RW-SVD: A surround view rough weather video anomaly dataset and a brief overview of existing datasets

Sachin Dube[1], Dinesh Kumar Tyagi[1], and Ramesh Babu Battula[1]

Malaviya National Institute of Technology Jaipur, Jaipur, India-302017
sachin.rdubey.2011@gmail.com, {dktyagi.cse,rbbattula.cse}@mnit.ac.in

Abstract. Modern surveillance societies constantly face bottlenecks due to manual monitoring of huge amounts of data generated by surveillance infrastructure. The limitation of manual monitoring is further aggravated by challenging weather conditions such as fog, rain, mist, etc. This gave rise to automated surveillance making Video anomaly detection (VAD) one of the most sought-after domains in computer vision. The availability of data that contain weather-induced variations is a key factor in the effectiveness of Data-driven approaches that rely on data for precise modeling. To this end, we have presented a brief review of previous datasets and their limitations on parameters such as size, scene variations, activities covered, effect of weather phenomena, etc. To leverage the intricate relationship between data and model we present a novel human-centric surround view dataset where each scripted activity is recorded simultaneously by 4 strategically placed cameras to capture effects of varying distance, angle, height, and illumination on the same scene. The proposed dataset is arranged into 4 abnormal classes namely fighting, snatching, panic running, and kidnapping. It contains complex backgrounds, real-life objects (cycle, motorbike, four-wheeler), both indoor and outdoor environments as well as illumination change. To tackle ambiguity during the transition from normal to abnormal or vice-versa we conducted voting (subjective evaluation) with 10 volunteers. We further augmented the dataset with two of the most common weather phenomena namely haze and rain to bridge the gap between real-world challenges and dataset.

Keywords: Anomaly Detection · Database · RW-SVD.

1 Introduction

Video anomaly detection (VAD) refers to autonomous identification as well as localisation of abnormal events in a video stream. It reduces dependency on manpower thereby increasing efficiency and effectiveness. It has become a necessary tool to realise the true potential of huge amounts of data collected from various sources. It forms the basis of various vision-based applications such as crime and violence detection [1,2], traffic monitoring and management [3], Accident detection [4], intelligent surveillance systems [5,6], disaster management [7]

etc. Data distribution happens to be core of various learning-based VAD approaches that can be further divided into 4 categories: *1.supervised learning:* based approaches [8,9] require well-defined labels and are often trained on small single scene datasets such as UCSD [10,11], UMN [12], Subway [13] etc.. *2.unsupervised learning*: based approaches [14,15] are generally powered by large scale datasets such as UCF-Crime [16] where, frame level labels are not feasible. These approaches usually try to build a latent representation of normality only from normal data and completely ignore abnormal data. *3.weakly supervised learning* : based approaches [17, 18] act as a bridge between the above two approaches and uses large-scale datasets such as XD-Violence [19]. Instead of frame; a label is assigned to videos. However, it may be possible for a video to contain multiple snippets that can be normal or anomalous, that anomaly decision criterion happens to be a field of active research.

Weather-affected data has seen constant interest in traffic management, road safety, and autonomous driving domains due to apparent reasons. many datasets such as X-MAN [20], AI city challenge [21],Dawn and Dusk [22] etc. have been proposed to cater to that need. However, this trend is being picked up in the anomaly detection domain but the availability of data, especially human-centric data remains a challenge. Apart from this very specific challenge issues of data imbalance, multi-view real-world data, appropriate labels, ambiguous labels, etc remain. To this end, we propose a human-centric surround view dataset that is augmented to reflect weather-induced challenges. We summarise our contribution through the following points:

- A brief review of most sought-after benchmark datasets. The review includes a mix of single-view, multi-view, large and small scale as well as weather-specific datasets. Information like technical details, strengths, and weaknesses is also provided
- A human-centric surround-view video anomaly dataset that contains the effect of rough weather conditions like haze and rain is proposed. It is obtained by augmenting previously proposed Anovil [23] by using statistical methods after fine-tuning acc. to data.
- Some samples of neutral weather, haze, and rain-affected frames for multiple combinations like normal-abnormal, indoor-outdoor, and different activities are presented along with frame-wise quantitative data.

The rest of the paper is divided into 3 sections. Section 2 gives a brief review of various datasets while the proposed dataset and augmentation technique are described in Section 3. It also contains snippets for subjective evaluation. Section 4 concludes the paper.

2 Review of existing benchmark datasets

Here we present a brief overview of the most commonly used benchmark datasets by dividing them into three categories namely: single-view datasets, multiple-view datasets, and weather-related datasets. We further discuss recently proposed weather-centric datasets.

