

CHAD: Charlotte Anomaly Dataset

Armin Danesh Pazho, Ghazal Alinezhad Noghre, Babak Rahimi Ardabili,
Christopher Neff, and Hamed Tabkhi

University of North Carolina at Charlotte, Charlotte NC 28223, USA
adaneshp@uncc.edu, galinezh@uncc.edu, brahimia@uncc.edu, cneff1@uncc.edu,
and htabkhiv@uncc.edu

Abstract. In recent years, we have seen a significant interest in data-driven deep learning approaches for video anomaly detection, where an algorithm must determine if specific frames of a video contain abnormal behaviors. However, video anomaly detection is particularly context-specific, and the availability of representative datasets heavily limits real-world accuracy. Additionally, the metrics currently reported by most state-of-the-art methods often do not reflect how well the model will perform in real-world scenarios. In this article, we present the Charlotte Anomaly Dataset (CHAD). CHAD is a high-resolution, multi-camera anomaly dataset in a commercial parking lot setting. In addition to frame-level anomaly labels, CHAD is the first anomaly dataset to include bounding box, identity, and pose annotations for each actor. This is especially beneficial for skeleton-based anomaly detection, which is useful for its lower computational demand in real-world settings. CHAD is also the first anomaly dataset to contain multiple views of the same scene. With four camera views and over 1.15 million frames, CHAD is the largest fully annotated anomaly detection dataset including person annotations, collected from continuous video streams from stationary cameras for smart video surveillance applications. To demonstrate the efficacy of CHAD for training and evaluation, we benchmark two state-of-the-art skeleton-based anomaly detection algorithms on CHAD and provide comprehensive analysis, including both quantitative results and qualitative examination. The dataset is available at <https://github.com/TeCSAR-UNCC/CHAD>.

Keywords: Anomaly Detection · Dataset · Computer Vision · Deep Learning.

1 Introduction

Video anomaly detection, which requires understanding if a video contains anomalous behaviors, is a popular but challenging task in computer vision. In addition to substantial research interest, many real-world applications greatly benefit from being able to determine if such anomalous behaviors are present. Parking lot surveillance is one such application, where being able to determine the presence of an anomalous action (e.g. fighting, theft, fainting) is paramount.

Current state-of-the-art (SotA) deep learning solutions take one of two approaches. The first is an appearance-based method, where the algorithm works directly on video frames. The second is the skeleton-based methodology, in which algorithms rely on extracted human pose data to understand human behaviors. Both methods require large amounts of quality data. Anomaly detection is particularly context-specific, so training data must also be representative of both the environment and the context of the target application. This need is amplified for unsupervised approaches, which try to learn the normal behaviors of a specific context and need many example frames to do so.

There are currently only a limited number of datasets for video anomaly detection. These datasets, while seeing continual growth in the amount of data provided, also tend to fall short regarding the number of normal frames per context (i.e., per scene). Additionally, no current video anomaly dataset provides the detection, tracking, and pose information required by skeleton-based methods, leaving them to rely on external algorithms to generate this data. Since there is no standard for this, it is difficult to determine how much of an approach’s error is due to the noise in this generated data or from the algorithm itself. This is further obfuscated by the inconsistency of the metrics used in reporting performance. Of the three main metrics for anomaly detection, discussed in Section 6, most SotA approaches only report one. However, all of them are necessary for a full understanding of an algorithm’s performance, especially in the real-world.

In this paper, we present the Charlotte Anomaly Dataset (CHAD), a high-resolution, multi-camera anomaly detection dataset in a parking lot setting. CHAD is designed to address the most challenging issues facing current video anomaly detection datasets. The first video anomaly dataset with multiple camera views of a single scene, CHAD has over 1.15 million frames capturing the same context. With over 1 million normal frames, CHAD places itself as the premiere video anomaly dataset for unsupervised methods, providing human detection, tracking, and pose annotations. Thanks to these annotations, CHAD allows for a more accurate standard, positioning itself as the best-in-class dataset for skeleton-based anomaly detection.

We also propose a new standard in the benchmarking and evaluation of real-world video anomaly detection. Included is a detailed discussion on metrics, the benefits and disadvantages of each, and how the use of all three is needed to truly understand an algorithm’s performance. To demonstrate the efficacy of CHAD, we train two SotA skeleton-based approaches, report both single camera and multi-camera performance, and compare to those methods trained on other datasets. Additionally, we perform cross-validation on CHAD, and the ShanghaiTech Campus Dataset [17], demonstrating CHAD’s suitability for enabling generalization and revealing it to be more challenging than its peers.

In summary, this paper has the following contributions:

- We introduce CHAD, a high resolution, multi-camera video anomaly detection dataset in a parking lot setting. With over 1.15 million frames of a single context and detection, tracking, and pose annotations, CHAD positions it-

self as the best-in-class dataset for both unsupervised and skeleton-based anomaly detection methods.

