INSTITUTE FOR HOMELAND SECURITY

Sam Houston State University™

ALGORITHMS AND DATASET FOR ANOMALY DETECTION IN SMART VIDEO SURVEILLANCE FOR INFRASTRUCTURE SAFETY

Institute for Homeland Security

Sam Houston State University

Hamed Tabkhivayghan

Algorithms and Dataset for Anomaly Detection in Smart Video Surveillance for Infrastructure Safety

Department of Electrical and Computer Engineering
University of North Carolina Charlotte
Charlotte, NC 28223, USA,
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 collected from continuous video, more than 2x the size of the next largest. 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.

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. 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 continued 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, 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, 2.9× that of the next largest comparable dataset, CHAD places itself as the premiere video anomaly dataset for unsupervised methods. Additionally, CHAD provides human detection, tracking, and pose annotations, allowing for a more accurate standard and 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.

2 Related Work

Anomaly Detection Algorithms Appearance-based methods utilize appearance and motion features generated directly from pixel data for detecting anomalies (Chu et al. 2019; Zhou et al. 2016; Sultani, Chen, and Shah 2018; Tian et al. 2021; Goodfellow et al. 2014; Ganokratanaa, Aramvith, and Sebe 2019; Ravanbakhsh et al. 2019; Liu et al. 2018). These methods generally achieve high accuracy in their context at the cost of high computation. Skeleton-based methods utilize high-level, low-dimensional human pose skeletons (Rodrigues et al. 2020; Luo, Liu, and Gao 2021; Markovitz et al. 2020; Morais et al. 2019; Li, Chang, and Liu 2022). 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 (Lu, Shi, and Jia 2013) 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 (Mahadevan et al. 2010) 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 (Adam et al. 2008) consists of two surveillance videos, the subway entrance, and exit. With a combined total of 139 minutes of video, this dataset counts IITB-Corridor (Rodrigues et al. 2020) was the largest single-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 (Pranav, Zhenggang, and K 2020) 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 (Acsintoae et al. 2022) is composed entirely of synthetically generated videos. With a total of 236,902 frames, ABnormal is moderately large compared to other anomaly datasets, though with 29 scenes the average number of frames per scene is fairly low.

UCF Crime (Sultani, Chen, and Shah 2018) and X-D Violence (Wu et al. 2020) 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 differ ent 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

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 un satisfying performance in the real-world (Alinezhad Noghre et al. 2022). 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).

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. This list has been curated in line with other state-of-the-art datasets (Liu, W. Luo, and Gao 2018; Lu, Shi, and Jia 2013; Adam et al. 2008; Ma hadevan et al. 2010). All other actions present in CHAD (e.g. walking, waving, talking, etc.) are considered normal.

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.

4.2 Person Annotations

One of the innovations that sets CHAD above its peers is the inclusion of person annotations. Person-annotations ensure they are more representative of a real-world situation (Chandra et al. 2019). It allows skeleton-based anomaly detection methods 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 (Bochkovskiy, Wang, and Liao 2020) 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 (Wojke, Bewley, and Paulus 2017) 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 personal 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 key points to use, CHAD follows the 17 keypoint methodology proposed by MS COCO (Lin et al. 2014). Using the localization provided by the previously generated bounding boxes, keypoints are extracted using HRNet (Sun et al. 2019), 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.