2.1 Single scene datasets

Fig. 1. Abnormal snippets from UCSD- Dataset

UCSD pedestrian Dataset for Anomaly Detection [10, 11]: consists of a grayscale video clip of a crowded walkway. It is shot by a static camera under two different scenes: Peds1 and Peds2 as shown in figure 1. Pedestrians walking on the path are considered to be normal and the movement of cyclists, skaters, bikers, carts, wheelchairs, etc. in the pathway is considered to be anomalous. Even the movement of pedestrians outside the designated pathway (lawn) is also considered to be anomalous. Peds 1 includes a total of 70 video clips ranging from 5 to 10 seconds (200-400 frames). The training set consists of 34 video clips while the test set consists of 36 video clips. Peds 2 dataset is even smaller and contains 28 video clips ranging from 2-4 seconds (120-180 frames). The training set consists of 16 video clips while the test set consists of 12 video samples. Both datasets contain annotations at both pixel and frame level. Hence, it facilitates the evaluation of the localization performance of SOTA methods.

UMN Dataset [12]: is primarily designed to capture crowd dynamics and includes panic-driven situations such as sudden transitions from walking to running. It contains about 8000 frames (320x240) from 3 scenarios namely: lawn, plaza, and indoor. It is developed around only two events: pedestrians walking and panic running while trying to escape. The normal scene is characterised by individuals casually walking in various directions, while the abnormal scene contains individuals running in disarray. These videos are captured in varying scenes and environments ensuring diversity in motion patterns.

PETS Dataset *[24]*: pets 2009 dataset is primarily focused on video surveillance applications such as human and vehicular motion tracking, crowd dynamics, etc. It features real-world video clips captured by multiple synchronised cameras at 576×768. this helps with 3D- tracking, occlusion, and re-identification problems. Captured events include; pedestrians walking, moving vehicles, crowd movement, etc. The dataset is meticulously annotated with bounding boxes around people, and vehicles along with corresponding labels. Major challenges presented by the dataset are illumination and environmental variations, occlusion, and crowd variability. The pets2010 dataset is an extension of the pets2009 datasets and includes more complex scenarios such as loitering, unattended objects, aggressive behaviour, etc.

Subway Dataset [13]: It is a real-world dataset shot at a relatively low resolution and longer duration. It contains two video sequences: *1) subway entrance:* focuses on people entering the subway station. It is 96 minutes long (144249 frames) and contains 66 abnormal activities such as individuals walking in the wrong direction, abrupt Running, and stopping. *2 subway exit:* captures people leaving the subway station. It is 43 minutes long (64900 frames) and encompasses 19 types of anomalous events, such as loitering near the exit. Major challenges posed by the dataset include occlusion, illumination variation, and abrupt motion.

CUHK Avenue Dataset [25]: is shot to capture events at the CUHK Campus to represent the crowded urban environment. It consists of 30 minutes of footage at 25 FPS with 640x360 resolution. The training set encompasses 16 training videos containing only normal events while the test set consists of 21 video clips that include both normal as well as anomalous events. Throwing bags, sudden running, stopping, or changing direction is considered to be an anomaly. To further increase challenges, camera shake is introduced deliberately in test video frames. The dataset is temporally and spatially annotated.

UBnormal Dataset [26]: is a software-generated synthetic dataset (30 FPS) that offers diverse real-world scenarios in a controlled environment to overcome issues with real-world videos such as privacy concerns, inconsistent annotations, etc. As it is computer generated it offers high-resolution videos with consistency. The dataset is split into training (268), validation (64), and testing (211) sets. It provides frame-level as well as video-level annotations especially abnormal events annotated at pixel-level making it suitable for fully supervised methods also. Disjoint sets of anomalies are maintained in both: training and test sets to facilitate open-set formulation.

UBI-Fights dataset [27]: is focused on fighting and contains diverse real-life scenarios in indoor-outdoor, coloured-grayscale, fixed-movable cameras with varied orientations. Irrelevant frames that could hinder the learning process are removed manually and the dataset is resized to 640x360 pixels at 30 FPS. It contains a total of 1000 videos amounting to 80 hours of playtime. However major chunk of the dataset is made up of normal videos (784) while the number of abnormal videos is less than 1/4th of the entire dataset (216). More than half of videos are less than 2 minutes while 98 videos have lengths of more than 10 minutes

MSAD (Multi-Scenario Anomaly Detection) dataset [28]: is a multi-view, multi-scenario dataset that contains a total of 720 videos of real-life scenarios captured by surveillance cameras. It contains 55 anomaly types: 20 non-human related and 35 Human related. Variations such as indoor-outdoor, effect of weather, multiple camera views, illumination variation as well diverse motion patterns present major challenges as shown in figure 2. Frame-level annotations are provided.

