An Exploratory Study on Human-Centric Video Anomaly Detection through Variational Autoencoders and Trajectory Prediction

Ghazal Alinezhad Noghre Armin Danesh Pazho Hamed Tabkhi
University of North Carolina at Charlotte
Charlotte, NC, USA
{galinezh, adaneshp, htabkhiv}@uncc.edu

Abstract

Video Anomaly Detection (VAD) represents a challenging and prominent research task within computer vision. In recent years, Pose-based Video Anomaly Detection (PAD) has drawn considerable attention from the research community due to several inherent advantages over pixel-based approaches despite the occasional suboptimal performance. Specifically, PAD is characterized by reduced computational complexity, intrinsic privacy preservation, and the mitigation of concerns related to discrimination and bias against specific demographic groups. This paper introduces TSGAD, a novel human-centric Two-Stream Graph-Improved Anomaly Detection leveraging Variational Autoencoders (VAEs) and trajectory prediction. TSGAD aims to explore the possibility of utilizing VAEs as a new approach for pose-based human-centric VAD alongside the benefits of trajectory prediction. We demonstrate TSGAD's effectiveness through comprehensive experimentation on benchmark datasets. TSGAD demonstrates comparable results with state-of-the-art methods showcasing the potential of adopting variational autoencoders. This suggests a promising direction for future research endeavors. The code base for this work is available at https://github.com/TeCSAR-UNCC/TSGAD.

1. Introduction

In recent years, surveillance cameras have been proliferation; nevertheless, the available human resources are insufficient for real-time monitoring and expeditious, judicious response to the voluminous video feed generated by these cameras [30]. Furthermore, there may exist an inherent bias in decisions made by humans. Hence, as Artificial Intelligence (AI) continues to advance, the integration of smart technologies for the detection of anomalous behaviors has garnered significant attention across various communities.

Anomaly detection can refer to a wide range of applications [5, 15, 17, 31, 34]. One of the main subsets is the domain of human-centric video anomaly detection that has been examined from two primary perspectives: pose-based video anomaly detection [27, 39] and pixel-based video anomaly detection [32, 37]. While pixel-based approaches typically demonstrate superior detection accuracy, Pose-based Anomaly Detection (PAD) has attracted considerable research interest due to reduced computational complexity, inherent privacy preservation, and robustness to visual variations and background noise [3, 10, 25]. However, PAD methods may suffer from limited information and reliance on accurate pose estimation. The choice between these approaches should consider the specific application's requirements, balancing the need for privacy, efficiency, and the nature of targeted anomalies. In this article, our focus is pose-based approaches.

Video anomaly detection presents an intrinsic challenge as it inherently constitutes an open-set problem characterized by the potential emergence of diverse normal and abnormal behaviors within the real-world setup, driven by the complexity of human behavior. The supervised training, often reliant on data that may inadequately represent the entirety of anomalies, suffers from limited generalizability [29]. In response to this challenge, the field employs unsupervised learning techniques as the most common approach [16] to improve the efficacy of anomaly detection models. We embrace the unsupervised paradigm to mitigate this challenge in line with prior scholarly investigations.

This study introduces TSGAD, a novel approach that amalgamates reconstruction, distribution analysis, and trajectory prediction, offering promising avenues to enhance PAD performance. TSGAD departs from conventional methodologies by incorporating Variational Autoencoders (VAEs) [20]. Additionally, we propose the use of a State-of-the-Art (SotA) trajectory prediction model [1] as a complementary branch for PAD. The trajectory, derived from the human pose, is defined as the temporal evolution of the spatial displacement of the central keypoint (middle of the hips). Motivations for using trajectory are two-fold; firstly, in instances involving individuals positioned at a significant distance from the camera, trajectory data exhibits reduced noise levels compared to pose. Secondly, the chosen trajectory prediction model can capture social interactions missing from most pose-based methods through Graph Isomorphism.

We evaluate the proposed TSGAD method through comprehensive experiments. To ensure a thorough analysis, unlike most previous works, we employ not only the conventional Area Under the Receiver Operating Characteristic Curve (AUC-ROC) metric but also supplementary metrics, including Area Under the Precision-Recall Curve (AUC-PR) and Equal Error Rate (EER). These additional metrics provide a more nuanced understanding of our model's performance, addressing aspects that AUC-ROC may not fully represent. TSGAD attains AUC-ROC values of 80.67%, 81.77%, and 69.55% on three well-known anomaly detection benchmarks namely ShanghaiTech [24], Human Related ShanghaiTech [28], and CHAD [11] respectively. This exploratory study and its results showcase our method's ability to compete with SotA models in the field of anomaly detection, affirming its potential as a novel avenue for future endeavors and enhancement in PAD.

The contributions of this paper are as follows:

- Investigating the fusion of prediction-based and reconstruction-based approaches utilizing Variational Autoencoders (VAE) for human-related anomaly detection.
- Exploring the benefits of using social interaction-aware trajectory prediction for anomaly detection and propose an integrated approach that combines pose and trajectory methods for comprehensive anomaly detection.
- Conducting a thorough evaluation of the proposed solutions using a comprehensive range of metrics to gain deeper insights into the merits and limitations of the design under consideration.
- Empirical analysis of different pose anomaly score formulations to assess their impact on anomaly detection performance.