- We propose a new standard in real-world video anomaly detection benchmarking and evaluation. We provide a detailed discussion on the metrics used, including the insights they provide.
- To validate the efficacy of CHAD, we train and evaluate two SotA skeleton-based models with our proposed methodology. We further perform cross-validation on CHAD and ShanghaiTech Campus [17], demonstrating that CHAD is robust enough for generalization while being more challenging.

2 Related Work

Anomaly Detection Algorithms Appearance-based methods utilize appearance and motion features generated directly from pixel data for detecting anomalies [6, 11, 12, 18, 27, 30, 32, 35]. These methods generally achieve high accuracies in their context at the cost of high computation. Skeleton-based methods utilizes high-level, low-dimensional human pose skeletons [14, 20, 23, 24, 28]. These skeletons are informative in the context of human behavior while requiring far less computation than working with raw video data. They are more privacy preserving, and they remove demographic biases. As such, researchers have found significant success in skeleton-based anomaly detection.

Anomaly Detection Datasets The CUHK Avenue Dataset [19] consists of nearly 31K frames captured from a single camera. Abnormal objects, walking in the wrong direction, and sudden movements are examples of anomalous behaviors in this dataset.

The UCSD Anomaly Detection Dataset [22] consists of 19K frames overlooking pedestrian walkways. UCSD has been categorized into two subsets, each one covering a different view. UCSD Ped1 sees pedestrian movement perpendicular to the camera, while UCSD Ped2 sees movement parallel to the camera. UCSD contains positional information for localizing anomalies.

The Subway dataset [2] consists of two surveillance videos, the subway entrance, and exit. With a combined total of 139 minutes of video, this dataset counts behaviors such as running, loitering, and walking in the opposite direction of the crowd as anomalous behaviors.

Street Scene [26] is a single scene anomaly detection dataset captured from a bird’s eye view of a two lane street. Compared to most other datasets, Street Scene is relatively large at over 200K frames. Street Scene also contains non-human anomalies, such as illegally parked cars, dogs on the sidewalk, and cars making u-turns.

The ShanghaiTech Campus dataset [17] contains 13 different scenes taken from a campus setting. With over 317K frames, ShanghaiTech is one of the largest and most popular anomaly detection datasets available. However, ShanghaiTech has relatively few frames per context.

IITB-Corridor [28] was the largest single-stationary-camera anomaly detection dataset that existed before CHAD. It contains nearly 440K frames in a campus setting. Recorded in high-resolution 1080p, it is the only continuous video anomaly detection dataset with a resolution comparable to CHAD.

The ADOC dataset [25] is captured from a single high-resolution camera over 24 hours in a campus setting. ADOC consists of 260K frames and adopts an approach of considering any low-frequency behavior to be anomalous. Assuming only walking is normal, they consider all other behaviors as anomalous, even relatively commonplace activities like walking with a briefcase, having a conversation, or a bird flying through the air. While this categorization works for ADOC’s context, it is inconsistent with how other datasets define anomalous behaviors.

Specifically for supervised anomaly detection, UBnormal [1] is comprised entirely of synthetically generated videos. With a total of 236,902 frames, UBnormal is moderately large compared to other anomaly datasets, though with 29 scenes the average number of frames per scene is fairly low.

The NOLA dataset [9] is another new dataset. Collected over an entire week, NOLA contains over 1.4 million frames including both day and night scenes. In contrast to most other anomaly datasets, NOLA uses a single moving camera instead of stationary cameras. The rapid movement of the camera introduces a massive change of context, making the video anomaly detection more challenging. Due to the way annotations are presented in the dataset and the lack of clarifying documentation, it is impossible to ascertain what constitutes an anomaly in the context of NOLA. As such, it is difficult to determine the efficacy of this dataset for anomaly detection, and fair comparison to other datasets is not feasible.

UCF Crime [30] and X-D Violence [34] collect video clips from many different sources in varying contexts, as opposed to continuous recordings. This allows them to be enormous by anomaly dataset standards but is so fundamentally different in problem formulation that it could be considered a different task altogether. XD-Violence provides both video and audio, making it unique among video anomaly datasets.

All of these datasets bring their own benefits and have helped advance the field of video anomaly detection. However, while they all have their own strengths, each of them also provides its own challenges when it comes to training networks for the real-world. Some datasets are too small, either in overall frames or frames per scene. Some of them have strict definitions of normal behaviors that would be undesirable in a real-world context. Some have to contend with domain shift, either from taking a large amalgamation of clips from entirely different contexts or from training with synthetic actors and moving to real persons and objects when used in a real-world context. And while many of these datasets provide multiple contexts, none of them provide different views of the same context, as would be fairly common in a surveillance setting. Further, none of these datasets provide the human detection, tracking, and pose annotations

needed for skeleton-based anomaly detection. It is impossible for a single dataset to fit every possible scenario.

