Evaluating the Effectiveness of Video Anomaly Detection in the Wild
Online Learning and Inference for Real-world Deployment

Shanle Yao
University of North Carolina Charlotte
syao@charlotte.edu

Ghazal Alinezhad Noghre
University of North Carolina Charlotte
galinezh@charlotte.edu

Armin Danesh Pazho
University of North Carolina Charlotte
adaneshp@charlotte.edu

Hamed Tabkhi
University of North Carolina Charlotte
htabkhiv@charlotte.edu

Abstract

Video Anomaly Detection (VAD) identifies unusual activities in video streams, a key technology with broad applications ranging from surveillance to healthcare. Tackling VAD in real-life settings poses significant challenges due to the dynamic nature of human actions, environmental variations, and domain shifts. Many research initiatives neglect these complexities, often concentrating on traditional testing methods that fail to account for performance on unseen datasets, creating a gap between theoretical models and their real-world utility. Online learning is a potential strategy to mitigate this issue by allowing models to adapt to new information continuously. This paper assesses how well current VAD algorithms can adjust to real-life conditions through an online learning framework, particularly those based on pose analysis, for their efficiency and privacy advantages. Our proposed framework enables continuous model updates with streaming data from novel environments, thus mirroring actual world challenges and evaluating the models' ability to adapt in real-time while maintaining accuracy. We investigate three state-of-the-art models in this setting, focusing on their adaptability across different domains. Our findings indicate that, even under the most challenging conditions, our online learning approach allows a model to preserve 89.39% of its original effectiveness compared to its offline-trained counterpart in a specific target domain.

1. Introduction

Video Anomaly Detection (VAD) is an essential area within computer vision, tasked with pinpointing atypical behaviors in specific scenes. It plays a pivotal role in a variety of sectors, including surveillance and healthcare, where identifying deviations from the norm is crucial. The scope of anomalies it addresses is wide, with a significant emphasis on detecting anomalies centered around human activities. VAD techniques are primarily categorized into two types: pixel-based methods, which analyze the raw data of pixels, and pose-based methods, which focus on the dynamics of joints and bodily movements. The latter is especially beneficial in scenarios where privacy is a major concern, as it prioritizes the analysis of skeletal movements over detailed pixel imagery. This approach minimizes privacy concerns and plays a vital role in mitigating biases, particularly those affecting marginalized communities, thus providing a fairer and more privacy-aware anomaly detection method.

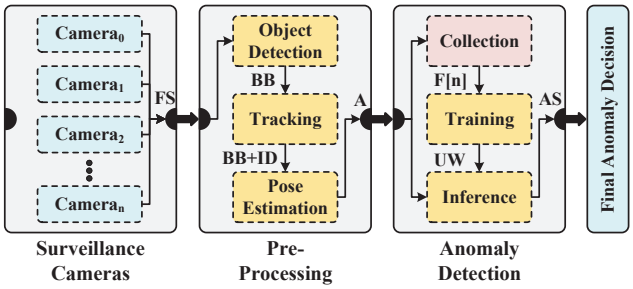


Figure 1. A conceptual overview of an end-to-end system with online unsupervised anomaly detection training. Frame sequences (FS) collected from surveillance cameras pass through a preprocessing phase to extract necessary annotations (A), including bounding boxes (BB), tracking information (ID), and pose information. This information consequently goes through anomaly detection, which is used for real-time inference and collection. The collection algorithm collects enough frame annotations (F[n]) for training. After training, Updated Weights (UW) are replaced for the next inference step.

Recognizing the wide range of normal and anomalous behaviors, a reflection of human behavior's complex na-

ture poses a significant challenge to the generalizability of VAD models in real-world scenarios. The inability of existing datasets to fully capture this breadth significantly hampers the applicability of VAD models outside laboratory conditions. This limitation has spurred a shift towards unsupervised learning in VAD, where models learn from unlabelled data, recognizing normal behavior patterns and identifying deviations without needing predefined anomalies. This move towards unsupervised methods represents a pivotal adaptation, promising to enhance the robustness and relevance of VAD models in diverse and unpredictable real-world environments by accommodating the full spectrum of human behaviors.

Nonetheless, human-centric VAD faces other inherent challenges, such as the context-specific nature of what constitutes an anomaly. This variability means that behaviors considered normal in one setting might be deemed anomalous in another. For instance, punching in a gym is typical, whereas the same behavior in a mall would be considered anomalous. Domain shift hurdle is more significant when transitioning anomaly detection models from controlled experimental settings to real-world applications of anomaly detection, in which anomalous behaviors are deviations from established norms of behavior. In real-world environments, ostensibly normal behaviors can often be misconstrued as anomalous by these models due to discrepancies arising from various factors, such as camera angles and distance, which were not accounted for during the training phase. Such discrepancies can lead to an inflated rate of false positives, substantially undermining VAD systems' practical utility and accuracy in natural settings. This vulnerability to domain shift underscores the need for more adaptive, context-aware machine learning models capable of dynamically recalibrating their parameters to the nuances of their operational environment, thereby enhancing their effectiveness and reliability in diverse real-world applications.

Existing video anomaly detection (VAD) methodologies often rely on offline learning paradigms, which inherently limit their ability to adapt to real-world situations' dynamic and unpredictable nature. The shift towards online learning for VAD anomaly detection is not merely a trend but a necessary evolution to address these limitations. By continually updating their knowledge base with new, unlabeled data encountered in their operational environment, online learning algorithms embody the adaptability required to tackle the complex nature of human behavior and the broad spectrum of what may be considered abnormal in different contexts. This capability to learn from streaming data in real-time allows for the detection system to remain relevant and practical, even as the nature of anomalies evolves.

