds, FASTANO and AI-VAD were close at 8 and 11 seconds respectively, and DMAD took around 13 seconds. On average, FASTANO consistently demonstrated the lowest inference across all datasets and frame sizes, followed by AI-VAD and ASTNET, with DMAD generally having the highest inference. Despite DMAD occasionally delivering higher AUC scores, its longer inference, especially noticeable with larger frame sizes in the Shanghai dataset, could be a significant drawback in time-sensitive scenarios.

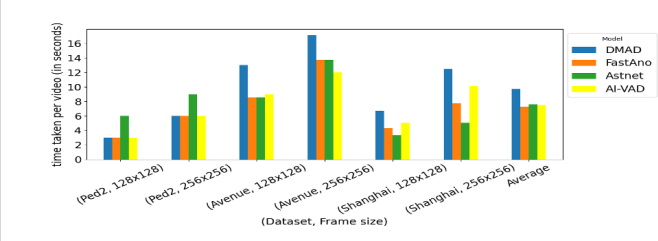

Fig. 19. Inference of various models by dataset and frame size per one input

V. Related Works

Video anomaly detection, a fundamental component of surveillance and security systems, entails the identification of abnormal events or behaviors within video streams. Over the years, researchers have explored many methodologies to address the complexities inherent in this task. Herein, we delve into a comprehensive review of the existing literature, encompassing traditional and contemporary approaches, challenges encountered, and recent advancements in video anomaly detection.

A. Statistical Techniques

In a paper [17], the author discussed statistical anomaly detection techniques, highlighting the principle that an anomaly is an observation suspected of being partially or wholly irrelevant because it does not conform to the assumed stochastic model. The fundamental assumption behind statistical anomaly detection techniques is that normal data instances reside in high-probability areas of a stochastic model, whereas anomalies occupy low-probability regions. Both parametric and non-parametric approaches are employed to develop these models. Depending on the assumed distribution type, parametric techniques can be further divided into:

1) Gaussian Model-Based: These techniques assume Gaussian-distributed parameters, with the parameters estimated using Maximum Likelihood Estimation (MLE). Anomaly scores are then derived based on deviations from this distribution.

2) Regression Model Based: Anomaly detection via regression has been applied in various contexts, such as linear regression models, high-dimensional outlier detection, categorical or mixed data, and time series data. The basic regression model-based anomaly detection involves two steps: first, a regression model is learned from the data. In the second step, the residuals, or the parts not explained by the regression model, are used to determine anomaly scores. The magnitude of these residuals, given a certain confidence level, serves as the anomaly score for test data. Non-parametric techniques, on the other hand, utilize non-parametric statistical models, meaning the model structure is not predefined but inferred from the data. These techniques make fewer assumptions about the data, such as the smoothness of density. For example: Histogram-Based: These techniques use histograms to maintain a profile of normal data, also known as frequency-based or counting-based methods.

B. Pattern-Based Techniques

Prior research addresses the challenge of developing robust semantic scene models for activity analysis in traffic scenarios. This involves the classification of detected moving objects as pedestrians or vehicles using a co-trained classifier that leverages multiview information. A graph-based approach is introduced to cluster these motion patterns by parameterizing trajectories and dividing the image into blocks, treated as nodes in the graph. On the other hand, Gaussian mixture models (GMM) are employed to extract primary motion patterns in each block, and a graph cut algorithm is used to group these patterns resulting in clustering of trajectories to learn semantic models [18]. Simultaneously, in a study on activity pattern recognition, a spatiotemporal feature descriptor named Histograms of Optical Flow Orientation and Magnitude (HOFM) is used to capture information from cuboids, encoding the magnitude and orientation of optical flow separately into histograms. The proposed descriptor enables a simple nearest-neighbor search to identify whether a given pattern should be classified as an abnormal event [19].

C. Machine learning techniques Techniques

In a study focusing on time-efficient anomaly detection and localization, a spatial-temporal cascade autoencoder (ST-CaAE) consisting of two neural networks, namely a spatial-temporal adversarial autoencoder (ST-AAE) and a spatial-temporal convolutional autoencoder (ST-CAE). A two-stream framework is employed to fuse appearance and motion cues, utilizing gradient and optical flow cuboids as inputs for each stream. The proposed ST-CaAE is evaluated on three public datasets, demonstrating superior performance compared to state-of-the-art methods, and showcasing its effectiveness in leveraging both spatial and temporal cues for anomaly detection [20]. Wu et al. identified a limitation in the baseline

MAE, which heavily relies on spatial cues and ignores temporal relations for frame reconstruction, leading to suboptimal temporal matching representations. To address the issue, Drop MAE is implemented, performing spatial-attention dropout during frame reconstruction to enhance temporal correspondence learning [21].