3 Data Collection and Setup

Since anomaly detection is such a context-specific task, it is important that the data used to train algorithms is representative of their real-world environments. Often the disconnect between training data, and inference data leads to unsatisfying performance in the real-world [3]. CHAD was designed to accurately mimic a real-world parking lot surveillance setting. The four cameras, as seen in Fig. 1, were positioned to cover the same general scene, though their perspectives give them each a unique context compared to the others. Each video is recorded in full HD (1920x1080, 30fps), except camera 4 which is in standard HD (1280x720, 30fps), as seen in Table 3.

There are thirteen actors present in CHAD. The actors represent diverse demographics (gender, age, ethnicity, etc.) and each participates in both normal and anomalous clips. There are 22 classes of anomalous behaviors in CHAD, which can be seen in Table 1. This list has been curated in line with other state-of-the-art datasets [2, 17, 19, 22]. All other actions present in CHAD (e.g. walking, waving, talking, etc.) are considered normal.

Table 1: Anomalous behaviors present in CHAD.

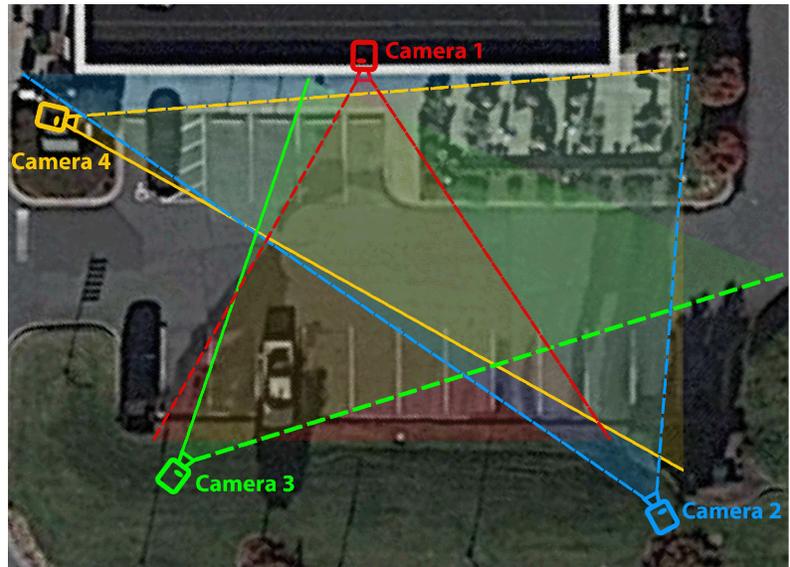
Type of Anomalous Behavior			
Group Activities		Individual Activities	
Fighting	Punching	Throwing	Running
Kicking	Pushing	Riding	Falling
Pulling	Slapping	Littering	Jumping
Strangling	Body Hitting	Hopping	Sleeping
Theft	Pick-Pocketing		
Tripping	Playing with Ball		
Chasing	Playing with Racket		

4 Annotation Methodology

CHAD contains four types of annotations: frame-level anomaly labels, person bounding boxes, person ID labels, and human keypoints.

4.1 Anomaly Annotations

We annotate anomalous behaviors at the frame level. This is, we mark the frame where the anomalous behavior begins, the frame where it ends, and every frame in between. This is done by hand, accounting for all the behaviors defined in Section 3. These frame-level labels are needed for both appearance-based and skeleton-based approaches. CHAD does not include anomaly localization labels.



(a) Birds-eye View



(b) Camera View

Fig. 1: Approximate position and the views of the cameras.

4.2 Person Annotations

One of the innovations that sets CHAD above its peers is the inclusion of person annotations. In real-world scenarios, there is no access to hand-annotated data. The annotations must be generated through available tools and are not always perfect. We include generated person-annotations to ensure they are more representative of a real-world situation. This is by design, as a certain amount of noise is desirable in the dataset to assist models in learning how to deal with unclean data inherent in real-world situations [5]. It also allows skeleton-based anomaly detection methods to have access to the processed data they need without having to spend time extracting it themselves. We hope this will make skeleton-based anomaly detection more accessible to researchers, leading to more innovation. It also sets a standard previously unavailable for how to generate this human detection, tracking, and pose information. With this standard, the variability based on the quality of input data is removed, leading to more precise and fair comparisons between approaches.

Bounding Boxes The bounding box of a person refers to the upper and lower x and y coordinate limits they occupy in an image. Having quality bounding boxes for each individual and for every frame is doubly important for CHAD, as this localization is needed for the extraction of both person ID labels and human keypoints as well. For this reason, CHAD utilizes the popular object detection algorithm YOLOv4 [4] for generating quality bounding boxes. Since CHAD is focused on anomalous human behavior, only the bounding boxes for people are used.

Person ID Labels Anomaly detection algorithms often utilize temporal information to understand the behaviors of people. Particularly for skeleton-based methods, it is necessary to be able to associate the different poses of a person to that specific person across frames. Person ID labels provide this information, allowing for temporal tracking of individual persons in each video clip. Given the bounding box information generated previously, DeepSORT [33] was utilized to provide tracking for persons through frames, generating unique person ID labels for each person in a video clip. For label stability, a three frame warm-up is used by DeepSORT before providing person ID labels. As such, the first two frames of each video clip are absent of person annotations.