To our knowledge, no existing research has shied away from online learning VAD, specifically within the domain of pose-based VAD, marking a significant gap in the literature. It is also important to separate the concept of "online learning" VAD from the broader concept of "online anomaly detection" as outlined in various studies focused on pixel-based analysis, such as those by [8, 16, 27], where the term is typically associated with the capacity for real-time decision-making. Unlike mere online anomaly detection, which implies immediate processing without learning from new data, online learning involves the algorithm's ability to continuously adapt and update its understanding, enhancing its predictive accuracy over time. Overall, we observe a notable oversight in mainstream VAD research, where the inherent benefits and necessities of online learning for anomaly detection are often overshadowed by results derived from the offline learning paradigm.

This study rigorously assesses the effectiveness and adaptability of current pose-based Violence Detection (VAD) methodologies, focusing on their application in online VAD environments that simulate real-life conditions. Our aim is not only to highlight the strengths and weaknesses of each model in the context of online VAD but also to reveal their adaptability and efficiency in transitioning to new domains. To this end, we design and implement an online learning framework that mirrors the actual world challenges of VAD. The proposed online learning VAD framework enables continuous model updates from novel environments, thereby testing their ability to adapt in real-time to new domains while maintaining high levels of accuracy and privacy advantages. We analyze the performance and efficiency of three state-of-the-art models, GEPC[18], STG-NF[12], and TSGAD[20], in an execution environment that emulates unseen streaming data in the real world. The findings from our experiments demonstrate the proposed online learning frameworks' effectiveness, where models are able to preserve between 89.39% and 99.20% of their performance, as measured by the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), in both the worst and best-case scenarios, respectively, relative to models trained offline in the target domain.

In summary, this study presents the following contributions:

- Development of an online learning framework tailored for pose-based anomaly detection.
- Evaluation of traditional offline learning methodologies to discover their efficacy and limitations within unseen online scenarios.
- Highlighting the research gaps and looking at the evolution and potential breakthroughs of video anomaly detection in the wild.

2. Related Works

Historically, traditional VAD methods predominantly utilized handcrafted features[5, 6, 24], which, while effective

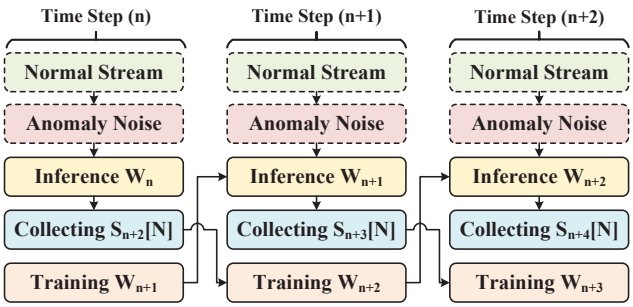


Figure 2. A conceptual overview of an end-to-end system with online unsupervised anomaly detection training.

in controlled settings, often faced challenges in generalizing to the diverse conditions of real-world applications. With the advent of deep learning, a paradigm shift occurred in VAD, leading to its classification into two primary strategies: pixel-based and pose-based approaches. Pixel-based methods[1–3, 10, 11, 14, 22, 25] analyze the raw pixel data to detect anomalies, whereas pose-based approaches concentrate on extracted skeletal information and monitoring the movements of individuals within the scene. This study specifically focuses on the exploration of pose-based VAD techniques. Consequently, we will provide a detailed examination of pose-based methods in the subsequent sections, highlighting their operational mechanisms and advantages.

In unsupervised learning environments, strategies are developed to establish tasks that inherently encourage models to assimilate normal behavior patterns. These tasks predominantly involve reconstructing the current timestep[4, 15, 18, 23] or predicting future or past sequences[13, 21, 26]. Several studies[19, 20] employ a multi-branch framework, leveraging both objectives to enhance anomaly detection capabilities, showcasing the diversity and adaptability of unsupervised methods in identifying deviations from established norms.

Among the pose-based models, GEPC[18], TSGAD[20], and STG-NF[12] distinguish themselves by having their source code publicly accessible. Consequently, these models were selected for our experimental analysis. The GEPC model[18] encodes input pose sequences into a latent graph space, followed by a clustering process. Anomalies are detected through a Dirichlet mixture model that evaluates the distribution of cluster-based action normalities. The STG-NF model[12] leverages normalizing flows to map input pose sequences into a standard distribution within latent space, with the degree of deviation from this distribution indicating potential anomalies. The TSGAD model[20] analyzes anomalies by examining both pose and trajectory data. It employs a graph variational autoencoder for pose analysis, generating scores based on deviations from the model's learned distribution, while the trajectory analysis predicts

Table 1. Number of Poses comparison for ShanghaiTech[17], CHAD[7], and different cameras views from CHAD [7]

Dataset	Number of Poses		
	Train	Test	Total
ShanghaiTech[17]	257,650	37,845	295,495
CHAD[7]	802,167	119,867	922,034
CHAD[7] Cam 0	111,230	21,074	132,304
CHAD[7] Cam 1	213,991	35,502	249,493
CHAD[7] Cam 2	245,436	35,727	281,163
CHAD[7] Cam 3	231,510	27,564	259,074

future movements, comparing these predictions to actual trajectories to produce a trajectory-based anomaly score. The overall anomaly detection is then determined by combining the scores from both the pose and trajectory analyses.