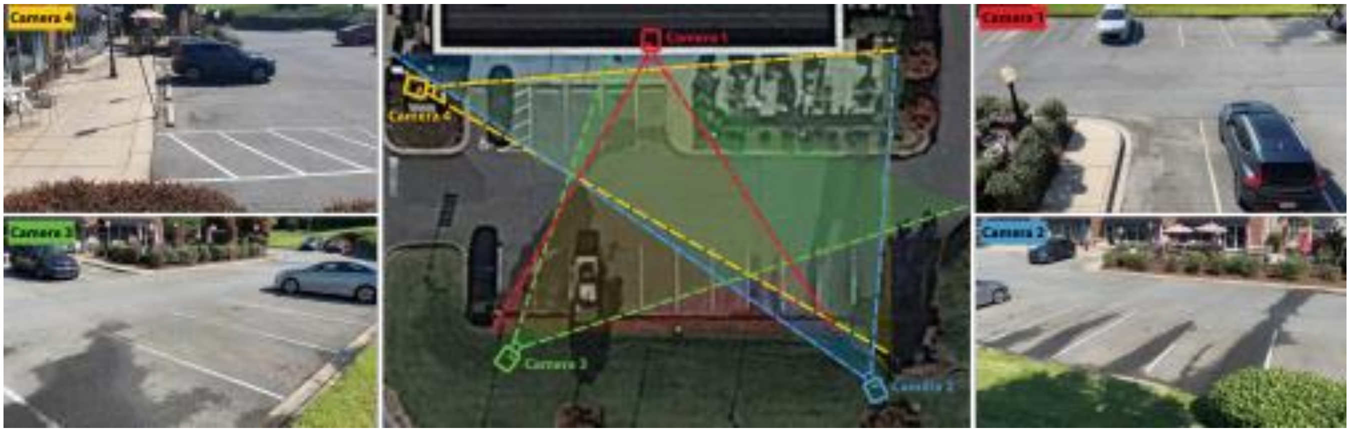

Figure 1: Approximate position and the views of the cameras.

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:

$$X_{i = (}X_N - X_{N - M}) \times i + X_M \quad (1)$$

where X_i refers to a missing point (either bounding box or keypoint coordinate) at frame i, X_M and X_N refer to the two nearest matching points at frames M and N respectively, and where $M < i < N$ and $N - M + 1 \leq 15$.

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 video. CHAD has more than 2× 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× more than can be found in other datasets, positioning CHAD as the best-in-class dataset for unsupervised approaches. The 59K anomalous frames in CHAD are the 22 anomalous behaviors presented. To facilitate supervised, unsupervised, and semi-supervised approaches, CHAD includes two splits for training and testing. The *unsupervised split* has a training set composed only of normal behaviors, while the test set contains both normal and anomalous behaviors. 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, 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 (Crowley and Parker 1984) is an image processing method that has been used to simulate how the human eye extracts visual details of an image for neural processing (Lv et al. 2015). 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 representa tion. 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 (Rodrigues et al. 2020) is the only other dataset with 1080p images. However, the Difference of Gaussian tells a different story.

Table: Annotation availability in Shanghai (Liu, W. Luo, and Gao 2018), CUHK (Lu, Shi, and Jia 2013), UCSD (Mahadevan et al. 2010), Subway (Adam et al. 2008), IITB (Rodrigues et al. 2020), Street Scene (Ramachandra and Jones 2020), UBnormal (Acsintoae et al. 2022), and CHAD (Ours). * partially annotated, − not annotated.

Frame-level Label	Pixel-level Label	Person Bounding Box	Person ID
✓	✓	−	−
✓	✓	−	−
✓	*	−	−
✓	✓	−	−
✓	✓	−	−
✓	✓	−	−
✓	✓	✓	−

While there are details present in the environment, they are comparably indistinct. Even in the brightened image, it is difficult to tell if there is a person in the image. This demonstrates a surprising lack of rich features in the IITB-Corridor, despite the resolution.

Resolution and frame rate in Shanghai (Liu, W. Luo, and Gao 2018), CUHK (Lu, Shi, and Jia 2013), UCSD (Ma hadevan et al. 2010), Subway (Adam et al. 2008), IITB (Rodrigues et al. 2020), Street Scene (Ramachandra and Jones 2020), UBnormal (Acsintoae et al. 2022), and CHAD (Ours). N/A means Not Available.

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 (Liu, W. Luo, and Gao 2018), 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 personal 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 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 (Fernandez et al. 2018). How- ´ ever, AUC-ROC is not indicative of the final decisions of a model. The metric does not indicate useful information about False Negative Rate (FNR), when an anomaly is classified as normal, which is important to understand for real world applications. Additionally, AUC-ROC is very sensitive to imbalances in data (He and Ma 2013), making it sub-optimal if one class is over represented, as is often the case with normal behaviors in anomaly datasets (Davis and Goadrich 2006).