2. Related Works

2.1. Pixel-based Approaches

[32] proposes a multi-branch design for anomaly detection. The proposed method is based on the idea that anomalies can be detected using abrupt changes in velocity, pose, and deep features extracted from input frames. [37] tackles anomaly detection by solving a spatio-temporal jigsaw puzzle. The jigsaw solver is only trained using normal videos. The permutation predictions from the solver are used as a measure for anomaly detection. [6] introduced a two-stream framework for anomaly detection. The context recovery stream predicts the future video frames and the knowledge recovery stream compares the video snippet to the knowledge gathered from training on normal videos. [12] also proposes a multi-branch multi-task design. The first branch learns to predict the arrow of time, assuming that detecting the arrow of time is harder for anomalous behaviors. The second branch tries to detect irregular motion in the input sequence. The final branch reconstructs the detected objects bounding boxes. The model is trained on normal samples, and in the inference stage, the score from all these tree branches is fused to form the final score.

2.2. Pose-based Approaches

Normal Graph [26] leverages spatial-temporal graph convolutional network for predicting future pose segments trained on the normal data. The predicted pose segments are then compared to the ground truth, and the Mean Squared Error (MSE) loss is used as the anomaly score. Since the model is only trained on normal data, it cannot predict anomalous movements accurately. Hence the drastic difference between the prediction and actual future movement can reveal anomalous behavior. [33] uses a similar prediction-based approach. [33] not only predicts future pose sequences but also leverages a past prediction module for multi-scale past/future prediction to enhance the accuracy of the final anomaly detection model. [40] similarly proposes a prediction-based method but with three branches for predicting future pose, trajectory, and the motion vector. MPED-RNN [28] proposes a Gated Recurrent Unit-based (GRU) encoder-decoder structure with two decoder heads: a predicting and a reconstructing head. The reconstruction and prediction scores are then fused to generate the anomaly score. Similar to MPED-RNN, [22] uses both reconstruction and prediction for anomaly detection, but [22] uses Long Short-Term Memory (LSTM) units instead of GRUs. MemWGAN-GP [23] also uses a similar dual-head decoder structure upgraded with a modified version of the Wasserstein generative adversarial network [2] for improving the prediction and reconstruction quality. GEPC [27] uses a spatio-temporal graph autoencoder combined with a clustering layer for assigning soft probabilities to the input pose segments. The output probabilities are a measure of anomalous behavior.

3. Preliminaries

3.1. Variational Autoencoders

Variational Autoencoder (VAE) first introduced by [20] is a type of generative model initially designed for probabilistically generating content. They have gained popularity for their capability to model complex data distribution in an unsupervised manner. Similar to Autoencoders (AEs), VAEs can be used for feature learning, dimensionality reduction, denoising, etc. However, VAEs can be more advantageous for anomaly detection due to their probabilistic nature and ability to capture intricate data distributions. In VAEs, the objective function consists of two terms: a reconstruction term similar to AEs and a regularization term as depicted by Eq. (1):

$$L_{\text{VAE}} = L_{\text{reconstruction}} + L_{\text{regularization}} \tag{1}$$

L_{VAE} indicates Evidence Lower Bound (ELBO) which is maximized during the training. The reconstruction term is defined as the log-likelihood of the observed data given the latent variable:

$$L_{reconstruction} = \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}_q \left[\log p\left(x_n \mid z\right)\right] \tag{2}$$

where N is the number of data samples, x_n is the data sample and z is the latent variable. On the other hand, the regularization term encourages the latent space to have a specific structure, typically a multivariate Gaussian distribution. This objective is achieved using Kullback-Leibler (KL) divergence between the approximate posterior distribution $q(z \mid x)$ and the prior distribution $p(z)$:

$$L_{regularization} = -\frac{1}{N}\sum_{n=1}^{N} \text{KL}\left(q\left(z \mid x_n\right) \| p(z)\right) \tag{3}$$

Another variant of VAEs named β-VAE [4] was later introduced to learn a more disentangled representation in the latent space. This goal is achieved by highlighting the regularization term in Eq. (1) by adding a multiplier β with setting β values greater than 1. Later, TC-VAE [7] introduces a new formulation for achieving better disentanglement:

$$\begin{aligned} L_{TC-VAE} &:= \mathbb{E}_{q(z|n)p(n)}[\log p(n \mid z)] - \alpha I_q(z;n) \\ -\beta\, \text{KL}&\left(q(z) \| \prod_j q\left(z_j\right)\right) - \gamma \sum_j \text{KL}\left(q\left(z_j\right) \| p\left(z_j\right)\right) \end{aligned} \tag{4}$$

Where n is a distinct integer index assigned to every training datapoint establishing a random variable that uniformly covers the range from 1 to N. Keep in mind that $q(z,n) = q(z|n)p(n) = q(z|n)\frac{1}{N}$. The first term corresponds to the reconstruction loss. The second term is the Index-Code Mutual Information, which is the mutual information between the data variable and latent variable or $I_q(z;n)$. It is shown that maximizing this term results in learning more disentangled latent representations based on previous studies [4, 8]. Total correlation calculated by the third term in Eq. (4) quantifies the dependence between variables. Optimizing this term leads to learning independent factors in the data distribution. The final term denoted as dimension-wise KL, is responsible for maintaining congruence between the latent dimensions and their respective prior distributions. α, β, and γ are adjustable multipliers chosen by the requirements.