Multiple pixel-based investigations[8, 16, 27] interpret "online anomaly detection" differently from our discussion in Sec. 1. They consider it as the ability of a model to dynamically render decisions in real time. Contrarily, our manuscript delves into online learning for anomaly detection, highlighting the model's continuous adjustment and training with streaming data to enhance detection accuracy. Furthermore, while several studies[7, 9] explore cross-domain evaluation under a zero-shot framework, they fall short of suggesting any strategies for domain adaptation, particularly concerning streaming data.

3. Methodology

To explore the viablilty of the online unsupervised anomaly detection training application with frameworks using existing pose-based algorithms, a three-stage pipeline was developed, emulating real-world scenarios. This pipelin design illustrated in Fig. 2 comprises an inference stage, a collection stage, and a training stage. At the inference stage, algotithm with source pretrained weight is used to process input stream to identify normal behavioral patterns with potential existing anomaly noise. Subsequently, sequences detected as 'normal' are formatted and collected for training purposes within the collection stage. Once the training data is accumulated to pre-defined volume, the training stage will start fine-tuning the pre-trained weights to adapt to the target domain weight with evaluation. Notably, the inference stage's weights are updated with these refined weights with a lag of two time steps from the initial state because of the nature of such a pipeline design.

3.1. Inference Methodology

3.1.1 Input Stream

As shown and discussed in Fig. 1 from Sec. 1, prerequisite for real-world applications of online anomaly detection is the capability to accurately detect and track indi-

vidual figures within video streams to extract sequential pose information. This process could be easily influenced by noise, primarily due to the variability introduced by diverse streaming conditions such as location, camera angles, and coverage area.These factors contribute to the domain-specific nature of anomaly detection tasks. For instance, jogging or running, which is normal behavior in a park setting, may be considered anomalous within a grocery store environment.

To minimize the impact of such variability and enhance the precision of extracted pose sequences, pose-based datasets are employed to emulate real-world streaming conditions. Among existing pose based anomaly datasets with continues pose sequences, CHAD dataset[7] offers a comprehensive collection of 922,034 count of pose instances captured from four different camera views and ShanghaiTech dataset[17], widely utilized resource, provides a total of 295,495 pose instances with thirteen camera views as shown in Tab. 1.

Despite ShanghaiTech's[17] diversity, including thirteen distinct scenes, its limited pose instance count per scene constrains its utility for exploration into the feasibility of online anomaly detection across varied domains. In this study, ShanghaiTech[17] is trained as the initial source weight for different models and four different camera views from CHAD[7] are used to replicate the stream inputs from four different domains. Notably, Cam 1 to 3 in CHAD[7] have at least 200k poses, providing a substantial volume of data conducive to effective online training and all the train set data are augmented with anomalous pose data extracted from the test set, at a ratio of 9.5:0.5 to mirror the nature of anomalies in typical surveillance scenarios. This strategy ensures that models are exposed to both normal and anomalous patterns.

3.1.2 Detection

The pre-processed pose sequences undergo actual inference phase, where they are analyzed within distinct temporal windows size to the architectural requisites of specific models—30 frames for TSGAD[20] and GEPC [18], and 24 frames for STG-NF[12] because of specific design. This window size is selected to align with the standard frame rate of surveillance cameras. This alignment ensures that the models are not only compatible with standard surveillance video characteristics but also optimized for detecting anomalies within a temporal context that mirrors real-world surveillance scenarios.

3.2. Collection Methodology

One primary limitation when utilizing pre-existing datasets for such online anomaly detection design is the constrained volume of pose instances. Moreover, a significant uncertainty in it is the quantity of data that can be classified as "normal" at inference stage with keeping updated weights. To mitigate these challenges and more accurately mirror real-world environments, the training data for each camera view is strategically partitioned into twelve distinct subsets. Once each subset has been inferenced and collected, the training phase would start training. This collecting mechanism allows for the systematic analysis of the models' adaptability and performance across varying conditions, effectively capturing the evolution of domain-specific characteristics.

3.3. Training Methodology

In the training phase for each algorithm, default settings were retained with the exception of window and stride sizes. The window size wasadjusted to match the frame rate of the input stream for each specific model, ensuring temporal alignment with the dynamics of the observed activities in one second. The stride size was uniformly set to 1 across all models, such decision aimed at achieving balance within the pipeline's design, particularly to optimize the efficiency and responsiveness of the inference stage.

Leveraging pre-existing datasets for fine-tuning offers the distinct advantage of enabling evaluation of each model's performance with different input subsets using the respective test sets. To thoroughly evaluate the performance of models and gain a multifaceted understanding of their strengths and weaknesses, especially in real-world scenarios, we selected a comprehensive suite of metrics. These metrics—Area Under the Receiver Operating Characteristic Curve (AUC-ROC), Area Under the Precision-Recall Curve (AUC-PR), and Equal Error Rate (EER)—are utilized to assess the efficacy of the models from complementary perspectives, ensuring a holistic analysis.

AUC-ROC is a performance measurement for binary classification problems at various threshold settings. The ROC curve is a graphical representation that plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at different thresholds, essentially showing the trade-off between sensitivity and specificity. The AUC represents the degree to which a model is capable of distinguishing between classes. The higher the AUC, the better the model is at discriminating between the classes. The AUC-ROC metric, while useful in many scenarios, can indeed be misleading in the context of highly imbalanced datasets, such as those typically found in anomaly detection. Thus, it is vital to use it in combination with other metrics to analyze the efficacy of anomaly detection models thoroughly.