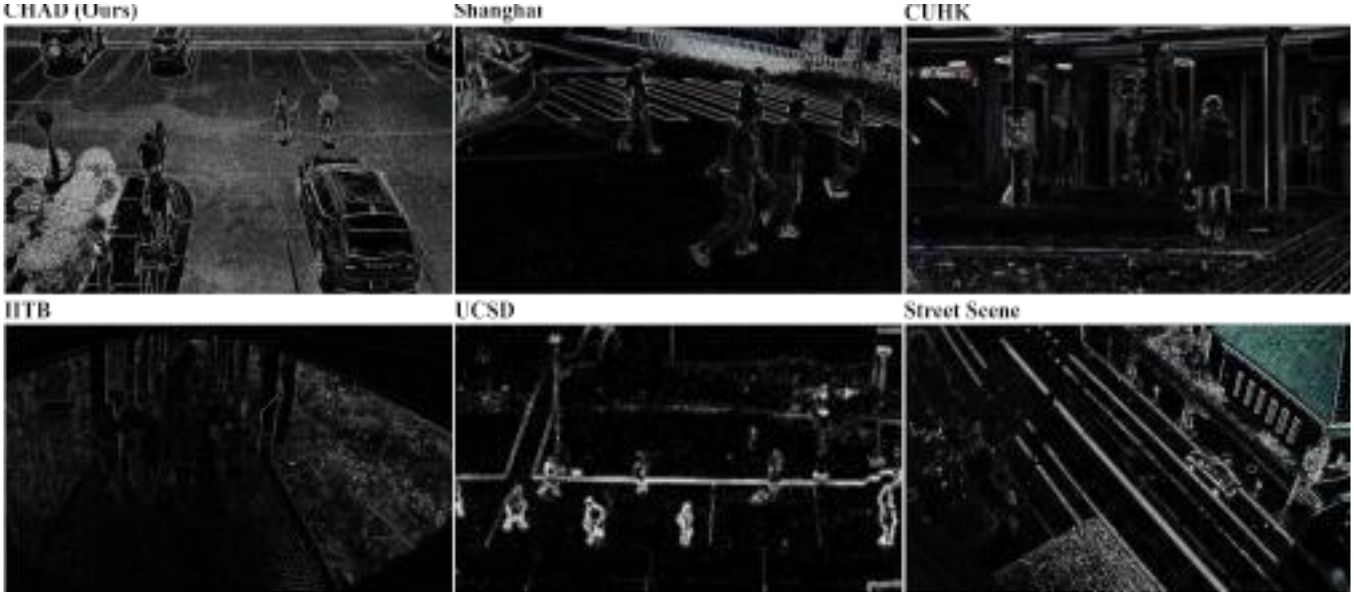

Figure 2: Visualization of Difference of Gaussian in Shanghai (Liu, W. Luo, and Gao 2018), CUHK (Lu, Shi, and Jia 2013), UCSD (Mahadevan et al. 2010), IITB (Rodrigues et al. 2020), Street Scene (Ramachandra and Jones 2020), and CHAD (Ours). UCSD cropped to fit. All brightened for readability.

Table: Dataset comparison for Shanghai (Liu, W. Luo, and Gao 2018), CUHK (Lu, Shi, and Jia 2013), UCSD (Mahadevan et al. 2010), Subway (Adam et al. 2008), IITB (Rodrigues et al. 2020), Street Scene (Ramachandra and Jones 2020), UBnormal (Acsintoae et al. 2022), and CHAD (Ours). CHAD uses an unsupervised split. N/A means Not Available.

Total	Train	Test	Normal	Anomalous	Number of scenes
317,398	274,515	42,883	300,308	17,090	13
30,652	15,328	15,324	26,832	3,820	1
18,560	9,050	9,210	12,919	5,641	2
208,925	27,500	181,425	205,805	3120	2
483,566	301,999	181,567	375,288	108,278	1

203,257	56,847	146,410	N/A	N/A	1
236,902	116,087	28,175	147,887	89,015	29

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 (Saito and Rehmsmeier 2015). Additionally, AUC-PR is better suited for highly imbalanced data (Saito and Rehmsmeier 2015), making it better at evaluation of the minority class (He and Ma 2013). As the minority class in anomaly detection usually refers to the anomalous behaviors, this is an important quality for this context. However, AUC-PR does not provide 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) (Li, Mahadevan, and Vasconcelos 2013). 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 (Sultani, Chen, and Shah 2018). However, when used as a complement to AUC-ROC and AUC-PR, a more complete understanding can be achieved.