NTU CCTV-Fights Dataset [29]: contains 1000 videos ranging from 5 to 720 seconds with a total runtime of well over 17 hours. These videos were recorded from CCTV, mobile, or dashboard cameras and were collected

Fig. 2. Snippets from MSAD dataset showing diversity and challenges offered by dataset [28].

from YouTube. There are a total of 2414 annotated fight instances where the video contains at least one fight instance from the following categories: Pushing, punching, kicking, and wrestling involving two or more individuals. Frame-level annotation along with the starting and end points of each fighting instance are provided.

DoTA dataset [30]: contains 4677 videos focused on traffic anomaly in terms of three aspects : 1) **When** represents temporal aspects as it includes videos recorded at different times of day, leading to diverse traffic conditions (ex. Peak traffic in Rush hours). 2) **Where** represents geographic locations such as intersections, highways, and urban areas. 3) **What** represents types of traffic anomalies, such as Accident, Traffic violations, Blockage, and Pedestrian-related incidents These anomalies are annotated according to category allowing frame and video level detection. It includes challenges presented by multiple viewpoints, and weather conditions such as rain, fog, low visibility, etc. It finds primary application in the area of autonomous driving and intelligence surveillance systems.

D^2-City Dataset [31]: is a collection of 10000 dashboard camera videos shot at 720 and 1080 P and is primarily dedicated to the development of ADAS and intelligent driving systems. It includes bounding boxes for 1000 videos along with tracking information. These annotations cover 12 types of objects ranging from pedestrians, bicycles, tricycles, cars, buses, etc.

QMUL Junction Dataset [32]: focuses on traffic scene analysis and is shot at a busy public road regulated by traffic lights. It includes two sets of videos shot at 25 FPS and 360×288 resolution. The first one is an hour long (90,000 frames) while the second one is 52 minutes long (78,000 frames). The major challenge comes from complex and diverse interactions between vehicles and pedestrians, variable flow of traffic along with illumination variations, shadows, and noise. Traffic violations such as Traffic interruption, illegal u-turns, stoppages, etc. constitute anomalous samples. However, it lacks formal partitioning into training and test sets.

Street Scene Dataset [33]: The Street Scene dataset presents urban environments, focusing on capturing scenes from city streets. The training set contains 46 (56,847 frames) videos and the test set contains 55 (146,410 frames)

videos shot at 1280x720 at 15 fps. Videos typically include a range of elements such as vehicles, pedestrians, buildings, traffic signs, and other street furniture. It contains 205 anomalous events and presents challenges like varied lighting conditions, weather, and urban layouts, making it valuable for developing robust AI systems that can interpret complex street environments.

CAVIAR dataset (Context-aware vision using image-based active recognition) [34]: is a publicly available dataset that can be used for pedestrian detection and tracking, anomaly detection, and activity recognition. It encompasses two scenarios shot at 25 FPS and 384 x 288 resolution at INIRA lab France and a shopping complex in Lisbon. It contains several scenarios such as walking, group interactions, Entering and leaving shops, window shopping, etc., and unusual activities like panic running, fighting, loitering, abandoning a package, etc. It has frame-level annotations of the activities mentioned above.

2.2 Multi-scene datasets

BEHAVE Dataset [35]: focuses on crime-oriented behaviour. It contains 4 video clips of staged scenarios recorded in a controlled outdoor environment at 25 FPS and 640 x 480 pixels by a static camera. Fighting, chasing, and running together form the major portion of anomalous instances. It includes detailed annotation by drawing a bounding box for each interacting individual. It includes individual as well as group activities. However, it also lacks formal portioning into train and test sets.

Web Dataset [36]: is a crowd oriented dataset collected from various websites. It contains videos of crowds in various urban scenarios. It includes a total of 20 high-quality documentary videos. 12 videos contain normal scenes like walking, and running, while the anomalous sample contains 8 videos with activities like panic escape, fighting and protesters clashing.

MIT Traffic Dataset [37]: is a 90-minute long video (20 clips) shot by a static camera at 30 FPS and 720 $\times$ 480 resolution to capture traffic activity. The scene contains multiple vehicles and pedestrians with a less dense but more erratic and unorganised flow of traffic. Ground truth annotations are provided from some frames focused on pedestrian detection.