AUC-PR is a metric that evaluates the trade-off between precision (the proportion of true positive results in all positive predictions) and recall (the proportion of true positive results in all actual positives) for different threshold values, without being affected by the distribution of class la-

Table 2. Evaluation of Models Pre-trained on Shanghaitech[17]

Model	Test	AUC-ROC	AUC-PR	EER
TSGAD[20]	ShanghaiTech[17]	0.742	0.602	0.315
	CHAD[7] Cam 0	0.549	0.550	0.494
	CHAD[7] Cam 1	0.561	0.487	0.467
	CHAD[7] Cam 2	0.477	0.382	0.507
	CHAD[7] Cam 3	0.638	0.696	0.498
GEPC[18]	ShanghaiTech[17]	0.729	0.614	0.318
	CHAD[7] Cam 0	0.623	0.608	0.409
	CHAD[7] Cam 1	0.622	0.491	0.407
	CHAD[7] Cam 2	0.592	0.494	0.437
	CHAD[7] Cam 3	0.680	0.693	0.370
STG-NF[12]	ShanghaiTech[17]	0.851	0.869	0.230
	CHAD[7] Cam 0	0.582	0.638	0.459
	CHAD[7] Cam 1	0.550	0.634	0.488
	CHAD[7] Cam 2	0.498	0.608	0.495
	CHAD[7] Cam 3	0.633	0.573	0.430

bels. This makes AUC-PR especially valuable for analyzing the efficacy of anomaly detection models, where data is often imbalanced. The AUC-PR encapsulates the model's ability to identify the rare positive cases (anomalies) correctly while minimizing false positives, which is crucial in anomaly detection scenarios where the primary concern is the accurate detection of these rare events.

EER represents the point at which the FPR and False Negative Rate (FNR) are equal. In the context of anomaly detection, the EER offers a singular, balanced threshold at which the likelihood of incorrectly labeling normal behavior as an anomaly equals the likelihood of failing to detect an actual anomaly. The EER aids in identifying the optimal operating point of a model, thereby facilitating more informed decisions in the deployment of anomaly detection systems.

4. Experiments and Evaluation

As mentioned in Sec. 3, ShanghaiTech[17] serves as the foundation for initial model training, providing pre-trained weights representative of the source domain. The efficacy of these initial weights, derived from various models, was evaluated across multiple test sets, including those from ShanghaiTech[17] and diverse camera views within the CHAD dataset[7]. The comparative performance analysis is summarized in Tab. 2

Notably, STG-NF[12] emerged as the best model within the ShanghaiTech[17] test set, outperforming others across all evaluated metrics, thereby affirming its status as a state-of-the-art (SoTA) algorithm. TSGAD(pose only)[20], which originally has pose and path branch, ranked second, showing its strength in AUC-ROC and EER metrics. However, when subjected to test sets from different domains, both STG-NF[12] and TSGAD(pose only)[20] experienced significant declines in accuracy, underscoring the challenge of domain shift. In contrast, GEPC[18] shows relatively stable performance across varied test environments, achieving the highest overall scores in all CHAD[7] camera views. Following with TSGAD(pose only)[20] performs the second in CHAD[7] CAM 1 and CAM 3 and third in CAM 0 and CAM 2.

This pattern of results, notable drop in accuracy in different domain, aligns with the expectation highlighted in Sec. 3 that anomaly detection algorithms are highly sensitive to contextual variations.

Tab. 3 presents a comparison of model evaluation result across three distinct training scenarios: baseline (no training), average performance across twelve online training, and outcomes following complete offline training. The evaluation spans four domains (Cam 0 to Cam 3), each offering unique insights into the adaptability and efficacy of the models under consideration: TSGAD(pose only)[20]. GEPC[18], and STG-NF[12].

Cam 0 In the online training scenario, GEPC[18] emerged as the top performer, with STG-NF[12] closely following, showing their adaptability to this specific domain. Conversely, the offline training scenario highlighted TSGAD's[20] fine performance, with notable improvements observed from the baseline to the online training phase, indicating TSGAD's[20] learning capability in this domain.

Cam 1 GEPC[18] consistently achieved good results across most evaluation cases, yet an unexpected drop in performance was observed transitioning from the no training to the online training scenario. This suggests a potential anomaly in GEPC's[18] learning curve or an overfitting issue within this specific context. STG-NF[12], while generally underperforming, exhibited a further decline in the offline training scenario, raising questions about its adaptability and efficacy.