3.2. Trajectory Prediction

A trajectory prediction problem is forecasting a subject's future position based on the observed trajectory. Trajectory prediction has many real-world computer vision applications such as pedestrian safety, transportation safety, intelligent traffic monitoring, and video surveillance [13, 36]. Predicted trajectory can also be used for anomaly detection; sudden changes in the trajectory can be an indication of anomalous events.

We leverage Pishgu [1] for the purpose of anomaly detection. Pishgu uses the Graph Isomorphism Network (GIN) to capture the interdependencies between subjects available in the scene and constructs latent representations considering both social interactions and the movement history of the subject. In the next step, an attentive Convolutional Neural Network (CNN) is used for capturing temporal relations and constructing the final predicted trajectories.

4. TSGAD

In this section, we introduce our proposed TSGAD methodology.

4.1. Problem Formulation

As depicted in Fig. 1, TSGAD has two branches. The top branch uses a window of size T_{in} of observed poses as input:

$$\mathcal{P}^i = [P^i_{t_0}, P^i_{t_0+1}, ..., P^i_{t_0+T_{in}-1}] \tag{5}$$

where $P^i_{t_0}$ shows the pose of person i in frame t_0. A sequence of the position of the center of a person is used as the input of the trajectory branch:

$$C^i = [C^i_{t_0}, C^i_{t_0+1}, ..., C^i_{t_0+T'_{in}-1}] \tag{6}$$

where C_{t_0} is the location of the center of person i in frame t_0 and T'_{in} is the input window size for the trajectory prediction branch. Pose branch and trajectory branch output S^i_{Pose} and S^i_{Traj} respectively associated with the i^{th} person. The final output of the model is constructed by combining normalized S^i_{Pose} and S^i_{Traj}:

$$S^i_{Final} = a \cdot Norm(S^i_{Pose}) + b \cdot Norm(S^i_{Traj}) \tag{7}$$

where a and b are multipliers calculated by dividing the AUC-ROC of each branch by the sum of the AUC-ROC of both branches. In the final step, the maximum score over all

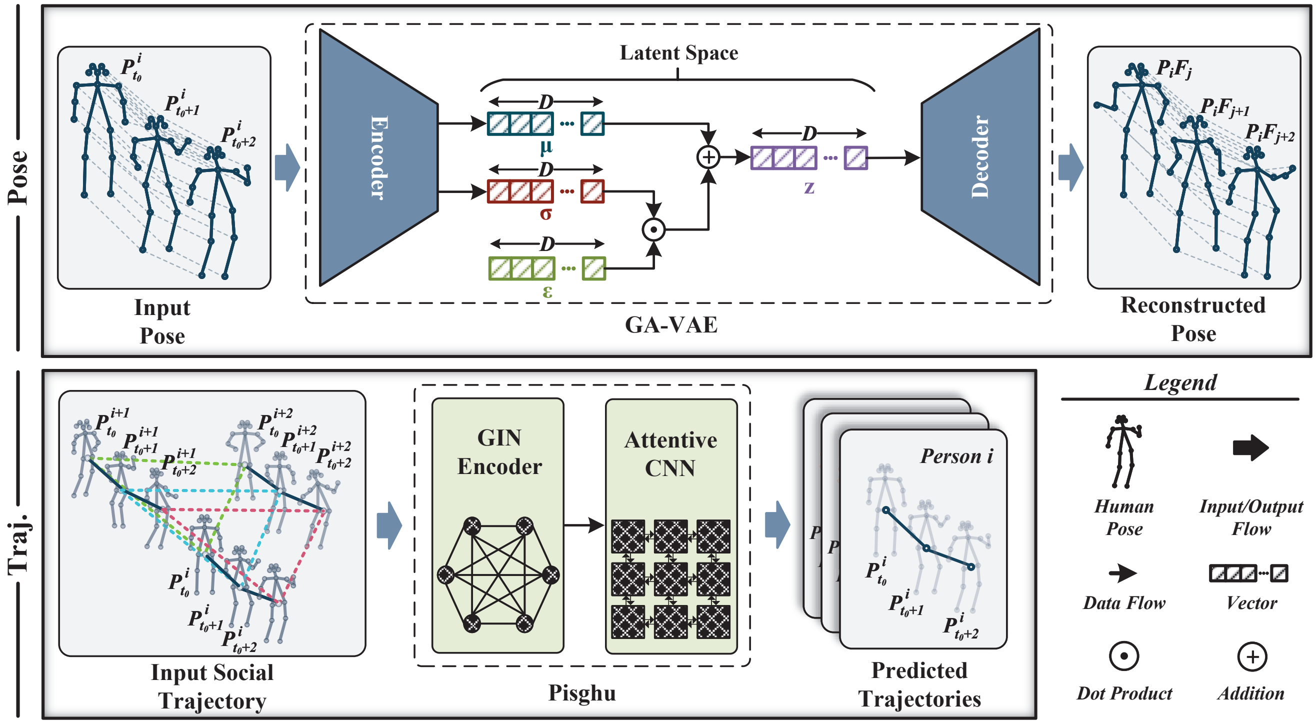

Figure 1. TSGAD architecture. The upper branch utilizes Graph Attentive Variational Autoencoder (GA-VAE) for learning the characteristics of normal human behavior distribution in an unsupervised manner. The lower branch leverages a SotA trajectory prediction method, namely Pishgu [1], for learning how to predict normal trajectories. P_t^i denotes the i^{th} person at time t, and D, μ, and σ refer to the latent representation's dimensions, mean, and variance. z follows a normal distribution with $z \sim (0, I)$, where I is the identity matrix.