UCF Crime [16]: dataset is a collection of 1900 variable length untrimmed surveillance footage of real-world incidents exceeding playtime of 128 hours. It is collected from various social media platforms and is further divided into 13 types of anomalous (criminal) events namely: Arrest, Abuse, Assault, Arson, Burglary, Fighting, Explosion, Robbery, Stealing, Shoplifting, Vandalism, Shooting, Road Accidents apart from Normal videos. It has a balanced distribution (950 normal and 950 abnormal videos) and is weakly labeled, i.e. each video is tagged with one type of anomaly without precise temporal annotations, posing a significant challenge for models to detect and localise anomalies. It includes variations in lighting, camera angles, background, and occlusions that are typical in real-world scenarios. The training set consists of 1610 videos while the Test set is made of 290 videos (150 abnormal and 140 abnormal).

XD-Violence Dataset [19]: is a large-scale, multimodal (includes aural and visual information), multi-scene dataset collected from movies, news reports, sports streaming, surveillance videos, etc. It's a comparatively balanced dataset that consists of 4,754 (2349 normal, 2405 abnormal) untrimmed videos totaling more than 217 hours of footage, and covers 6 types of violent activities namely: Fighting, Rioting, Arson, Explosions, Abuse, and Shooting. It is also weakly supervised; however, it contains video-level annotation for the training set but frame-level annotations for the test set. The training set is made up of 3954 videos (2049 normal, 1905 abnormal) while the test set is made of 800 videos (300 normal, 500 abnormal). Currently, it happens to be the most comprehensive and voluminous public dataset for real-world VAD detection and introduces a range of challenges, such as varying camera angles, lighting conditions, and background noise.

2.3 Weather related datasets

While datasets shot outdoors are prone to weather phenomena and have a few weather-induced variations eg. UCF-Crime [19], street scene [33], CUHK-Avenue [25] etc. yet, but to cater to research in weather-sensitive application few dedicated datasets are proposed as follows.

X-MAN: (Extreme Weather Anomaly Dataset) [20]: is focused on video Anomalies under extreme weather conditions such as heavy rain, snow, fog, etc. It directly addresses how extreme weather affects the visibility and detection of anomalies.

AI City Challenge Dataset (Track 3: Anomaly Detection) [21]: focuses on Traffic and urban anomaly detection and includes weather conditions like rain, fog, and cloudy weather. It explicitly claims to include weather variability, with the goal of simulating real-world driving and traffic anomalies that occur under different weather conditions.

Dawn and Dusk Dataset [22]: focuses on Anomaly detection in driving scenarios for autonomous driving. It includes challenging lighting conditions; Under the influence of fog, and rain. It also includes low-visibility scenarios in addition to varying light conditions such as dawn and dusk.

While the above datasets cater to the need for intelligent transport systems and ADAS-based applications There is no dedicated dataset for human-centric anomaly detection in urban settings under the influence of weather. We aim to fill this gap with RW-SVD.

3 Proposed Dataset RW-SVD

Information conveyed by visual data i.e. images and videos is highly dependent on various environmental factors as well as on composition. Change of perspective/viewpoint can change context; which, in turn, influences the interpretation of events recorded in a scene. We leverage this property and propose

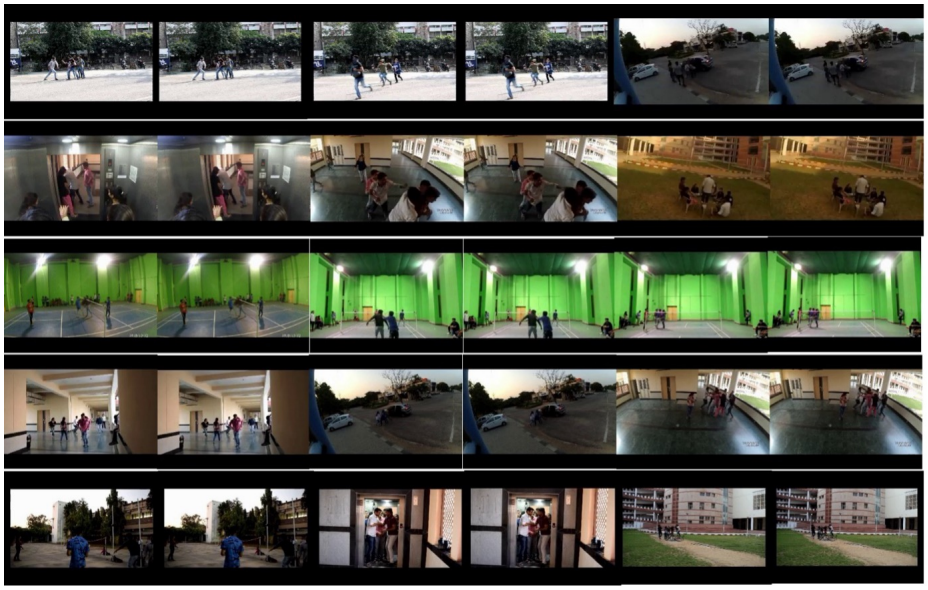

Fig. 3. Abnormal snippets from normal weather scenarios showing diverse backgrounds, viewpoints, illumination conditions, and action classes

a Human-Centric surround-view dataset that captures a particular event of interest simultaneously with 4 different cameras placed at varying angles, heights, and distances from the subject as shown in fig.4.