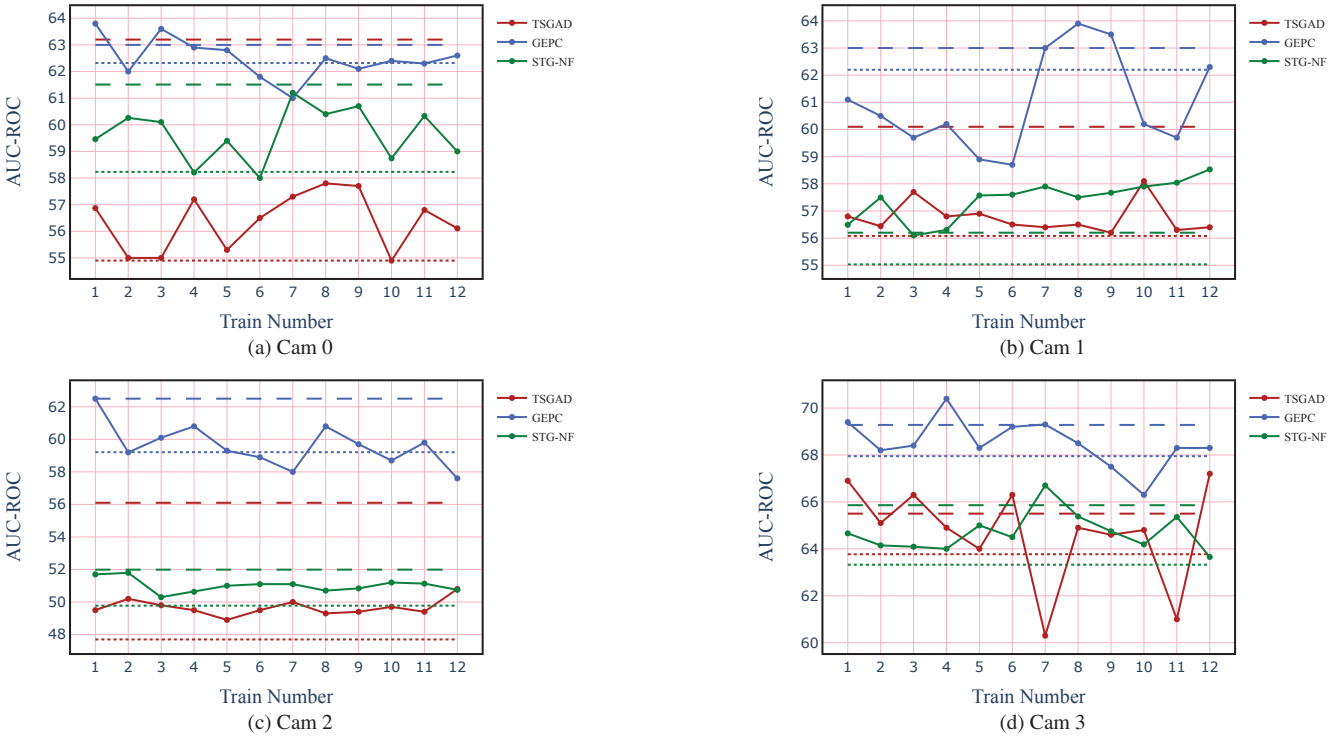
Cam 2 Despite a general decline in model performance in this domain, GEPC[18] maintained its lead, followed by TSGAD(pose only)[20] and STG-NF[12]. This consistent pattern across models suggests inherent challenges within the Cam 2 domain.

Cam 3 TSGAD(pose only)[20] and STG-NF[12] demonstrated remarkably similar performance metrics, indicating a convergence in their learning capabilities for this camera view. GEPC[18], maintaining its pattern, stood out with its performance, reinforcing its robustness across varied conditions.

Detailed analyses of the models' performance are visualized in Fig. 3, Fig. 4, and Fig. 5, each illustrating variations in evaluation metrics across training numbers. GEPC[18] consistently exhibits strong performance in AUC-ROC and EER metrics, while STG-NF[12] excels in AUC-PR across multiple training stages. Notably, STG-NF[12] model consistently shows high EER and low AUC-ROC, yet achieving high AUC-PR. Despite this, STG-NF[12] could be use-

Table 3. Evaluation of TSGAD[20], GEPC[18], and STG-NF[12] on different camera views from CHAD[7] in cases of Baseline (No Train), Online Training, and Offline Training.

Model	Case	AUC-ROC	AUC-PR	EER	Model	Case	AUC-ROC	AUC-PR	EER
Cam 0					**Cam 1**				
TSGAD[20]	No Train	0.549	0.550	0.494	**TSGAD**[20]	No Train	0.561	0.487	0.467
	Online	0.565	0.548	0.466		Online	0.568	0.488	0.461
	Offline	0.632	0.608	0.409		Offline	0.601	0.506	0.430
GEPC[18]	No Train	0.623	0.608	0.409	**GEPC**[18]	No Train	0.622	0.491	0.407
	Online	0.625	0.625	0.419		Online	0.609	0.495	0.422
	Offline	0.630	0.635	0.415		Offline	0.630	0.494	0.383
STG-NF[12]	No Train	0.582	0.638	0.459	**STG-NF**[12]	No Train	0.550	0.634	0.488
	Online	0.596	0.651	0.442		Online	0.574	0.661	0.456
	Offline	0.615	0.667	0.429		Offline	0.562	0.654	0.464
Cam 3					**Cam 4**				
TSGAD[20]	No Train	0.477	0.382	0.507	**TSGAD**[20]	No Train	0.638	0.696	0.498
	Online	0.497	0.393	0.493		Online	0.647	0.706	0.397
	Offline	0.561	0.490	0.422		Offline	0.655	0.656	0.385
GEPC[18]	No Train	0.592	0.494	0.437	**GEPC**[18]	No Train	0.680	0.693	0.370
	Online	0.596	0.505	0.429		Online	0.685	0.704	0.350
	Offline	0.625	0.510	0.391		Offline	0.693	0.695	0.338
STG-NF[12]	No Train	0.498	0.608	0.495	**STG-NF**[12]	No Train	0.633	0.573	0.423
	Online	0.510	0.621	0.495		Online	0.647	0.585	0.399
	Offline	0.520	0.626	0.495		Offline	0.659	0.600	0.392



(a) Cam 0 (b) Cam 1

(c) Cam 2 (d) Cam 3

Figure 3. Model AUC-ROC percentage Trend Comparison by Training Number: Long dashes indicate Offline Training, solid lines indicate Online Training, and dots indicate the Baseline (No training).