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.

6.4 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) (Markovitz et al. 2020)

Both models were trained on each of CHAD's four cam era views individually. 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 be 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.

6.5 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 (Liu, W. Luo, and Gao 2018) 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 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.

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. This is in part due to the additional noise and distractions 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× that of Shanghaitech), this makes for a very challenging dataset for current anomaly detection algorithms.

7. Conclusion

This paper presented the Charlotte Anomaly Dataset (CHAD). Consisting of more than 1.15*m* high-resolution frames of a single scene, CHAD is the largest anomaly detection dataset recording from continuous video available. 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 bench marking 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.

References

Acsintoae, A.; Florescu, A.; Georgescu, M.; Mare, T.; Sume drea, P.; Ionescu, R. T.; Khan, F. S.; and Shah, M. 2022. UBnormal: New Benchmark for Supervised Open-Set Video Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Adam, A.; Rivlin, E.; Shimshoni, I.; and Reinitz, D. 2008. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE transactions on pattern anal ysis and machine intelligence*, 30(3): 555–560.

Alinezhad Noghre, G.; Danesh Pazho, A.; Sanchez, J.; He witt, N.; Neff, C.; and Tabkhi, H. 2022. ADG-Pose: Automated Dataset Generation for Real-World Human Pose Estimation. In *International Conference on Pattern Recognition and Artificial Intelligence*, 258–270. Springer.

Bochkovskiy, A.; Wang, C.-Y.; and Liao, H.-Y. M. 2020. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*.

Chandra, R.; Bhattacharya, U.; Roncal, C.; Bera, A.; and Manocha, D. 2019. Robusttp: End-to-end trajectory prediction for heterogeneous road-agents in dense traffic with noisy sensor inputs. In *ACM Computer Science in Cars Symposium*, 1–9.

Chu, W.; Xue, H.; Yao, C.; and Cai, D. 2019. Sparse Coding Guided Spatiotemporal Feature Learning for Abnormal Event Detection in Large Videos. *IEEE Transactions on Multimedia*, 21(1): 246–255.

Crowley, J. L.; and Parker, A. C. 1984. 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.

Davis, J.; and Goadrich, M. 2006. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*, 233– 240.

Fernandez, A.; Garc ´ ´ıa, S.; Galar, M.; Prati, R. C.; Krawczyk, B.; and Herrera, F. 2018. *Learning from imbalanced data sets*, volume 10. Springer.

Ganokratanaa, T.; Aramvith, S.; and Sebe, N. 2019. Anomaly Event Detection Using Generative Adversarial Network for Surveillance Videos. In *2019 Asia-Pacific Sig nal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 1395–1399.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N.; and Weinberger, K., eds., *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

He, H.; and Ma, Y. 2013. *Imbalanced learning: foundations, algorithms, and applications*. Wiley-IEEE Press. Li, N.; Chang, F.; and Liu, C. 2022. Human-related anomalous event detection via spatio-temporal graph convolu tional autoencoder with embedded long short-term memory network. *Neurocomputing*, 490: 482–494.

Li, W.; Mahadevan, V.; and Vasconcelos, N. 2013. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence*, 36(1): 18–32.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ra manan, D.; Dollar, P.; and Zitnick, C. L. 2014. Microsoft ´ coco: Common objects in context. In the European *conference on computer vision*, 740–755. Springer.

Liu, W.; Luo, W.; Lian, D.; and Gao, S. 2018. Future Frame Prediction for Anomaly Detection – A New Baseline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liu, W.; W. Luo, D. L.; and Gao, S. 2018. Future Frame Prediction for Anomaly Detection – A New Baseline. In *2018 IEEE Conference on Computer Vision and Pattern Recogni tion (CVPR)*.

Lu, C.; Shi, J.; and Jia, J. 2013. Abnormal Event Detection at 150 FPS in Matlab.