Feeds received from different cameras show drastic variations in visual information due to changes in viewpoint, illumination, size, and background as shown in Fig.-5. The salient features of the proposed dataset can be summarized as follows:

- Shot on four strategically placed cameras simultaneously to capture multiple views of the same scene to cover viewpoint changes shown in Figure 6
- Shot on six different locations like a park, indoor/outdoor courts, hallway, road, and lift at a different time of day.
- Shot from varied distances and heights to capture subjects with varying size
- Scenes involve instances of kidnapping, snatching, Fighting, and panic running along with normal scenes during activities like strolling, playing, cycling, etc.
- Scenes were shot indoors/outdoors, in natural light-artificial light, and low light as well to cover illumination changes.

Fig. 6 presents snippets of the same incident taken from different cameras to highlight the difference in visual information. The difference is so drastic that an anomalous event may look normal from a certain angle or vice versa. The wrong perspective can be deceiving at times. This combined with the transition from

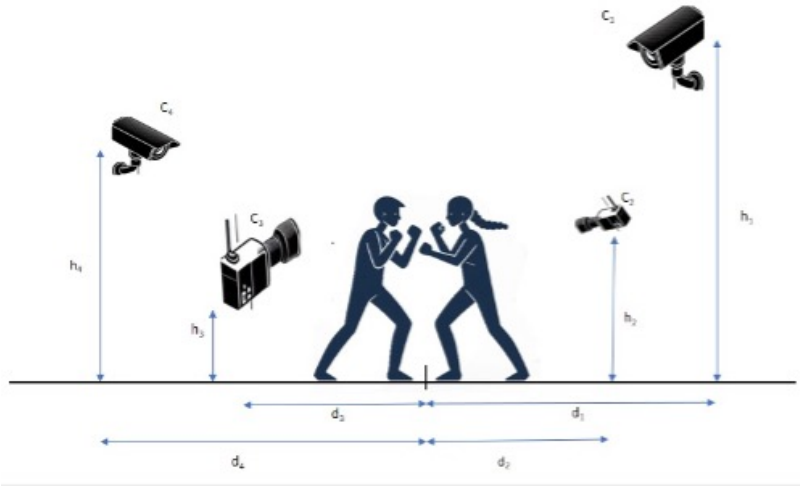

Fig. 4. Shows two subjects fighting are simultaneously being recorded by 4 different cameras C1, C2, C3, and C4. They are placed at different heights h1,h2,h3, and h4 from the ground (hence subjects) as well as different distances d1,d2,d3, and d4 from the subjects respectively.

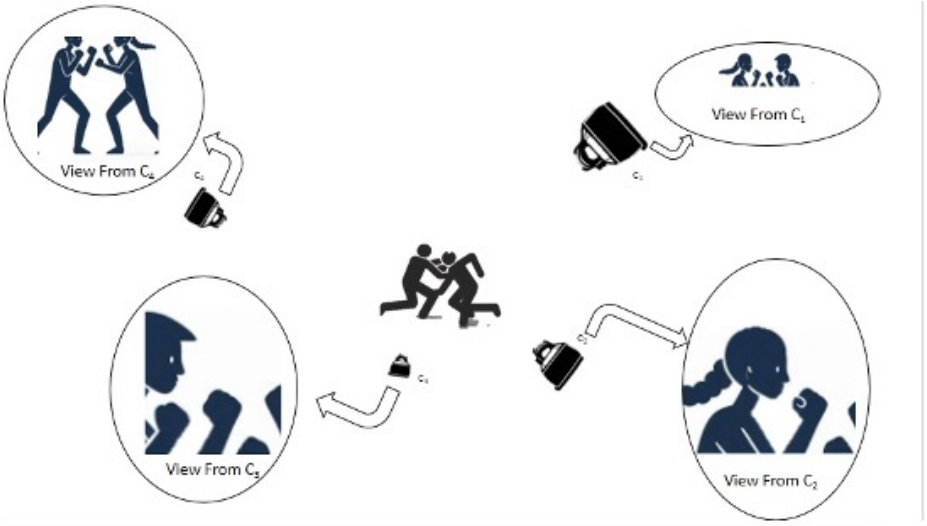

Fig. 5. demonstrates the significance of angle and viewpoint while capturing the same scene. To further highlight the cumulative effect of height, view-point, and distance; symbolic frames are given in ovals