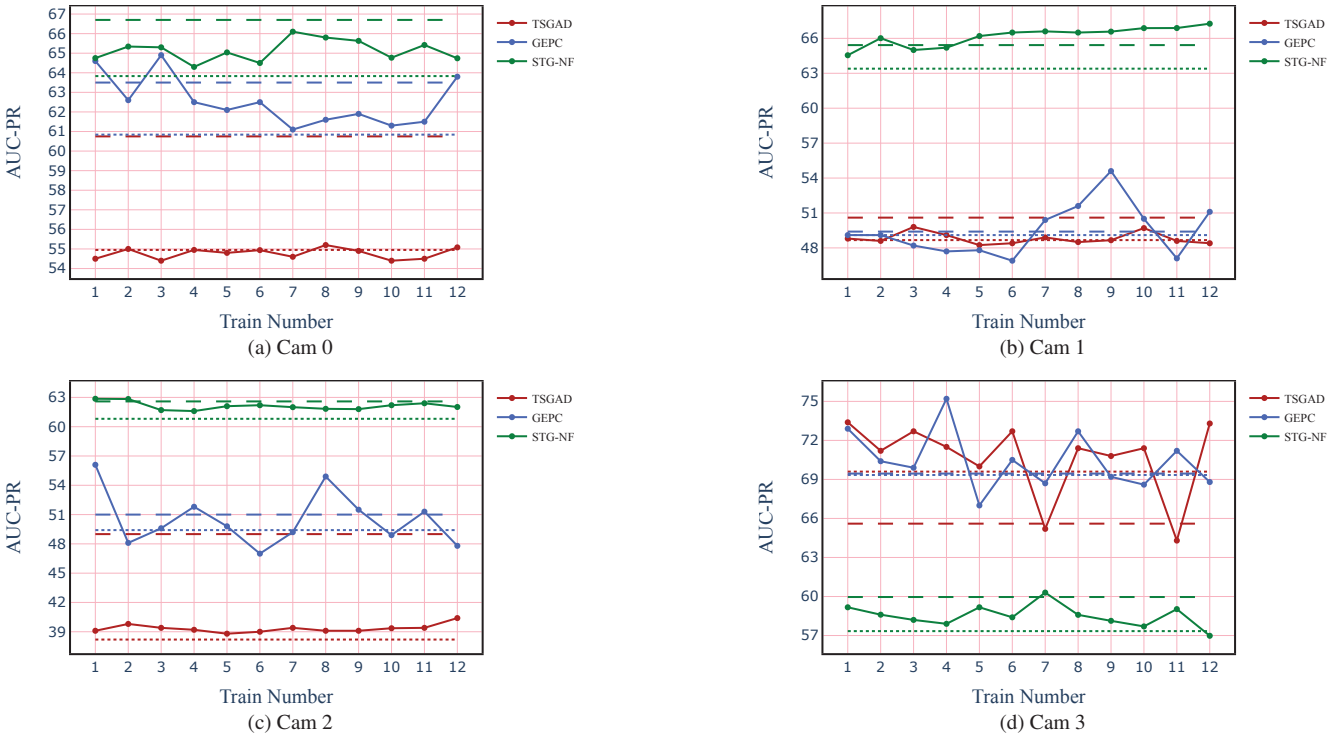


Figure 4. Model AUC-PR percentage Trend Comparison by Training Number: Long dashes indicate Offline Training, solid lines indicate Online Training, and dots indicate the Baseline (No training).