Luo, W.; Liu, W.; and Gao, S. 2021. Normal graph: Spatial temporal graph convolutional networks based prediction network for skeleton based video anomaly detection. *Neurocomputing*, 444: 332–337.

Lv, Y.; Jiang, G.; Yu, M.; Xu, H.; Shao, F.; and Liu, S. 2015. Difference of Gaussian statistical features based blind image quality assessment: A deep learning approach. In *2015 IEEE International Conference on Image Processing (ICIP)*, 2344–2348. IEEE.

Mahadevan, V.; Li, W.; Bhalodia, V.; and Vasconcelos, N. 2010. Anomaly detection in crowded scenes. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1975–1981.

Markovitz, A.; Sharir, G.; Friedman, I.; Zelnik-Manor, L.; and Avidan, S. 2020. Graph embedded pose clustering for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10539–10547.

Morais, R.; Le, V.; Tran, T.; Saha, B.; Mansour, M.; and Venkatesh, S. 2019. Learning Regularity in Skeleton Trajectories for Anomaly Detection in Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Pranav, M.; Zhenggang, L.; and K, S. S. 2020. A Day on Campus - An Anomaly Detection Dataset for Events in a Single Camera. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*.

Ramachandra, B.; and Jones, M. J. 2020. Street Scene: A new dataset and evaluation protocol for video anomaly detection. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2558–2567.

Ravanbakhsh, M.; Sangineto, E.; Nabi, M.; and Sebe, N. 2019. Training Adversarial Dis Criminators for Cross Channel Abnormal Event Detection in Crowds. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1896–1904.

Rodrigues, R.; Bhargava, N.; Velmurugan, R.; and Chaudhuri, S. 2020. Multi-timescale Trajectory Prediction for Abnormal Human Activity Detection. In *Proceedings of the*

IEEE/CVF Winter Conference on Applications of Computer Vision (WACV).

Saito, T.; and Rehmsmeier, M. 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3): e0118432.

Sultani, W.; Chen, C.; and Shah, M. 2018. Real-World Anomaly Detection in Surveillance Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sun, K.; Xiao, B.; Liu, D.; and Wang, J. 2019. Deep high resolution representation learning for human pose estimation. Proceedings *of the IEEE/CVF conference on computer vision and pattern recognition*, 5693–5703.

Tian, Y.; Pang, G.; Chen, Y.; Singh, R.; Verjans, J. W.; and Carneiro, G. 2021. Weakly-Supervised Video Anomaly Detection With Robust Temporal Feature Magnitude Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4975–4986.

Wojke, N.; Bewley, A.; and Paulus, D. 2017. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, 3645–3649. IEEE.

Wu, P.; Liu, J.; Shi, Y.; Sun, Y.; Shao, F.; Wu, Z.; and Yang, Z. 2020. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In the European *conference on computer vision*, 322–339. Springer.

Zhou, S.; Shen, W.; Zeng, D.; Fang, M.; Wei, Y.; and Zhang, Z. 2016. Spatial–temporal convolutional neural networks for anomaly detection and localization in crowded scenes. *Signal Processing: Image Communication*, 47: 358–368.

INSTITUTE FOR
HOMELAND SECURITY

Sam Houston
State University ™

The Institute for Homeland Security at Sam Houston State University is focused on building strategic partnerships between public and private organizations through education and applied research ventures in the critical infrastructure sectors of Transportation, Energy, Chemical, Healthcare, and Public Health.

The Institute is a center for strategic thought with the goal of contributing to the security, resilience, and business continuity of these sectors from a Texas Homeland Security perspective. This is accomplished by facilitating collaboration activities, offering education programs, and conducting research to enhance the skills of practitioners specific to natural and human caused Homeland Security events.

Institute for Homeland Security
Sam Houston State University

Tabkhivayghan, H. (2022). **Algorithms and Dataset for Anomaly Detection in Smart Video Surveillance for Infrastructure Safety** (Report No. IHS/CR-2022-2038). The Sam Houston State University Institute for Homeland Security. https://ihsonline.org/Research/Technical-Papers/Algorithm-and-Dataset-for-Anomaly-Detection-in-Smart-Video-Surveillance-for-Infrastructure-Safety