Fig. 6. Snippet from surround view dataset in normal weather of the same scene. It clearly presents the effect of viewpoint, distance, height, and background variation while keeping the same core activity and subjects.

normality to abnormality or vice versa may lead to ambiguity and degradation in the performance of the model. To this end, we conducted voting for subjective assessment and labeling with the help of 10 volunteers. Fig. 3 presents snippets of abnormal scenarios from proposed RW-SVD under normal weather conditions. They highlight the diversity encompassed by the dataset in terms of Spatiotemporal information, the effect of viewpoint and distance, Illumination variation, Diverse motion patterns, subjects, etc. under normal weather conditions. Frame-level annotations as normal or abnormal are provided for each video. The dataset contains a total of 101 videos with a good balance between normal (49) and abnormal videos (52). The training set consists of 53 videos (24 normal and 29 abnormal) while the test set contains 47 videos (25 normal and 22 abnormal). Hence, we have maintained a balance between normal and abnormal data as well as training and testing data. In extending our dataset to incorporate weather noise we artificially add haze and rain effects in video frames. The addition of weather noise like haze and rain makes our dataset more practical for real-world scenarios where environmental noises are more frequent. We have thoroughly extended our previously proposed AnoVIL [23] dataset by incorporating the needs of recent computer vision applications that are sensitive to weather conditions like haze and rain. So, we have meticulously augmented AnoVIL [23] to depict frames in rainy and hazy weather [38–44]. We have introduced haze and rain with a procedural image generation method with randomness and geometric principles. Statistical Augmentation is performed to incorporate effects in visibility due to haze and rain as follows:

3.1 Haze generation model

For haze addition we utilize the given methods to incorporate haze.

$$I(x) = J(x) \cdot t(x) + A \cdot (1 - t(x))$$

Where: $I(x)$ stands for the observed hazy image, $J(x)$ stands for the scene radiance (haze-free image), $t(x)$ stands for the transmission map, A stands for the global atmospheric light.

Transmission Map (Haze Map) Generation The transmission map $H(x)$ is generated using a random gradient based on image coordinates, which is given as:

$$H(x) = \alpha_1 \cdot xv + \alpha_2 \cdot yv \tag{1}$$

Where: xv and yv are 2D gradient fields based on the image coordinates, α_1 and α_2 are random weights controlling the gradient's direction and intensity.

Fig. 7. heavy haze conditions contained in proposed dataset both in indoor and outdoor conditions in case of normal activities

Haze output The final hazy image is computed as follows:

$$I_{hazy}(x) = I(x) \cdot (1 - H(x)) + 255 \cdot H(x)$$

Where: 255 represents the white haze layer.
Haze output can be seen in Fig.-7,8

3.2 Rain Generation model

For rain generation, we introduce rain sticks by calculating their positions and orientations. Gaussian blurring is applied to enhance the raindrop streaks.

Raindrop statistics To generate a raindrop, we compute its endpoint based on its length and orientation angle:

$$x_{i+1} = x_i + L \cdot \cos(\theta)$$

$$y_{i+1} = y_i + L \cdot \sin(\theta)$$

Fig. 8. heavy haze conditions contained in proposed dataset both in indoor and outdoor conditions in case of 4 different types of abnormal activities with different background, scene density, and rate of motion

Fig. 9. Effect rain of in case of normal outdoor videos

Where: (x_{i+1}, y_{i+1}) is the endpoint of the raindrop, L is the length of the raindrop, θ is the angle of orientation.

Raindrop Streaks generation with Gaussian Blur To simulate the appearance of raindrops, we apply Gaussian blur, which is mathematically represented as:

$$G(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$$

Where σ is the standard deviation that controls the spread of the raindrop streaks.

Fig. 10. effect of rain in case of abnormal outdoor videos.

Rainy output The final rainy image is computed by blending the original image with the raindrop layer:

$$I_{rainy}(x) = I(x) \cdot (1 - \beta) + R(x) \cdot \beta$$

Where: $I(x)$ stands for the original image, $R(x)$ stands for the generated raindrop layer, β stands for the blending factor that controls the intensity of the rain effect.