Human Keypoints CHAD contains pose information in the form of human pose skeletons. These skeletons are made up of human keypoints, or points of interest on the human body. While there are several methods for defining what keypoints to use, CHAD follows the 17 keypoint methodology proposed by MS COCO [16]. Using the localization provided by the previously generated bounding boxes, keypoints are extracted using HRNet [31], a prolific algorithm for human pose estimation used by many. To ensure we only provide quality keypoint annotations, we remove any person with low confidence (<50%) for at

least half of their keypoints (9+). While this leads to some frames where people are not detected, it helps reduce the overall noise of the data that is present.

4.3 Annotation Smoothing

The algorithms used to annotate CHAD are imperfect, and there are instances where people are completely missed at either the object detection or keypoint extraction stage. Combined with our purposeful removal of overly noisy data, this results in an undesirable number of missed persons. To compensate for this, we introduce annotation smoothing to CHAD, using high confidence annotations to help fill in the missing information.

Given the relatively high frame rate of CHAD at 30 frames per second, it is a reasonable assumption that the positions and skeletons of a person will not drastically change between consecutive frames. As such, we can use linear interpolation to approximate the bounding box coordinates of each individual, assuming we have accurate detection at the start and end of the missing frames, and the number of missing frames is not too large. We choose 15 frames, or half a second, as a qualitative analysis showed this to be long enough to provide a significant benefit to annotation consistency, but not so long that the data it produced became unreliable. We apply the same smoothing technique to the keypoint annotations, with the same frame limitations. The details of smoothing are provided in the following equation:

$$\mathbb{X}_i = \left(\frac{\mathbb{X}_{\mathcal{N}} - \mathbb{X}_{\mathcal{M}}}{\mathcal{N} - \mathcal{M}} \right) \times i + \mathbb{X}_{\mathcal{M}} \quad (1)$$

where \mathbb{X}_i refers to a missing point (either bounding box or keypoint coordinate) at frame i , $\mathbb{X}_{\mathcal{M}}$ and $\mathbb{X}_{\mathcal{N}}$ refer to the two nearest matching points at frames \mathcal{M} and \mathcal{N} respectively, and where $\mathcal{M} < i < \mathcal{N}$ and $\mathcal{N} - \mathcal{M} + 1 \leq 15$.

The added consistency in annotations created by this smoothing is particularly useful in the context of unsupervised learning. However, the confidence scores of keypoints generated by this smoothing are set to Null, so they can be easily discarded if undesired.

5 CHAD Statistics

With over 1.15 million frames, CHAD is the largest anomaly detection dataset available that is recorded from continuous videos captured from stationary cameras, and includes person annotations. As shown in Table 4, CHAD has more than $2\times$ the number of frames as the next largest dataset, providing a substantial amount of learnable data. Additionally, CHAD has over 1 million frames of purely normal behaviors, which are required for unsupervised methods that rely on learning the normal to understand the anomalous. This is nearly $3\times$ more than can be found in other datasets. The 59K anomalous frames in CHAD are comprised of the 22 anomalous behaviors presented in Table 1. To facilitate supervised, unsupervised, and semi-supervised approaches, CHAD includes two

Table 2: Annotation availability in ShanghaiTech [17], CUHK [19], UCSD [22], Subway [2], IITB [28], Street Scene [26], UBnormal [1], and CHAD (Ours). * partially annotated, – not annotated.

Dataset	Anomaly Annotations		Person Annotations		
	Frame-level	Pixel-level	Bounding Box	ID Number	Keypoints
ShanghaiTech	✓	✓	–	–	–
CUHK	✓	✓	–	–	–
UCSD	✓	*	–	–	–
Subway	✓	✓	–	–	–
IITB	✓	✓	–	–	–
Street Scene	✓	✓	–	–	–
UBnormal	✓	✓	✓	–	–
CHAD (Ours)	✓	–	✓	✓	✓

splits for training and testing. The *unsupervised split* has a training set comprised only of normal behaviors, while the test set contains both normal and anomalous behaviors. The details of the *unsupervised split* can be found in Table 4. For the *supervised split*, the normal and anomalous frames were distributed uniformly between the training and test sets, with 60% of each belonging to the training set and 40% to the test set.

More than just the amount of data, CHAD benefits from having high quality image data. As discussed in Section 3, CHAD was recorded from four high-resolution cameras with an overlapping view of a scene. Recorded at 30 FPS, CHAD not only boasts a higher resolution and frame rate than other datasets, shown in Table 3, but also presents data in a format representative of modern real-world surveillance systems. While resolution and frame rate are indicators of overall video quality and the amount of data present in each frame, they can not convey how much of that data is actually useful for learning. Difference of Gaussian [7] is an image processing method that has been used to simulate how the human eye extracts visual details of an image for neural processing [21]. More simply, it creates a visual illustration of the density and richness of the features in an image. This allows us to visually analyze the quality of the data present in each dataset by comparing the Difference of Gaussian between them.