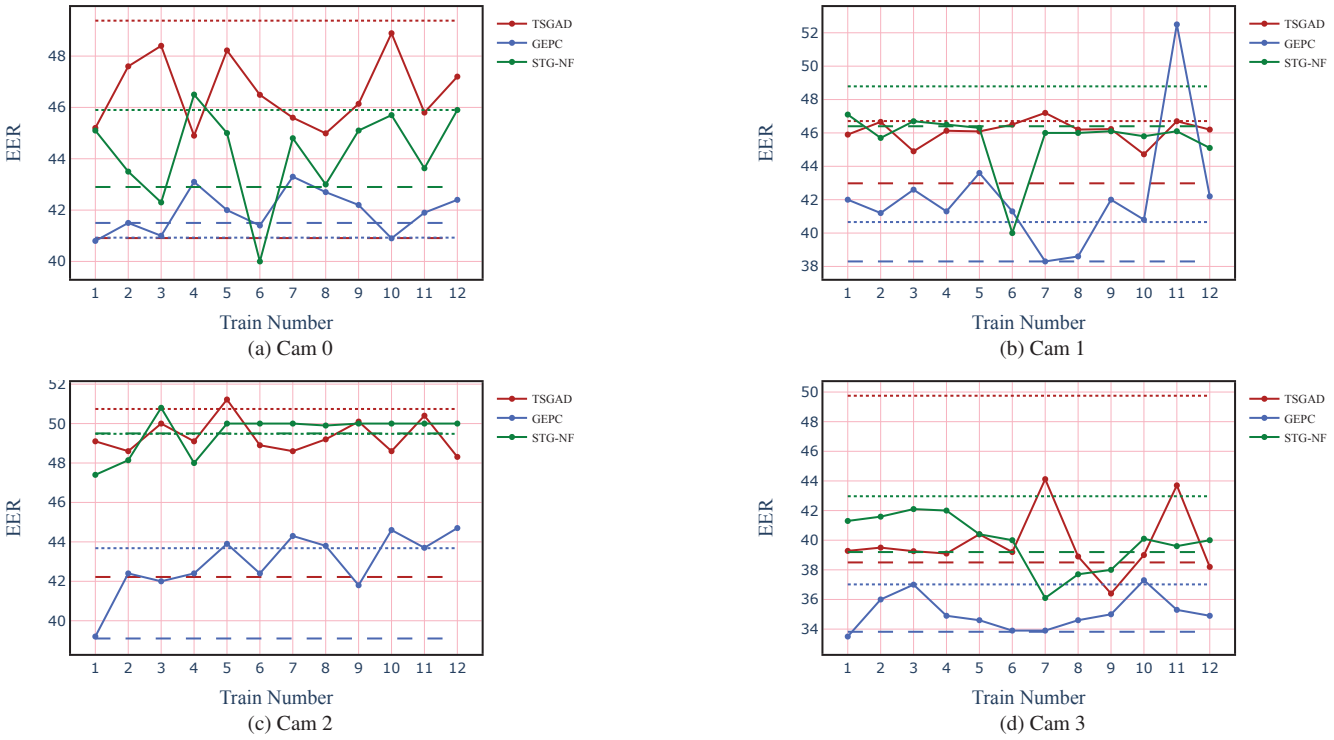


Figure 5. Model EER percentage Trend Comparison by Training Number: Long dashes indicate Offline Training, solid lines indicate Online Training, and dots indicate the Baseline (No training).

ful in situations where minimizing false negatives is more important than avoiding false positives. The model's tendency to mistakenly classify normal instances as anomalies, leading to high EER and low AUC-ROC, may be due to its simplistic assumption of a normal distribution in the latent space, which struggles to capture complex patterns in large datasets like CHAD[7].

While GEPC[18] generally exhibited superior performance across multiple evaluation scenarios, its incremental learning gains from baseline to complete offline training were modest, rarely exceeding a 2% improvement. This phenomenon likely caused by the design of GEPC's[18] algorithm, which incorporates a sequencial training mechanism with encoding, decoding, and clustering phases. The necessity to start each training cycle with encoder fine-tuning, effectively initializing decoder and cluster components without pre-trained weights, might be not ideal for the demands of online anomaly training. This design choice, while beneficial in certain contexts, appears to constrain GEPC's[18] capacity in online anomaly training design.

On the other hand, STG-NF[12] demonstrated incredible performance within the controlled setting of the ShanghaiTech dataset[17] during offline training. However, its adaptability to the varied domains represented in the CHAD dataset[7] was less consistent, with improvements from the baseline to offline training being modest, at approximately 4%. This suggests that while STG-NF[12] is capable of learning and improving, its architecture may not be suited to the diverse and dynamic nature of real-world surveillance scenarios in online anomaly training.

TSGAD(pose only)[20], although not consistently surpassing GEPC[18] in domain-specific evaluations, exhibited notable progressions from baseline through online training to offline training. This trajectory of improvement highlights TSGAD's[20] potential compatibility with online anomaly training frameworks. The model's ability to learn effectively suggests a structural or algorithmic adaptability that could be optimized for the continuous, evolving nature of online anomaly training.

These observations underscore the nuanced relationship between model architecture, training methodology, and domain specificity in anomaly detection. The varied performance and learning trajectories of GEPC[18], STG-NF[12], and TSGAD(pose only)[20] across different training stages and domains illuminate the critical considerations necessary for tailoring anomaly detection models to the specific requirements and challenges of online training environments.

5. Research Questions and Future Directions

This study highlights the viability and potential of employing online training strategy with pose-based anomaly detection models utilizing existing datasets. As mentioned in Sec. 3, the reliance on current datasets introduces uncertainties related to the volume of training data each training time, complicating the exploration of time constraints within online training frameworks.

It highlights the challenges posed by variable training data volumes and underscores the necessity of an integrated end-to-end system for live stream processing and training execution in future research. Despite these challenges, our method retains 89.38% effectiveness relative to offline training, suggesting the potential to not only match but exceed offline training efficacy with further enhancements. Critical to this ambition is the precise measurement of the online learning timing in a balanced system, essential for demonstrating the practicality of such systems in the wild applications. The objective is to identify and surpass the limitations of existing approaches, thereby designing more effective strategies to tackle the domain-specific barrier encountered in the detection of anomalous actions, thereby advancing the state of research in this specialized area.

6. Conclusion

This study underscores the significant challenges inherent in applying Video Anomaly Detection (VAD) in real-world scenarios, due to the dynamic and unpredictable nature of human behavior, environmental contexts, and domain shifts. Through an evaluation of SOTA VAD algorithms within an online learning framework, our research highlights the potential of pose-based approaches to not only address these challenges but also to offer privacy-conscious solutions suitable for practical applications. The adaptability of these models to continuously learn and update from streaming data represents a critical step forward in bridging the gap between theoretical research and real-world applicability. Our findings demonstrate that even under the most challenging conditions, the proposed online learning method enables models to maintain a high degree of effectiveness, retaining up to 89.39% of their original performance. This work paves the way for future research in enhancing the robustness and adaptability of VAD systems, ensuring their reliability and efficacy in the wild environments.

Acknowledgment

This research is supported by the National Science Foundation (NSF) under Award No. 1831795.

References

[1] Ghazal Alinezhad Noghre, Armin Danesh Pazho, Vinit Katariya, and Hamed Tabkhi. Understanding the challenges and opportunities of pose-based anomaly detection. In *Proceedings of the 8th International Workshop on Sensor-Based Activity Recognition and Artificial Intelligence*, New York, NY, USA, 2023. Association for Computing Machinery. 3

[2] Antonio Barbalau, Radu Tudor Ionescu, Mariana-Iuliana Georgescu, Jacob Dueholm, Bharathkumar Ramachandra, Kamal Nasrollahi, Fahad Shahbaz Khan, Thomas B Moeslund, and Mubarak Shah. Ssmtl++: Revisiting self-supervised multi-task learning for video anomaly detection. *Computer Vision and Image Understanding*, 229:103656, 2023.