Rainy output can be seen in Fig.-9,10

Videos were shot at 1080P however resolution was lowered to 320X240 to meet computation constraints. RW-SVD covers various routine activities like walking, playing, running, and playing as normal scenarios. Whereas, fighting, panic running, and snatching events are recorded as abnormal events. To tackle ambiguity in similar-looking actions, we have recorded the same scenario from 4 different perspectives. Frequently used objects such as bags, motorbikes, cycles, and cars are also included. Detailed frame-wise description is given in Table 1

4 Conclusion

We have developed a dataset that fills existing gaps such as the effect of weather; towards advancing the field. The proposed dataset focuses on human-centric anomalies through routine activities like walking, playing, and running as normal scenarios, and abnormal events such as fighting, panic running, kidnapping, and snatching. To address the challenges posed by varying viewpoints, the same scenarios are recorded from four different angles, ensuring comprehensive coverage of different perspectives. The dataset includes 49 normal and 52 abnormal

Table 1. Details of AnoVIL: No. of frames for 3 categories: normal weather, haze, and rain with Training, Testing set for Anomaly and Normal video frames are given below

Weather	Action	Training Frames	Testing Frames	Total
Normal Weather	Fighting	7090	4927	12017
	Kidnapping	1339	1014	2353
	Snatching	595	938	1533
	Panic Running	746	831	1577
	Normal	32855	60270	93125
Fogg	Fighting	7090	4927	12017
	Kidnapping	1339	1014	2353
	Snatching	595	938	1533
	Panic Running	746	831	1577
	Normal	32855	60270	93125
Rain	Fighting	6914	4837	11751
	Kidnapping	1339	1014	2353
	Snatching	595	938	1533
	Panic Running	746	831	1577
	Normal	29798	58213	88011
	Total	124642	201793	326435

videos, divided into training and evaluation sets, providing a robust benchmark for human-centric anomaly detection in video data.

References

1. Ke Xu, Tanfeng Sun, and Xinghao Jiang. Video anomaly detection and localization based on an adaptive intra-frame classification network. *IEEE Transactions on Multimedia*, 22(2):394–406, 2019. 1
2. Shuning Chang, Yanchao Li, Shengmei Shen, Jiashi Feng, and Zhiying Zhou. Contrastive attention for video anomaly detection. *IEEE Transactions on Multimedia*, 24:4067–4076, 2021. 1
3. M Zaigham Zaheer, Arif Mahmood, M Haris Khan, Mattia Segu, Fisher Yu, and Seung-Ik Lee. Generative cooperative learning for unsupervised video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14744–14754, 2022. 1
4. Anima Pramanik, Sobhan Sarkar, and J Maiti. A real-time video surveillance system for traffic pre-events detection. *Accident Analysis & Prevention*, 154:106019, 2021. 1
5. Kun Xia, Le Wang, Yichao Shen, Sanpin Zhou, Gang Hua, and Wei Tang. Exploring action centers for temporal action localization. *IEEE Transactions on Multimedia*, 25:9425–9436, 2023. 1
6. Yuanhao Zhai, Le Wang, Wei Tang, Qilin Zhang, Nanning Zheng, and Gang Hua. Action coherence network for weakly-supervised temporal action localization. *IEEE Transactions on Multimedia*, 24:1857–1870, 2021. 1
7. Iván Maza, Fernando Caballero, Jesús Capitán, José Ramiro Martínez-de Dios, and Aníbal Ollero. Experimental results in multi-uav coordination for disaster man-

agement and civil security applications. *Journal of intelligent & robotic systems*, 61:563–585, 2011. 1

8. Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. Ubnormal: New benchmark for supervised open-set video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20143–20153, 2022. 2
9. Kuldeep Biradar, Sachin Dube, and Santosh Kumar Vipparthi. Dearest: deep convolutional aberrant behavior detection in real-world scenarios. In *2018 IEEE 13th international conference on industrial and information systems (ICIIS)*, pages 163–167. IEEE, 2018. 2
10. Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):18–32, 2013. 2, 3
11. LiWeixin mahadevanV et al. Anomaly detectionincrowdedscenes. *IEEEconferenceoncomputerVisionandPatternRecog? nition (cVPR)*, 2010. 2, 3
12. Unusual crowd activity dataset of university of minnesota, 2006. http://mha.cs.umn.edu/proj_events.shtml/. Accessed: 2024-09-14. 2, 3
13. Amit Adam, Ehud Rivlin, Ilan Shimshoni, and Daviv Reinitz. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE transactions on pattern analysis and machine intelligence*, 30(3):555–560, 2008. 2, 4
14. Sachin Dube, Kuldeep Biradar, Santosh Kumar Vipparthi, and Dinesh Kumar Tyagi. Mag-net: A memory augmented generative framework for video anomaly detection using extrapolation. In *International Conference on Computer Vision and Image Processing*, pages 426–437. Springer, 2021. 2
15. Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. *Advances in neural information processing systems*, 15, 2002. 2
16. 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. 2, 6
17. Yidan Fan, Yongxin Yu, Wenhuan Lu, and Yahong Han. Weakly-supervised video anomaly detection with snippet anomalous attention. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 2
18. Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. Mist: Multiple instance self-training framework for video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14009–14018, 2021. 2
19. Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 322–339. Springer, 2020. 2, 7
20. Yifan Zhang, Jian Li, and Xia Liu. X-man: A dataset for anomaly detection in extreme weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1234–1243, 2022. 2, 7
21. Hongwei Yu, Yalin Li, Yiwei Ma, and Andrew Todd. The 4th ai city challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 229–239, 2020. 2, 7