available subjects in the scene is calculated and considered as the anomaly score associated with a frame:

$$S_{Final} = max_{i \in N}(S^{i}_{Final}) \tag{8}$$

4.2. Archietcture

As depicted in Fig. 1, the proposed model consists of two branches; the top branch uses pose data and Graph Attentive Variational Autoencoder (GA-VAE) to capture the distribution of normal behavior. The bottom branch uses the state-of-the-art trajectory prediction method Pishgu [1] to predict future trajectories. Deviation from predicted trajectories is used as a measure of anomaly detection.

4.2.1 GA-VAE

In order to capture the relationships between joints in human pose, we choose to represent the human pose using a spatio-temporal graph formulation. The joints are considered the nodes of the graph, and the edges represent the physical limbs and learned motion dependencies necessary for modeling the human pose effectively. In the context of video anomaly detection, it is imperative to incorporate temporal edges to represent the temporal interdependencies among frames. Thus, the resulting graph is a spatio-temporal graph that formulates human motion. We propose building a deep variational autoencoder leveraging Spatial-Temporal Graph Convolution (ST-GCN) blocks [38]. [27] extended ST-GCN blocks by adding more sophisticated spatial attention, including three GCN blocks for better capturing physical relations, dataset-level keypoint relations, and sample-specific relations. We chose a symmetric design for the VAE with both Graph Attentive Probabilistic encoder (GA-VAE encoder) and Graph Attentive Probabilistic decoder (GA-VAE decoder) having 9 layers of modified ST-GCN blocks, demonstrated in Fig. 2. Unlike previous works that use ST-GCN for processing input pose data, we adopt a probabilistic approach using ST-GCN blocks for constructing a VAE. We consider the prior distribution to be a normal distribution to match real-world human behavior. The probabilistic design is instrumental in capturing the inherent distribution of input data, thereby enhancing the modeling of normal behavior and consequently leading to improved performance in the context of anomaly detection.

The training procedure is conducted in an unsupervised fashion, wherein the GA-VAE is trained on the training set, which exclusively includes normal behavior exemplars. During the GA-VAE training phase, we implement the Ev-

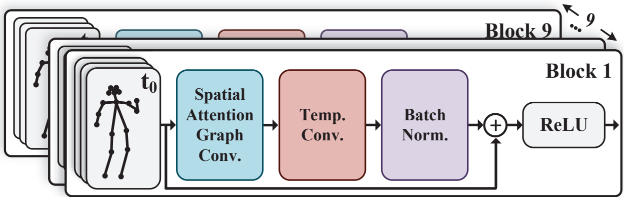

Figure 2. Nine layers of spatio-temporal graph convolution blocks are stacked forming the GA-VAE encoder. Each block consists of a spatial attention graph convolution followed by temporal convolution, batch normalization, a residual connection, and a final activation function.

idence Lower Bound (ELBO) loss as introduced in Eq. (4) with the same setup for multipliers as [7]. Throughout the training, the model minimizes the negative L_{GA-VAE} to encourage the understanding of normal behavior patterns. After the model is trained, for each datapoint in the training set, the corresponding normal distribution parameters (μ_n and σ_n) are concatenated and averaged over the training set to find an Aggregated Parameter Index (API) that is a single vector representing the characteristics of normal behavior:

$$API = \frac{1}{N}\sum_{n \in N}(\mu_n \,||\, \sigma_n) \tag{9}$$

where N is the number of datapoints in the training set.

Fig. 3 shows the inference process. Each datapoint in the test set is passed through the trained GA-VAE encoder to map to a normal distribution parameterized by μ_n and σ_n. To calculate the anomaly score, we measure the deviation from API:

$$S_{Pose} = \sqrt{\sum_{j=1}\left((\mu_n \,||\, \sigma_n)_j - API_j\right)^2} \tag{10}$$

where $(\mu_n \,||\, \sigma_n)_j$ is the j^{th} dimension of the latent parameters of input datapoint.

4.2.2 Trajectory Prediction for Anomaly Detection

As illustrated in Fig. 1, we advocate the incorporation of a trajectory prediction model within the context of anomaly detection. The primary goal is to introduce the dynamics of interactions between subjects, thereby extracting valuable insights for anomaly detection. The trajectory prediction branch, fundamentally concerned with modeling the collective movements of individuals within the scene, complements the pose-based anomaly detection approach. This approach is introduced to mitigate challenges associated with pose estimation inaccuracies. Consequently, the trajectory perspective provides a complimentary and holistic representation of the scene, contributing to improved anomaly detection capabilities.