Convolutional autoencoders (CAEs) have emerged as a prominent approach for learning hierarchical representations of video data and detecting anomalies therein (Lu et al., 2013; Hasan et al., 2016). CAEs consist of two main components: an encoder and a decoder. The encoder compresses the input video frames into a latent space representation, while the decoder reconstructs the original frames from the latent representations. Anomalies are detected based on the reconstruction error, i.e., the disparity between the input frames and their reconstructed counterparts. Notably, CAEs leverage the spatial locality and translation invariance properties of convolutional neural networks (CNNs) to capture spatial correlations within video data effectively.

Regenerative adversarial networks (GANs) have garnered attention for their ability to generate realistic samples from a given data distribution (Ravanbakhsh et al., 2017; Akcay et al., 2018). In the context of video anomaly detection, GANs can be leveraged to learn representations of normal video sequences and identify anomalies as deviations from the learned distribution. Specifically, GAN-based anomaly detection models consist of a generator network that generates normal video frames and a discriminator network that distinguishes between real and generated frames. Anomalies are detected based on the discriminator's ability to discriminate between real and generated frames. Recurrent neural networks (RNNs) with Long Short-Term Memory (LSTM) cells offer a powerful framework for modeling temporal dependencies within video sequences (Pimentel et al., 2014; Sabokrou et al., 2018). Unlike feedforward neural networks, RNNs maintain internal state representations that capture sequential information over time. LSTM cells, a variant of RNNs, are designed to mitigate the vanishing gradient problem and facilitate the learning of long-term dependencies. In the context of video anomaly detection, LSTM-based models encode temporal dynamics by processing sequential frames iteratively and detecting anomalies based on deviations from learned temporal patterns.

D. Hyperspectral Anomaly Detection Techniques

Focusing on hyperspectral anomalies, Lin et al. aimed to overcome the contamination of the potential anomaly dictionary by background pixels in existing methods. This involves adaptive inner window–based saliency detection to create a coarse binary map and background estimation network generates a fine binary map. Both maps guide the construction of pure background and potential anomaly dictionaries based on superpixels from the first stage [22]. Improving this, a more novel approach method is demonstrated in a study that focuses on constructing an effective background dictionary and utilizing sparse representation, considering the properties of anomalies in both spectral and spatial domains [23]. Different from the above two studies, Lv et al. proposed a hyperspectral anomaly detection (HAD) method based on a spatial-spectral joint approach using a two-branch 3D convolutional autoencoder and spatial filtering. The proposed method uses a two-branch 3D convolutional autoencoder to fully extract spatial-spectral joint features and spectral interband features, achieving a high area under the curve (AUC) value above 0.9, surpassing those of other methods [24].

VI. Conclusion

This survey evaluated the efficiency of various anomaly detection models using publicly available datasets of varying sizes. We assessed the performance of multiple machine learning algorithms in detecting anomalies in video streams, comparing a range of algorithms employing distinct anomaly detection methodologies. Our analysis focuses on two main aspects: a) the accuracy of these algorithms in identifying anomalous events, and b) computational efficiency, measured by the time required for both training and testing phases. The survey also emphasizes the importance of tailoring anomaly detection systems to specific datasets and frame sizes. Overall, we found that AI-VAD is the most accurate model while only having slightly higher computational overhead compared to the fastest model. We hope our study provides insights into the model, training dataset size, and frame size selections for other VAD users.

References

[1] Chandola, Varun& Banerjee, Arindam & Kumar, Vipin. (2009). Anomaly Detection: A Survey. ACM Comput. Surv.. 41. 10.1145/1541880.1541882.

[2] Yuxing Yang, Zeyu Fu, Syed Mohsen Naqvi, Abnormal event detection for video surveillance using an enhanced two-stream fusion method, Neurocomputing, Volume 553, 2023, 126561, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2023.126561.

[3] Yang Wang, Tianying Liu, Jiaogen Zhou, Jihong Guan, Video anomaly detection based on spatio-temporal relationships among objects, Neurocomputing, Volume 532, 2023, Pages 141-151, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2023.02.027.

[4] Khan, A.A.; Nauman, M.A.; Shoaib, M.; Jahangir, R.; Alroobaea, R.; Alsafyani, M.; Binmahfoudh, A.; Wechtaisong, C. Crowd Anomaly Detection in Video Frames Using Fine-Tuned AlexNet Model. Electronics 2022, 11, 3105. https://doi.org/10.3390/electronics11193105

[5] Le, Viet-Tuan & Kim, Yong-Guk. (2022). Attention-based residual autoencoder for video anomaly detection. Applied Intelligence. 52. 1-15. 10.1007/s10489-022-03613-1.