22. Luma Akram Harbawee. *Artificial Intelligence Tools for Facial Expression Analysis.* PhD thesis, University of Exeter (United Kingdom), 2019. 2, 7
23. Kuldeep Marotirao Biradar, Murari Mandal, Sachin Dube, Santosh Kumar Vipparthi, and Dinesh Kumar Tyagi. Triplet-set feature proximity learning for video anomaly detection. *Image and Vision Computing*, 150:105205, 2024. 2, 10
24. Anna Ellis and James Ferryman. Pets2010: Dataset and challenge. *AVSS, 00 (undefined)*, pages 143–150, 2010. 3
25. Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pages 2720–2727, 2013. 4, 7
26. Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. 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. 4
27. Bruno Degardin and Hugo Proença. Human activity analysis: Iterative weak/self-supervised learning frameworks for detecting abnormal events. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–7. IEEE. 4
28. Liyun Zhu, Lei Wang, Arjun Raj, Tom Gedeon, and Chen Chen. Advancing video anomaly detection: A concise review and a new dataset, 2024. 4, 5
29. Mauricio Perez, Alex C Kot, and Anderson Rocha. Detection of real-world fights in surveillance videos. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2662–2666. IEEE, 2019. 4
30. Yu Yao, Xizi Wang, Mingze Xu, Zelin Pu, Yuchen Wang, Ella Atkins, and David Crandall. Dota: unsupervised detection of traffic anomaly in driving videos. *IEEE transactions on pattern analysis and machine intelligence*, 2022. 5
31. Z Che, G Li, T Li, B Jiang, X Shi, X Zhang, Y Lu, G Wu, Y Liu, and J Ye. D2-city: A large-scale dashcam video dataset of diverse traffic scenarios. arxiv 2019. *arXiv preprint arXiv:1904.01975*. 5
32. Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1207–1216, 2019. 5
33. Bharathkumar Ramachandra and Michael Jones. Street scene: A new dataset and evaluation protocol for video anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2569–2578, 2020. 5, 7
34. Enrique Chavez. Caviar (context aware vision using image-based active recognition) dataset. http://homepages.inf.ed.ac.uk/rbf/CAVIAR/. University of Edinburgh, School of Informatics. 6
35. Scott Blunsden and Bob Fisher. The behave video dataset: ground truthed video for multi-person behavior classification. *Annals of the BMVA*, 2010(4):1–11. 6
36. Xinyi Yao and Zhen Liu. Webdataset: A dataset for anomaly detection in web-based scenarios. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5625–5634. IEEE, 2021. 6
37. Silvio Savarese, Rogerio Feris, and Josef Sivic. Mit traffic dataset. Massachusetts Institute of Technology, 2009. Available at http://web.mit.edu/. 6
38. Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2341–2353, 2010. 10

39. Kshitiz Garg and Shree K Nayar. Vision and rain. *International Journal of Computer Vision*, 75:3–27, 2007. 10
40. Xueyang Fu, Jiabin Huang, Delu Zeng, Yue Huang, Xinghao Ding, and John Paisley. Removing rain from single images via a deep detail network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3855–3863, 2017. 10
41. Gaofeng Meng, Ying Wang, Jiangyong Duan, Shiming Xiang, and Chunhong Pan. Efficient image dehazing with boundary constraint and contextual regularization. In *Proceedings of the IEEE international conference on computer vision*, pages 617–624, 2013. 10
42. He Zhang and Vishal M Patel. Density-aware single image de-raining using a multi-stream dense network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 695–704, 2018. 10
43. Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing*, 28(1):492–505, 2018. 10
44. He Zhang, Vishwanath Sindagi, and Vishal M Patel. Image de-raining using a conditional generative adversarial network. *IEEE transactions on circuits and systems for video technology*, 30(11):3943–3956, 2019. 10