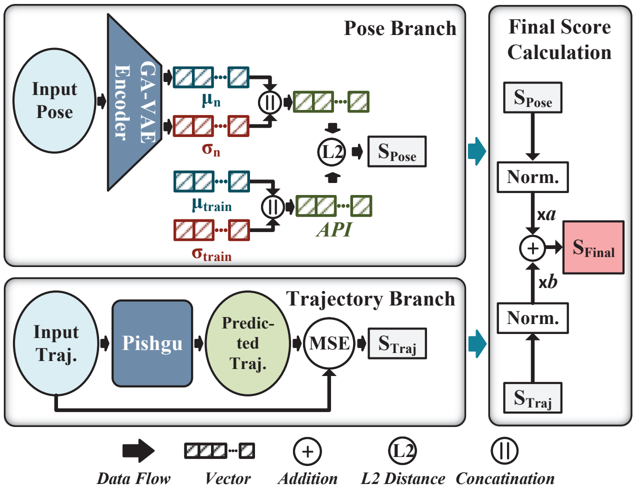

Figure 3. The inference phase. The deviation from API in the latent space is used for calculating the pose score (S_{Pose}). The difference between the predicted trajectory and the actual trajectory measured by MSE is used to form a trajectory score (S_{Traj}). The weighted sum of these normalized scores forms the final anomaly score. μ_n, σ_n, and API refer to the mean, and variance of the latent representation and Aggregated Parameter Index defined in Eq. (9) respectively.

We adopt the state-of-the-art trajectory prediction model Pishgu, as introduced by [1] for the specific application of anomaly detection. We train Pishgu exclusively on the training set, comprising instances of normal behavior, with a focus on capturing the nuanced features of typical movements. The training process involves the optimization of the Mean Squared Error (MSE) loss function.

$$\text{MSE} = \frac{1}{N}\sum_{n=1}^{N}\left(Y_n - \hat{Y}_n\right)^2 \tag{11}$$

where $\hat{Y}$ is the predicted trajectory and Y is the actual coordinates of a person.

In the inference phase, as illustrated by Fig. 3, the predicted trajectories for each datapoint are compared to the corresponding actual trajectories. The deviation is measured using MSE loss and used as an anomaly score S_{Traj}.

5. Experimental Setup

In this section, we focus on different aspects of our experimental setup. All the trainings and evaluations have been conducted on a computational workstation featuring three NVIDIA RTX A6000 GPUs and an AMD EPYC 7513 32-core CPU with 256 gigabytes of memory.

5.1. Datasets

5.1.1 ShanghaiTech Campus (SHT)

The ShanghaiTech Campus (SHT) dataset [24] serves as the principal benchmark for video anomaly detection. This dataset comprises more than 317,000 frames spanning 13 distinctive scenes. SHT is partitioned into an unsupervised subset, including in excess of 274,000 normal training frames and 42,883 normal and anomalous frames for testing purposes. For the purposes of our investigation, we adopt the unsupervised split to facilitate a meaningful comparison with prior research endeavors. PAD methodologies assess their models using this dataset due to its expansive scale and the presence of videos with sufficient quality recorded at 24 frames per second (FPS) for pose extraction. Additionally, the types of anomalies in this dataset are also a good representative of real-world scenarios. Accordingly, we integrate the SHT dataset into our experimental framework, employing a methodology consistent with previous SotA studies [39] for pose extraction and tracking [21], thus ensuring equitable comparisons.

5.1.2 HR-ShanghaiTech (HR-SHT)

The dataset introduced in [28] represents a specialized adaptation of the ShanghaiTech Campus (SHT) dataset, specifically tailored for human-related anomaly detection. It is essential to underscore that the sole differentiation between this dataset and the original SHT dataset lies in its exclusive concentration on anomalies related to human activities.

5.1.3 Charlotte Anomaly Dataset (CHAD)

Pazho et.al. [11] introduce the Charlotte Anomaly Dataset (CHAD), a high-resolution multi-camera dataset for video anomaly detection in real-world scenarios. It comprises approximately 1.15 million frames, encompassing 1.09 million normal frames and 59,000 anomalous frames with detailed annotations for human detection, tracking, and pose. It is suitable for both unsupervised and skeleton-based anomaly detection methods and emphasizes the use of multiple metrics for benchmarking, discussed in Sec. 5.3. CHAD simulates a real parking lot surveillance environment with four high-resolution cameras recording at 30 FPS, and a diverse set of actors engaging in normal and anomalous behaviors across 22 different anomaly classes. It stands out as the largest anomaly detection dataset with pose and tracking annotations. We exclusively employ the official unsupervised split of CHAD. We choose to conduct experiments on this dataset due to its role in establishing a standardized benchmark for pose-based anomaly detection, thereby mitigating variations due to the quality of extracted poses by different methods.

5.2. Training Setup

To optimize hyperparameters, a grid search methodology has been employed, systematically exploring parameter combinations. All trainings have been conducted 5 times to ensure stability. Adam optimizer is used in all training setups.

5.2.1 SHT and HR-SHT

For SHT and HR-SHT datasets, a batch size of 256 is employed for processing the data, while a dropout rate of 0.3 is applied to regularize the neural network. The model undergoes training for 20 epochs with an initial learning rate of 0.005 and a decay factor of 0.99. Also, a weight decay of 5.0e-05 to prevent overfitting. The input window size for set to 24 frames, allowing the model to capture meaningful patterns within the data over a span of 1 second.