We visualize the Difference of Gaussian for a single frame of each dataset in Fig. 2. We set a Gaussian blur radius of one pixel to maximize the precision of the resulting representation. Looking at the images, CHAD very clearly presents the most detail. This was anticipated due to its high resolution, but the amount by which it surpasses the other datasets far exceeded expectations. Fine details in the persons, clothing, vehicles, and the environment are clear, granting an accurate perception of the original image. IITB-Corridor [28] is the only other dataset with 1080p images. However, the Difference of Gaussian tells a different story. While there are details present in the environment, they are comparably indistinct. Even in the brightened image, it is difficult to tell there is a person

in the image. This demonstrates a surprising lack of rich features in the IITB-Corridor, despite the resolution.

Table 3: Resolution and frame rate in Shanghai [17], CUHK [19], UCSD [22], Subway [2], IITB [28], Street Scene [26], UBnormal [1], and CHAD (Ours). N/A means Not Available.

Dataset		Resolution (Pixels)	Frame Rate (FPS)
Shanghai		856*480	N/A
CUHK		640*360	25
UCSD	Ped1	238*158	N/A
	Ped2	360*240	N/A
Subway		N/A	N/A
IITB		1920*1080	25
Street Scene		1280*720	15
UBnormal		varies	30
CHAD (Ours)	Scene 1-3	1920*1080	30
	Scene 4	1280*720	30

Street View, at the next highest resolution, shows much more detail and clarity than IITB-Corridor, though nowhere near the level of CHAD. What is most interesting is that while the building, car, and street boundaries are clear, it is difficult to notice the two people in the bottom left of the image. This is perhaps due to their relative size compared to the other objects mentioned and not necessarily indicative of a lack of features. Unsurprisingly, the lower resolution datasets, UCSD and CUHK Avenue, show sharp focal points (bright white pixels) but very little overall detail. Interestingly for ShanghaiTech [17], despite its slightly higher resolution, it presents a similar level of detail as Street Scene. However, due to the different camera perspectives, this translates into Shanghai providing better features for people, which is beneficial for its context.

Overall, we can see that CHAD not only has the best-in-class resolution and frame rate among anomaly detection datasets but also that the videos in CHAD are extremely feature rich, unrivaled among its peers. Additionally, there is a significant amount of background information irrelevant to person behaviors. The brightest spot in the Difference of Gaussian for CHAD is the foliage in the bottom left. This is noise - a distractor from information pertinent to anomaly detection. This means CHAD is not only more informative than other datasets but also suggests that it is more challenging as well. This level of challenge is needed if algorithms are to perform well in real-world scenarios, which are notorious for being more demanding than dataset benchmarks.

6 Metrics and Measurements

There are three main metrics used for evaluating performance on anomaly detection datasets: Area Under the Receiver Operating Characteristic Curve, Area

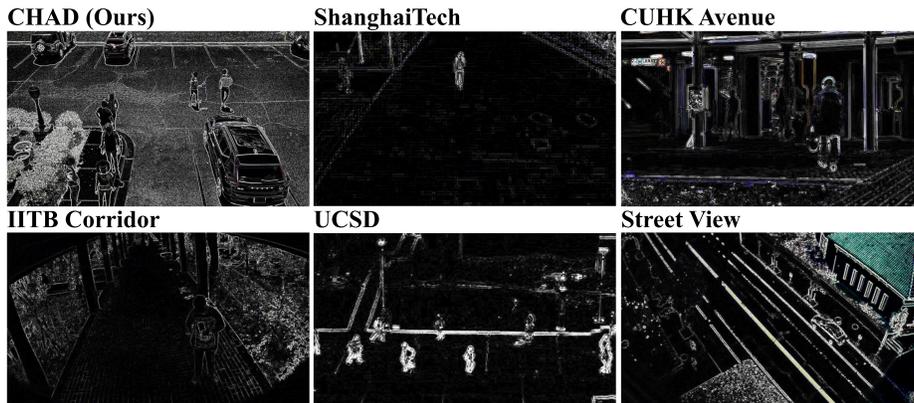


Fig. 2: Visualization of Difference of Gaussian in Shanghai [17], CUHK [19], UCSD [22], IITB [28], Street Scene [26], and CHAD (Ours). UCSD cropped to fit. All brightened for readability.

Table 4: Dataset comparison for ShanghaiTech [17], CUHK [19], UCSD [22], Subway [2], IITB [28], Street Scene [26], UBnormal [1], and CHAD (Ours). CHAD uses unsupervised split. N/A means Not Available.