[6] Reiss, T., & Hoshen, Y. (2022). Attribute-based Representations for Accurate and Interpretable Video Anomaly Detection. ArXiv. /abs/2212.00789

[7] Park, C., Cho, M., Lee, M., & Lee, S. (2021). FASTANO: Fast Anomaly Detection via Spatio-temporal Patch Transformation. ArXiv. /abs/2106.08613

[8] Liu, W., Chang, H., Ma, B., Shan, S., & Chen, X. (2023). Diversity-Measurable Anomaly Detection. ArXiv. /abs/2303.05047

[9] Thudumu, S., Branch, P., Jin, J. et al. A comprehensive survey of anomaly detection techniques for high dimensional big data. J Big Data 7, 42 (2020). https://doi.org/10.1186/s40537-020-00320-x

[10] Shuhan Yi, Zheyi Fan, Di Wu, Batch feature standardization network with triplet loss for weakly-supervised video anomaly detection, Image and Vision Computing, Volume 120, 2022, 104397, ISSN 0262-8856, https://doi.org/10.1016/j.imavis.2022.104397.

[11] Shen, Guodong & Ouyang, Yuqi & Sanchez, Victor. (2022). Video Anomaly Detection via Prediction Network with Enhanced Spatio-Temporal Memory Exchange. 3728-3732. 10.1109/ICASSP43922.2022.9747376.

[12] Altayeva, Aigerim & Omarov, Nurzhan & Tileubay, Sarsenkul & Zhaksylyk, Almash & Bazhikov, Koptleu & Kambarov, Dastan. (2023). Convolutional LSTM Network for Real-Time Impulsive Sound Detection and Classification in Urban Environments. International Journal of Advanced Computer Science and Applications. 14. 10.14569/IJACSA.2023.0141164.

[13] Luque Sánchez F, Hupont I, Tabik S, Herrera F. Revisiting crowd behaviour analysis through deep learning: Taxonomy, anomaly detection, crowd emotions, datasets, opportunities and prospects. Inf Fusion. 2020 Dec;64:318-335. doi: 10.1016/j.inffus.2020.07.008. Epub 2020 Jul 29. PMID: 32834797; PMCID: PMC7387290.

[14] Manoharan, Madhiarasan & Louzazni, Mohamed. (2022). Analysis of Artificial Neural Network: Architecture, Types, and Forecasting Applications. Journal of Electrical and Computer Engineering. 2022. 1-23. 10.1155/2022/5416722.

[15] Berahmand, K., Daneshfar, F., Salehi, E.S. et al. Autoencoders and their applications in machine learning: a survey. Artif Intell Rev 57, 28 (2024). https://doi.org/10.1007/s10462-023-10662-6

[16] Zhaoyang Niu, Guoqiang Zhong, Hui Yu, A review on the attention mechanism of deep learning, Neurocomputing, Volume 452, 2021, Pages 48-62, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2021.03.091.

[17] Xuanfan Wu, R. Jain, Techniques for Statistical and Structural Anomaly Detection, 2017. https://www.cse.wustl.edu/jain/cse567-17/ftp/mttad/index.html.

[18] Zhang, T.; Lu, H.; Li, S.Z. Learning semantic scene models by object classification and trajectory clustering. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 1940–1947.

[19] R. V. H. M. Colque, C. A. C. Júnior and W. R. Schwartz, "Histograms of Optical Flow Orientation and Magnitude to Detect Anomalous Events in Videos," 2015 28th SIBGRAPI Conference on Graphics, Patterns and Images, Salvador, Brazil, 2015, pp. 126-133, doi: 10.1109/SIBGRAPI.2015.21.

[20] N. Li, F. Chang and C. Liu, "Spatial-Temporal Cascade Autoencoder for Video Anomaly Detection in Crowded Scenes," in IEEE Transactions on Multimedia, vol. 23, pp. 203-215, 2022, doi: 10.1109/TMM.2020.2984093.

[21] Q. Wu, et al., "DropMAE: Masked Autoencoders with Spatial-Attention Dropout for Tracking Tasks," in 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023 pp. 14561-14571. doi: 10.1109/CVPR52729.2023.01399

[22] Lin, S.; Zhang, M.; Cheng, X.; Wang, L.; Xu, M.; Wang, H. Hyperspectral Anomaly Detection via Dual Dictionaries Construction Guided by Two-Stage Complementary Decision. Remote Sens. 2022, 14, 1784. https://doi.org/10.3390/rs14081784

[23] Zhu, L.; Wen, G. Hyperspectral Anomaly Detection via Background Estimation and Adaptive Weighted Sparse Representation. Remote Sens. 2018, 10, 272. https://doi.org/10.3390/rs10020272

[24] Lv, S.; Zhao, S.; Li, D.; Pang, B.; Lian, X.; Liu, Y. Spatial–Spectral Joint Hyperspectral Anomaly Detection Based on a Two-Branch 3D Convolutional Autoencoder and Spatial Filtering. Remote Sens. 2023, 15, 2542. https://doi.org/10.3390/rs15102542

[25] Bahri, M., Salutari, F., Putina, A. et al. AutoML: state of the art with a focus on anomaly detection, challenges, and research directions. Int J Data Sci Anal 14, 113–126 (2022). https://doi.org/10.1007/s41060-022-00309-0

[26] H. Xu, S. Xu, and W. Yang, 'Unsupervised industrial anomaly detection with diffusion models', Journal of Visual Communication and Image Representation, vol. 97, p. 103983, 2023.