For training the trajectory branch, we follow the setup of [1]; we used the input window size of 16 frames and the output prediction horizon of 14 frames and trained the model for 80 epochs with a learning rate of 0.02 and batch size of 64.

5.2.2 CHAD

For the CHAD dataset, a similar configuration is employed. A batch size of 256 is used, and a dropout rate of 0.3 is applied for regularization. The training also spans 20 epochs, with a slightly higher initial learning rate of 0.01, which still decays by a factor of 0.99. Similar to the first setup a weight decay of 5.0e-05 is employed for regularization. In this case, the input window size is set to 1 second or 30 frames.

When training the trajectory branch on CHAD, we use similar parameters as training on SHT with a window size set to 16 for input data and a prediction horizon of 14 for the output. The training process spanned 80 epochs, employing a learning rate of 0.02, a batch size of 64, and the Adam optimizer.

5.3. Metrics

5.3.1 AUC-ROC

The Area Under the Receiver Operating Characteristic (AUC-ROC) curve is used to evaluate the accuracy of binary classification models. The ROC curve represents the trade-off between a model's true positive rate (sensitivity) and its false positive rate (1-specificity) at various classification thresholds. A higher AUC-ROC score indicates a better model performance.

5.4. AUC-PR

The Area Under the Precision-Recall Curve (AUC-PR) quantifies the quality of binary classification, particularly in

Table 1. AUC-ROC compared on SHT [24], HR-SHT [28], and CHAD [11] datasets.

Methods	SHT	HR-SHT	CHAD
MPED-RNN [28]	73.40	75.40	-
GEPC [27]	75.50	-	64.90
PoseCVAE [19]	74.90	75.70	-
MSTA-GCN [9]	75.90	-	-
MTP [33]	76.03	77.04	-
HSTGCNN [40]	81.80	83.40	-
STGformer [18]	82.90	86.97	-
TSGAD-Pose (Ours)	80.59	81.52	59.30
TSGAD-Traj (Ours)	67.78	68.45	69.55
TSGAD (Ours)	80.67	81.77	66.49

imbalanced datasets. AUC-PR measures the area under the precision-recall curve, where precision represents the ratio of true positive predictions to the total positive predictions, and recall (sensitivity) quantifies the model's ability to capture all true positive instances. A higher AUC-PR value indicates a better-performing model, as it reflects a higher precision-recall trade-off, signifying superior discrimination and a more effective approach for problems where false positives can be costly or misleading.

5.5. EER

The Equal Error Rate (EER) quantifies the point at which the False Acceptance Rate (FAR) and False Rejection Rate (FRR) are equal for a binary classification model. EER identifies the threshold at which the decision boundary balances the rate of false positives vs. false negatives. Please note that EER alone is insufficient [35] but when combined with other metrics, offers valuable insights. A lower EER signifies improved system accuracy, as it indicates a better equilibrium between security and usability.

6. Results

6.1. Comparison with State-of-the-Art Approaches

The AUC-ROC metric stands as the most extensively investigated performance measure within the realm of PAD. Tab. 1 presents the investigation across diverse datasets revealing promising results for TSGAD, approaching SotA performance. This underscores the potential for further exploration of the innovative approach that integrates VAEs and probabilistic modeling, along with the fusion of pose and trajectories.

As illustrated in Table 1, it is evident that distinct branches exert varying influences, and the outcomes manifest their complementary nature. For instance, in the context of the SHT dataset [24], the pose and trajectory branches individually attain AUC-ROC values of 80.59% and 67.78%, respectively. However, the combination of these branches yields the most favorable results, implying that each branch addresses a complementary subset of anomalies, and their combined operation aims at mutual enhancement. The same trend can be seen for HR-SHT [28] as well.

Conversely, when examining the results for CHAD [11], the AUC-ROC values distinctly indicate that a trajectory-based approach is notably more well-suited for this particular environment. This observation may serve as an indicator of a potentially noisier environment or less precise pose annotations. As previously discussed, the inherent robustness of trajectory data, when contrasted with pose information, likely contributes to the superior results achieved with trajectory-based anomaly detection in this scenario. The incorporation of both pose and trajectory components appears to diminish the overall performance. Consequently, the selection of the most appropriate model depends on the intrinsic characteristics of the given environment. It is imperative to undertake comprehensive explorations involving diverse branches to tailor the model effectively to the specific requirements of the environment under consideration.

6.2. Detailed Analysis of Supplementary Metrics

While the AUC-ROC metric provides valuable insights into the effectiveness of binary classifiers, its applicability diminishes when confronted with imbalanced datasets [14]. Conversely, AUC-PR exhibits greater resilience in the presence of imbalanced data, thereby aiding in a deeper comprehension of the model's underlying characteristics. In addition to examining AUC-PR, we report EER to not only gain better insights into the sensitivity-specificity balance but also assess the real-world practicality of our model. AUC-PR and EER are only compared to GEPC [27] as it is the only model that its performance was reported for these metrics in [11].