Dataset	Number of Frames					Scene(s)	Camera(s)
	Total	Train	Test	Normal	Anomalous		
ShanghaiTech	317,398	274,515	42,883	300,308	17,090	13	13
CUHK	30,652	15,328	15,324	26,832	3,820	1	1
UCSD	18,560	9,050	9,210	12,919	5,641	2	2
Subway	208,925	27,500	181,425	205,805	3120	2	2
IITB	483,566	301,999	181,567	375,288	108,278	1	1
Street Scene	203,257	56,847	146,410	N/A	N/A	1	1
UBnormal	236,902	116,087	28,175	147,887	89,015	29	-
CHAD (Ours)	1,152,649	1,026,174	126,475	1,093,477	59,172	1	4

Under the Precision-Recall Curve, and Equal Error Rate. While none of these metrics are truly representative of overall performance, they each have their strengths and weaknesses, and, taken together, they can provide a comprehensive understanding of how an algorithm truly performs.

6.1 Receiver Operating Characteristic Curve

The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is simply the area under the curve when plotting the True Positive Rate (TRP) over the False Positive Rate (FPR) over various thresholds. This metric is specific to binary classification, such as determining if a video does or does not contain anomalous behavior. Generally, a higher AUC-ROC indicates that the model is better at separating inputs into their corresponding classes. The ROC curve itself also helps give insight into the trade-off between TPR and FPR at different thresholds [10]. However, AUC-ROC is not indicative of the final decisions of a model. The metric reports a final calculated number, and conclud-

ing useful information about the actual amount of False Negative Rate (FNR), when an anomaly is classified as normal is almost unfeasible. FNR is particularly important for real-world applications, and reporting it separately is crucial. Additionally, AUC-ROC is very sensitive to imbalances in data [13], making it sub-optimal if one class is over represented, as is often the case with normal behaviors in anomaly datasets [8].

6.2 Precision-Recall Curve

Precision is the fraction of correct positive guesses over all positive guesses, while Recall is the fraction of correct positive guesses over all positive samples. The Precision-Recall Curve (PR) is useful for understanding how to balance Precision and Recall, while the area under this curve summarizes all the information represented in it. While AUC-PR heavily focuses on the positive class, it still accounts for the False Negative Rate (FNR) – that is when the model classifies an anomaly as normal. As such, AUC-PR is a better metric for understanding the prediction ability of a model when compared to AUC-ROC [29]. Additionally, AUC-PR is better suited for highly imbalanced data [29], making it better at evaluating the minority class [13]. As the minority class in anomaly detection usually refers to anomalous behaviors, this is an important quality for this context. However, AUC-PR is a final calculated number and it does not provide direct insight into the correct classification of negative samples, nor does it provide a measure for the number of incorrect decisions a model makes. Thus, much like AUC-ROC, AUC-PR provides an incomplete understanding of a model’s performance.

6.3 Equal Error Rate

Another useful metric is the Equal Error Rate (EER) [15]. Plotting the FNR and FPR over various thresholds produces two curves that intersect at one point. The value at the intersection is the EER and shows what threshold value allows the model to achieve a balance between FNR and FPR. In the context of video anomaly detection, the EER illustrates how many false alarms a model will raise and how many anomalous frames it will miss when at equilibrium. On its own, this metric offers little insight into the overall performance of a model [30]. However, when used as a complement to AUC-ROC and AUC-PR, a more complete understanding can be achieved.

7 Evaluation

All experiments were conducted on a server containing two Intel Xeon Silver 4114, one V100 GPU, and 256 GB of RAM. We performed each experiment (training and testing) five times, averaging the results to remove any potential skew due to variability. For each model, training is performed exactly as described in their respective papers unless otherwise specified.

7.1 Standard Validation

To demonstrate CHAD’s viability as an anomaly detection dataset, we train and evaluate two state-of-the-art skeleton-based models using the *unsupervised split*. We select Graph Embedded Pose Clustering (GEPC) [23] and Message-Passing Encoder-Decoder Recurrent Neural Network (MPED-RNN) [24] for their high accuracy and model availability. GEPC utilizes a spatio-temporal graph autoencoder, while MPED-RNN uses a two-headed structure with reconstruction and prediction.

Both models were trained on each of CHAD’s four camera views individually, the results reported in Table 5. The most obvious observation is that both models were able to learn on CHAD. GEPC achieved an average AUC-ROC of 0.663 and AUC-PR of 0.619, while MPED-RNN achieved an average AUC-ROC of 0.718 and AUC-PR of 0.635. For both models, the AUC-ROC is noticeably higher than the AUC-PR. This is largely due to the overwhelming majority of normal frames in the data, which if properly classified will a significant boost to the AUC-ROC. AUC-PR, on the other hand, does not count True Negatives, and as such gives a more measured result for the imbalanced data. Additionally, GEPC achieved an EER of 0.378 and MPED-RNN an EER of 0.339. This means that, given the threshold at equilibrium, both models can expect to see between 34% and 38% of both normal frames and anomalous frames to be misclassified. This is important to understand when targeting real-world applications, where misclassification rates are more important than class separability.