[3] Dongyue Chen, Lingyi Yue, Xingya Chang, Ming Xu, and Tong Jia. Nm-gan: Noise-modulated generative adversarial network for video anomaly detection. *Pattern Recognition*, 116:107969, 2021. 3

[4] Xiaoyu Chen, Shichao Kan, Fanghui Zhang, Yigang Cen, Linna Zhang, and Damin Zhang. Multiscale spatial temporal attention graph convolution network for skeleton-based anomaly behavior detection. *Journal of Visual Communication and Image Representation*, 90:103707, 2023. 3

[5] Kai-Wen Cheng, Yie-Tarng Chen, and Wen-Hsien Fang. Gaussian process regression-based video anomaly detection and localization with hierarchical feature representation. *IEEE Transactions on Image Processing*, 24(12):5288–5301, 2015. 2

[6] Serhan Coşar, Giuseppe Donatiello, Vania Bogorny, Carolina Garate, Luis Otavio Alvares, and François Brémond. Toward abnormal trajectory and event detection in video surveillance. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(3):683–695, 2016. 2

[7] Armin Danesh Pazho, Ghazal Alinezhad Noghre, Babak Rahimi Ardabili, Christopher Neff, and Hamed Tabkhi. Chad: Charlotte anomaly dataset. In *Scandinavian Conference on Image Analysis*, pages 50–66. Springer, 2023. 3, 4, 5, 6, 8

[8] Keval Doshi and Yasin Yilmaz. Online anomaly detection in surveillance videos with asymptotic bound on false alarm rate. *Pattern Recognition*, 114:107865, 2021. 2, 3

[9] Keval Doshi and Yasin Yilmaz. Towards interpretable video anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2655–2664, 2023. 3

[10] Mariana-Iuliana Georgescu, Antonio Barbalau, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. Anomaly detection in video via self-supervised and multi-task learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12742–12752, 2021. 3

[11] Mariana Iuliana Georgescu, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. A background-agnostic framework with adversarial training for abnormal event detection in video. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):4505–4523, 2021. 3

[12] Or Hirschorn and Shai Avidan. Normalizing flows for human pose anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13545–13554, 2023. 2, 3, 4, 5, 6, 8

[13] Chao Huang, Yabo Liu, Zheng Zhang, Chengliang Liu, Jie Wen, Yong Xu, and Yaowei Wang. Hierarchical graph embedded pose regularity learning via spatio-temporal transformer for abnormal behavior detection. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 307–315, 2022. 3

[14] Xiangyu Huang, Caidan Zhao, Jinhui Yu, Chenxing Gao, and Zhiqiang Wu. Multi-level memory-augmented appearance-motion correspondence framework for video anomaly detection. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2717–2722. IEEE, 2023. 3

[15] Yashswi Jain, Ashvini Kumar Sharma, Rajbabu Velmurugan, and Biplab Banerjee. Posecvae: Anomalous human activity detection. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2927–2934. IEEE, 2021. 3

[16] Hamza Karim, Keval Doshi, and Yasin Yilmaz. Real-time weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6848–6856, 2024. 2, 3

[17] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6536–6545, 2018. 3, 4, 5, 8

[18] Amir Markovitz, Gilad Sharir, Itamar Friedman, Lihi Zelnik-Manor, and Shai Avidan. Graph embedded pose clustering for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10539–10547, 2020. 2, 3, 4, 5, 6, 8

[19] Romero Morais, Vuong Le, Truyen Tran, Budhaditya Saha, Moussa Mansour, and Svetha Venkatesh. Learning regularity in skeleton trajectories for anomaly detection in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[20] Ghazal Alinezhad Noghre, Armin Danesh Pazho, and Hamed Tabkhi. An exploratory study on human-centric video anomaly detection through variational autoencoders and trajectory prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 995–1004, 2024. 2, 3, 4, 5, 6, 8

[21] Royston Rodrigues, Neha Bhargava, Rajbabu Velmurugan, and Subhasis Chaudhuri. Multi-timescale trajectory prediction for abnormal human activity detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2626–2634, 2020. 3

[22] Guodong Wang, Yunhong Wang, Jie Qin, Dongming Zhang, Xiuguo Bao, and Di Huang. Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles. In *European Conference on Computer Vision*, pages 494–511. Springer, 2022. 3

[23] Shoubin Yu, Zhongyin Zhao, Haoshu Fang, Andong Deng, Haisheng Su, Dongliang Wang, Weihao Gan, Cewu Lu, and Wei Wu. Regularity learning via explicit distribution modeling for skeletal video anomaly detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 3

[24] Yuan Yuan, Jianwu Fang, and Qi Wang. Online anomaly detection in crowd scenes via structure analysis. *IEEE transactions on cybernetics*, 45(3):548–561, 2014. 2

[25] M Zaigham Zaheer, Arif Mahmood, M Haris Khan, Mattia Segu, Fisher Yu, and Seung-Ik Lee. Generative cooperative learning for unsupervised video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14744–14754, 2022. 3

[26] Xianlin Zeng, Yalong Jiang, Wenrui Ding, Hongguang Li, Yafeng Hao, and Zifeng Qiu. A hierarchical spatio-temporal graph convolutional neural network for anomaly detection in videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(1):200–212, 2021. 3

[27] Yuxing Zhang, Jinchen Song, Yuehan Jiang, and Hongjun Li. Online video anomaly detection. *Sensors*, 23(17):7442, 2023. 2, 3