As evident from the comprehensive metrics presented in Tab. 2, the behavior of the AUC-PR aligns with the trends observed in the AUC-ROC, as elucidated in Tab. 1, across various branches. Nevertheless, it is noteworthy that, across all scenarios, AUC-PR consistently registers values lower than AUC-ROC. This can be attributed to the inherent optimism of the AUC-ROC metric in imbalanced data, thereby complicating the translation of AUC-ROC results to real-world scenarios. The observed discrepancy signifies a misclassification event pertaining to the minority class (anomaly instances). Thus, a judicious calibration of the model to achieve an optimal trade-off between precision and recall is imperative, depending upon the specific requirements of the given use case.

In the field of anomaly detection, achieving a delicate balance between the False Acceptance Rate (FAR) and the False Rejection Rate (FRR) is critical, as it balances the im-

Table 2. AUC-PR and EER of our design compared to GEPC [27] on SHT [24], HR-SHT [28], and CHAD [11] datasets.

	SHT		HR-SHT		CHAD	
	AUC-PR ↑	**EER ↓**	**AUC-PR ↑**	**EER ↓**	**AUC-PR ↑**	**EER ↓**
GEPC [27]	65.70	0.31	-	-	58.70	0.38
TSGAD-Pose (Ours)	72.20	0.25	72.07	0.25	53.69	0.41
TSGAD-Traj (Ours)	61.26	0.38	61.32	0.38	66.97	0.36
TSGAD (Ours)	73.86	0.25	74.20	0.25	62.18	0.38

Table 3. Comparing different methods of calculating pose anomaly score (S_{Pose}) on SHT [24], HR-SHT [28], and CHAD [11] datasets.

	SHT			HR-SHT			CHAD		
	AUC-ROC ↑	**AUC-PR ↑**	**EER ↓**	**AUC-ROC ↑**	**AUC-PR ↑**	**EER ↓**	**AUC-ROC ↑**	**AUC-PR ↑**	**EER ↓**
GA-VAE	80.59	72.20	0.25	81.52	72.07	0.25	59.30	53.69	0.41
GA-VAE-ELBO	76.08	69.14	0.30	76.91	69.33	0.30	59.77	54.19	0.43

perative need for high sensitivity to detect anomalies and high specificity to minimize false alarms. The EER serves as the main metric to pinpoint this equilibrium. Notably, as depicted in Table 2, a substantial 34.2% reduction in the EER is observed in both SHT and HR-SHT datasets within the pose and combined models compared to the trajectory, signifying a more harmonious approach to anomaly detection. This trend is reversed in the context of CHAD, where the trajectory branch excels in achieving a balanced anomaly detection model, underscoring the contextual nuances inherent in different environments.

6.3. Ablation Study

In this section, we focus on the pose branch of the TSGAD model. As previously discussed, during the training of the GA-VAE, we employ the ELBO loss function. Maximizing ELBO forces the model to acquire more semantically significant representations within the latent space and to achieve a more precise approximation of the true posterior distribution. As a result, it ensures that the reconstructed data closely resembles the original data, preserving faithfulness in reconstruction.

In the context of unsupervised anomaly detection, when applied to the training dataset consisting of normal videos, a higher ELBO signifies stronger conformity to normal video patterns. After training, when we transition to the test dataset, which contains anomalous frames in addition to normal ones, a low ELBO serves as an indicator of deviations from the learned normal behavior, signifying abnormality. This intrinsic quality of the ELBO can be used as a standalone metric for detecting anomalies.

Therefore, as an alternative to the previously described approach outlined in Sec. 4, which involved the construction of a distribution of distributions, we have adopted the utilization of the Evidence Lower Bound (ELBO) inherent to the GA-VAE model as a singular measure. This approach allows us to quantitatively assess the advantages gained through the aforementioned distribution of distributions technique. Tab. 3 presents a summary of the performance enhancements accomplished through the utilization of the specified technique. In the context of the SHT and HR-SHT datasets, it is evident that GA-VAE outperforms GA-VAE-ELBO in a statistically significant manner. Conversely, in the case of the CHAD dataset, we observe a more subtle differentiation, with GA-VAE-ELBO exhibiting a slight advantage in terms of both AUC-ROC and the AUC-PR, albeit demonstrating inferior performance with respect to the EER.

7. Conclusion

Our investigation in this paper delved into the efficacy of variational autoencoders in combination with trajectory prediction for pose-based anomaly detection. Through a series of experiments conducted across multiple benchmark datasets, we have unveiled compelling evidence that this approach holds significant promise. The demonstrated effectiveness of this approach, with consistent performance on diverse datasets, indicates that it represents a worthwhile avenue for future exploration and development within the field of anomaly detection.

Acknowledgement

This research is supported by the National Science Foundation (NSF) under Award Number 2329816.