Table 5: Evaluation of GEPC [23] and MPED-RNN [24] on CHAD (Ours).

Model	Camera	AUC-ROC	AUC-PR	EER
GEPC	1	0.673	0.636	0.363
	2	0.660	0.566	0.382
	3	0.661	0.586	0.384
	4	0.656	0.688	0.382
MPED-RNN	1	0.747	0.715	0.303
	2	0.691	0.567	0.349
	3	0.771	0.584	0.331
	4	0.662	0.674	0.372

7.2 Cross Validation

To illustrate CHAD’s ability to train models that can generalize, we perform cross validation experiments with another anomaly dataset in the same domain. We choose the popular ShanghaiTech Campus Dataset [17] for its relatively large size, its similar context to CHAD, and its proven track record in anomaly detection research. For these experiments, we use GEPC, as its multi-camera training methodology allows for a simple conversion to cross validation. For both CHAD and ShanghaiTech, a single model is trained for all cameras in one dataset, then tested on both datasets. The results can be seen in Table 6.

Table 6: Cross-validation of GEPC [23] on ShanghaiTech [17] and CHAD (Ours).

Model	Train	Test	AUC-ROC	AUC-PR	EER
GEPC	CHAD (Ours)	CHAD (Ours)	0.649	0.587	0.385
		ShanghaiTech	0.728	0.637	0.326
	ShanghaiTech	CHAD (Ours)	0.639	0.572	0.399
		ShanghaiTech	0.741	0.657	0.315

The first thing to notice is that models trained on CHAD perform well on ShanghaiTech, and models trained on ShanghaiTech perform well on CHAD. This is logical, as the contexts for the two datasets (i.e. setting, camera views, anomalous behaviors) are quite similar. In all metrics, the validation of models across datasets performs within 1-2% of models validated on their parent datasets, showing that models trained on either can generalize quite well given their similar contexts.

Another trend seen in Table 6 is that for all metrics, models tend to achieve lower scores (or higher in the case of EER) on CHAD than they do on ShanghaiTech. Since both models performed equally well in cross validation, the logical assumption is that CHAD’s test set is more challenging than ShanghaiTech’s. This is in part due to the additional noise and distractors present in CHAD, as explained in Section 5. The other major factor is the inclusion of very subtle and complex anomalies in CHAD. Pick-pocketing is subtle by design, as most pick-pockets are trying not to be seen. Littering is also quite complex to learn, especially for a model that relies solely on human keypoints. Combined with the sheer size of CHAD’s test set ($3\times$ that of ShanghaiTech’s), this makes for a very challenging dataset for current anomaly detection algorithms.

8 Conclusion

This paper presented the Charlotte Anomaly Dataset (CHAD). Consisting of more than $1.15m$ high-resolution frames of a single scene, CHAD is the largest available anomaly detection dataset consisting of continuous video from stationary cameras. In addition to frame-level anomaly labels, CHAD goes further than other datasets and provides bounding-box, person ID, and human keypoints annotations, enabling a unified benchmarking standard for both skeleton and appearance-based anomaly detection. Additionally, this paper assesses three metrics for anomaly detection and proposes their use in combination as a new standard for real-world video anomaly detection.

Acknowledgements This research is supported by the National Science Foundation (NSF) under Award No. 1831795 and NSF Graduate Research Fellowship Award No. 1848727. In addition, this research is IRB approved under Document Number IRBIS-17-0307.

References

1. Acsintoae, A., Florescu, A., Georgescu, M., Mare, T., Sumedrea, P., Ionescu, R.T., Khan, F.S., Shah, M.: Ubnormal: New benchmark for supervised open-set video anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2022)
2. Adam, A., Rivlin, E., Shimshoni, I., Reinitz, D.: Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE transactions on pattern analysis and machine intelligence* **30**(3), 555–560 (2008)
3. Alinezhad Noghre, G., Danesh Pazho, A., Sanchez, J., Hewitt, N., Neff, C., Tabkhi, H.: Adg-pose: Automated dataset generation for real-world human pose estimation. In: International Conference on Pattern Recognition and Artificial Intelligence. pp. 258–270. Springer (2022)
4. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
5. Chandra, R., Bhattacharya, U., Roncal, C., Bera, A., Manocha, D.: Robusttp: End-to-end trajectory prediction for heterogeneous road-agents in dense traffic with noisy sensor inputs. In: ACM Computer Science in Cars Symposium. pp. 1–9 (2019)
6. Chu, W., Xue, H., Yao, C., Cai, D.: Sparse coding guided spatiotemporal feature learning for abnormal event detection in large videos. *IEEE Transactions on Multimedia* **21**(1), 246–255 (2019). <https://doi.org/10.1109/TMM.2018.2846411>
7. Crowley, J.L., Parker, A.C.: A representation for shape based on peaks and ridges in the difference of low-pass transform. *IEEE transactions on pattern analysis and machine intelligence* (2), 156–170 (1984)
8. Davis, J., Goadrich, M.: The relationship between precision-recall and roc curves. In: Proceedings of the 23rd international conference on Machine learning. pp. 233–240 (2006)
9. Doshi, K., Yilmaz, Y.: Rethinking video anomaly detection—a continual learning approach. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 3961–3970 (2022)
10. Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B., Herrera, F.: Learning from imbalanced data sets, vol. 10. Springer (2018)
11. Ganokratanaa, T., Aramvith, S., Sebe, N.: Anomaly event detection using generative adversarial network for surveillance videos. In: 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). pp. 1395–1399 (2019). <https://doi.org/10.1109/APSIPAASC47483.2019.9023261>
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K. (eds.) *Advances in Neural Information Processing Systems*. vol. 27. Curran Associates, Inc. (2014), <https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf>
13. He, H., Ma, Y.: Imbalanced learning: foundations, algorithms, and applications. Wiley-IEEE Press (2013)
14. Li, N., Chang, F., Liu, C.: Human-related anomalous event detection via spatial-temporal graph convolutional autoencoder with embedded long short-term memory network. *Neurocomputing* **490**, 482–494 (2022)
15. Li, W., Mahadevan, V., Vasconcelos, N.: Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence* **36**(1), 18–32 (2013)

16. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
17. Liu, W., Luo, D.L., Gao, S.: Future frame prediction for anomaly detection – a new baseline. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
18. Liu, W., Luo, W., Lian, D., Gao, S.: Future frame prediction for anomaly detection – a new baseline. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
19. Lu, C., Shi, J., Jia, J.: Abnormal event detection at 150 fps in matlab (2013)
20. Luo, W., Liu, W., Gao, S.: Normal graph: Spatial temporal graph convolutional networks based prediction network for skeleton based video anomaly detection. *Neurocomputing* **444**, 332–337 (2021). <https://doi.org/https://doi.org/10.1016/j.neucom.2019.12.148>, <https://www.sciencedirect.com/science/article/pii/S0925231220317720>
21. Lv, Y., Jiang, G., Yu, M., Xu, H., Shao, F., Liu, S.: Difference of gaussian statistical features based blind image quality assessment: A deep learning approach. In: 2015 IEEE International Conference on Image Processing (ICIP). pp. 2344–2348. IEEE (2015)
22. Mahadevan, V., Li, W., Bhalodia, V., Vasconcelos, N.: Anomaly detection in crowded scenes. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 1975–1981 (2010). <https://doi.org/10.1109/CVPR.2010.5539872>
23. Markovitz, A., Sharir, G., Friedman, I., Zelnik-Manor, L., Avidan, S.: Graph embedded pose clustering for anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10539–10547 (2020)
24. Morais, R., Le, V., Tran, T., Saha, B., Mansour, M., Venkatesh, S.: Learning regularity in skeleton trajectories for anomaly detection in videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
25. Pranav, M., Zhenggang, L., K, S.S.: A day on campus - an anomaly detection dataset for events in a single camera. In: Proceedings of the Asian Conference on Computer Vision (ACCV) (November 2020)
26. Ramachandra, B., Jones, M.J.: Street scene: A new dataset and evaluation protocol for video anomaly detection. In: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 2558–2567 (2020). <https://doi.org/10.1109/WACV45572.2020.9093457>
27. Ravanbakhsh, M., Sangineto, E., Nabi, M., Sebe, N.: Training adversarial discriminators for cross-channel abnormal event detection in crowds. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1896–1904 (2019). <https://doi.org/10.1109/WACV.2019.00206>
28. Rodrigues, R., Bhargava, N., Velmurugan, R., Chaudhuri, S.: Multi-timescale trajectory prediction for abnormal human activity detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (March 2020)
29. Saito, T., Rehmsmeier, M.: The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one* **10**(3), e0118432 (2015)
30. Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)

31. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5693–5703 (2019)
32. Tian, Y., Pang, G., Chen, Y., Singh, R., Verjans, J.W., Carneiro, G.: Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4975–4986 (October 2021)
33. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: 2017 IEEE international conference on image processing (ICIP). pp. 3645–3649. IEEE (2017)
34. Wu, P., Liu, J., Shi, Y., Sun, Y., Shao, F., Wu, Z., Yang, Z.: Not only look, but also listen: Learning multimodal violence detection under weak supervision. In: European conference on computer vision. pp. 322–339. Springer (2020)
35. Zhou, S., Shen, W., Zeng, D., Fang, M., Wei, Y., Zhang, Z.: Spatial-temporal convolutional neural networks for anomaly detection and localization in crowded scenes. *Signal Processing: Image Communication* **47**, 358–368 (2016). <https://doi.org/https://doi.org/10.1016/j.image.2016.06.007>, <https://www.sciencedirect.com/science/article/pii/S0923596516300935>