References

[1] Ghazal Alinezhad Noghre, Vinit Katariya, Armin Danesh Pazho, Christopher Neff, and Hamed Tabkhi. Pishgu: Universal path prediction network architecture for real-time cyber-physical edge systems. In *Proceedings of the ACM/IEEE 14th International Conference on Cyber-Physical Systems (with CPS-IoT Week 2023)*, pages 88–97, 2023. 1, 3, 4, 5, 6

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR, 06–11 Aug 2017. 2

[3] Francois Buet-Golfouse and Islam Utyagulov. Towards fair unsupervised learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1399–1409, 2022. 1

[4] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae. *arXiv preprint arXiv:1804.03599*, 2018. 3

[5] Francisco Caetano, Pedro Carvalho, and Jaime Cardoso. Deep anomaly detection for in-vehicle monitoring—an application-oriented review. *Applied Sciences*, 12(19):10011, 2022. 1

[6] Congqi Cao, Yue Lu, and Yanning Zhang. Context recovery and knowledge retrieval: A novel two-stream framework for video anomaly detection. *arXiv preprint arXiv:2209.02899*, 2022. 2

[7] Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31, 2018. 3, 5

[8] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29, 2016. 3

[9] 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. 7

[10] Mickael Cormier, Aris Clepe, Andreas Specker, and Jürgen Beyerer. Where are we with human pose estimation in real-world surveillance? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 591–601, 2022. 1

[11] 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. 2, 6, 7, 8

[12] 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. 2

[13] Harris Georgiou, Sophia Karagiorgou, Yannis Kontoulis, Nikos Pelekis, Petros Petrou, David Scarlatti, and Yannis Theodoridis. Moving objects analytics: Survey on future location & trajectory prediction methods. *arXiv preprint arXiv:1807.04639*, 2018. 3

[14] Haibo He and Yunqian Ma. *Imbalanced learning: foundations, algorithms, and applications*. Wiley-IEEE Press, 2013. 7

[15] Thi Kieu Khanh Ho and Narges Armanfard. Self-supervised learning for anomalous channel detection in eeg graphs: application to seizure analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7866–7874, 2023. 1

[16] Hadi Hojjati, Thi Kieu Khanh Ho, and Narges Armanfard. Self-supervised anomaly detection: A survey and outlook. *arXiv preprint arXiv:2205.05173*, 2022. 1

[17] Hadi Hojjati, Mohammadreza Sadeghi, and Narges Armanfard. Multivariate time-series anomaly detection with temporal self-supervision and graphs: Application to vehicle failure prediction. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 242–259. Springer, 2023. 1

[18] 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. 7

[19] 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. 7

[20] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1, 2

[21] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10863–10872, 2019. 6

[22] Nanjun Li, Faliang Chang, and Chunsheng Liu. Human-related anomalous event detection via spatial-temporal graph convolutional autoencoder with embedded long short-term memory network. *Neurocomputing*, 490:482–494, 2022. 2

[23] Nanjun Li, Faliang Chang, and Chunsheng Liu. Human-related anomalous event detection via memory-augmented wasserstein generative adversarial network with gradient penalty. *Pattern Recognition*, 138:109398, 2023. 2

[24] 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. 2, 6, 7, 8

[25] Ximeng Liu, Lehui Xie, Yaopeng Wang, Jian Zou, Jinbo Xiong, Zuobin Ying, and Athanasios V Vasilakos. Privacy

and security issues in deep learning: A survey. *IEEE Access*, 9:4566–4593, 2020. 1

[26] Weixin Luo, Wen Liu, and Shenghua Gao. Normal graph: Spatial temporal graph convolutional networks based prediction network for skeleton based video anomaly detection. *Neurocomputing*, 444:332–337, 2021. 2

[27] 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. 1, 2, 4, 7, 8

[28] 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*, pages 11996–12004, 2019. 2, 6, 7, 8

[29] Ghazal Alinezhad Noghre, Armin Danesh Pazho, Vinit Katariya, and Hamed Tabkhi. Understanding the challenges and opportunities of pose-based anomaly detection. *arXiv preprint arXiv:2303.05463*, 2023. 1

[30] Armin Danesh Pazho, Christopher Neff, Ghazal Alinezhad Noghre, Babak Rahimi Ardabili, Shanle Yao, Mohammadreza Baharani, and Hamed Tabkhi. Ancilia: Scalable intelligent video surveillance for the artificial intelligence of things. *IEEE Internet of Things Journal*, 2023. 1

[31] Armin Danesh Pazho, Ghazal Alinezhad Noghre, Arnab A Purkayastha, Jagannadh Vempati, Otto Martin, and Hamed Tabkhi. A survey of graph-based deep learning for anomaly detection in distributed systems. *IEEE Transactions on Knowledge and Data Engineering*, 2023. 1

[32] Tal Reiss and Yedid Hoshen. Attribute-based representations for accurate and interpretable video anomaly detection. *arXiv preprint arXiv:2212.00789*, 2022. 1, 2

[33] 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. 2, 7

[34] Oumaima Sliti, Maxime Devanne, Sophie Kohler, Naim Samet, Jonathan Weber, and Christophe Cudel. f-anogan for non-destructive testing in industrial anomaly detection. In *Sixteenth International Conference on Quality Control by Artificial Vision*, volume 12749, pages 297–304. SPIE, 2023. 1

[35] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488, 2018. 7

[36] Chujie Wang, Lin Ma, Rongpeng Li, Tariq S Durrani, and Honggang Zhang. Exploring trajectory prediction through machine learning methods. *IEEE Access*, 7:101441–101452, 2019. 3

[37] 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. 1, 2

[38] Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875*, 2017. 4

[39] 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. 1, 6

[40] 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*, 2021. 2, 7