Multimedia Datasets for Anomaly Detection: A Review

Pratibha Kumari, Anterpreet Kaur Bedi and Mukesh Saini, Indian Institute of Technology Ropar, India

Abstract—Multimedia anomaly datasets play a crucial role in automated surveillance. They have a wide range of applications expanding from outlier objects/ situation detection to the detection of life-threatening events. For more than 1.5 decades, this field has attracted a lot of research attention, and as a result, more and more datasets dedicated to anomalous actions and object detection have been developed. Tapping these public anomaly datasets enable researchers to generate and compare various anomaly detection frameworks with the same input data. This paper presents a comprehensive survey on a variety of video, audio, as well as audio-visual datasets based on the application of anomaly detection. This survey aims to address the lack of a comprehensive comparison and analysis of multimedia public datasets based on anomaly detection. Also, it can assist researchers in selecting the best available dataset for benchmarking frameworks. Additionally, we discuss gaps in the existing dataset and insights for future direction towards developing multimodal anomaly detection datasets.

Index Terms—Anomaly datasets survey, Concept drift, Multimodal anomaly detection, Long term surveillance

I. INTRODUCTION

Anomaly detection corresponds to identifying unexpected or rare events in a dataset [1]. Detection of anomalous events from a scenario can offer important insights into a large number of monitoring and safety-critical real-world applications such as disaster forecast, detection of extreme climate event, mechanical fault, disease outbreak, fire/blast, fraud, etc. Recently, there has been a significant rise in the use of audio and video sensors for monitoring situations [2]–[4]. Monitoring of scenes by humans is unscalable and error-prone due to inherent human limitations [5], [6]. Therefore, we need to automatically analyse audio/video data to detect anomalous events and objects. Consequently, many frameworks for anomaly detection have been proposed. In order to evaluate and validate a framework, researchers require a comprehensive dataset. Towards this end, numerous datasets have been created in literature for anomaly detection. These are either based on a specific type of anomaly, or for generic anomaly detection.

In this survey paper, we present a comprehensive review on video, audio, and audiovisual datasets available for evaluating anomaly detection frameworks. We compare these datasets based on various perspectives, which will help researchers to choose suitable datasets for an application. To the best of our knowledge, there exists only one short survey paper [7] on video datasets published in 2016. We could not find any survey on audio or audio-visual anomaly datasets. We believe that a consolidated review of audio, video, and audio-visual datasets will promote use of multimedia for anomaly detection. Additionally, there are more than fifteen new datasets contributed to the video anomaly detection domain after the year 2016, which need to be reviewed.

A comprehensive taxonomy of anomaly datasets is given in Figure 1. Based on the recording media, anomaly datasets can be categorized into video, audio, and audio-visual datasets. Further, we categorize them on the basis of four attributes, viz., area of application, scene-type, anomaly induction mode, and labeling type. The characterization is expanded to include additional attributes in multiple figures and tables across Sections II, III, and IV. This enables a quick comparison of existing anomaly datasets, and thus helpful in selecting most suitable dataset to evaluate the given work. Further, to facilitate an easy access of datasets, we assemble them together and provide brief description including common information such as how and where the data was collected, what are the anomalies, merit/ demerit identified by researchers, etc., along with link to their website via supplementary file attached with this manuscript.

Rest of the paper is organized as follows. Public video datasets for anomaly detection are discussed in the Section II. A discussion about the audio datasets has been provided in Section III. Further, a comparison of public audio-visual datasets is given in Section IV. Section V discusses on future directions. Finally, we conclude the paper in Section VI.

II. VIDEO ANOMALY DATASETS

There are more than 30 publicly available video datasets that are currently being used for the purpose of anomaly detection. Initial datasets are comprised of simple events and scenarios with a very constrained amount of anomalies. The anomalies are performed by a group of actors and therefore lack the natural flow of events. They are of very short duration as well. The presence of a few rare objects or events in unimodal background events was regarded as anomaly in these datasets. Some examples of such datasets are Canoe [8] (a boat occurring once in the scene is regarded anomaly), UMN [9] (few people acting for sudden evacuation is regarded anomaly here), Web [9] (panic-escape and crowd fighting regarded as anomaly), Subway entrance/exit [10] (movement...
Later datasets consist of more types of anomalies and scene variability such as UCSD [11], AVENUE [12], ARENA [13], ShanghaiTech [14]. In contrast, some recently developed datasets are gigantic in terms of duration and variability of scene and events such as UCF-Crime [15], VIRAT [16], LV [17], ADOC [18], HTA [19], Rodriguez’s [20], etc. They leverage the online sources of CCTV footage and public videos and curate large databases of realistic video clips from our daily life. They are equipped with a variety of events and actions in with complex scene settings. Some datasets are suitable for generic scene monitoring, such as UCSD [11], ADOC [18], AVENUE [12], ShanghaiTech [14], etc., while others are suitable for specific classes of anomaly detection such as Subway entrance/exit [10], UMN [9], Web [9], i-Lids [21], etc. Further, few datasets such as UCF-Crime, LV, and NVIDIA AI CITY [22] contain specific anomaly classes that are important for usage in frameworks for detecting life-threatening events.

Table I presents a comparison of video anomaly datasets on various attributes of recorded videos such as continuity, duration, number of frames, number of video clips, resolution, frame rate, and camera motion. The datasets are listed in chronological order of their release year. Continuity (yes if untrimmed) information helps to see the applicability of a dataset for the task at hand. If the aim is to examine the context adaptation of anomaly detection framework, then the footage should be untrimmed as well as recorded at a single location [4], [6]. The adaptive models need sufficient data to learn the context on their own and adapt over time; hence a dataset with too short clips from different spatial and temporal aspects is not useful for evaluation. The approximate amount of duration is also specified in the table. Otherwise, the total number of clips is specified there. Some datasets, mainly scraped from online sources (YouTube, Pond5, Getty Images), have multiple resolutions; therefore, they need to be scaled on the same scale before processing. Also, for the same reason, these datasets may have multiple FPS. Some datasets have camera motion present too. For some of the datasets, motion is intentionally added to make the dataset more challenging and realistic (surveillance in moving trains such as Train [23], etc., surveillance with head-mounted cameras such as HTA [19], etc.). In the following subsections, we present a systematic comparison and remark on publicly available video datasets based on (1) area of application, (2) scene type, (3) anomaly induction mode, and (4) labelling procedure. We describe and compare video datasets through Figures 2 to 5, following the categorization basis mentioned in Fig. 1. At the end of this section, we provide a more detailed overview of each dataset in Section II-E with Tables II to IV.

A. Area of application

Based on the area of application, the video datasets can be broadly categorized into traffic monitoring and public monitoring applications. Some datasets have a mix of utility for both traffic as well as public monitoring; we keep them in the miscellaneous application category. This categorization is shown in Fig. 2. The datasets built for traffic monitoring application have only traffic-related anomalies and hence owns exclusive application only to traffic surveillance. This category has seven datasets, and all of them possess mainly safety-critical anomalies, e.g., accident, tire skidding, wrong turn, etc. On the other hand, the public monitoring application has diversity in the dataset recording location and can be used across indoor to outdoor public place monitoring, e.g., mall, railway station, park, subway, street, etc. Most of the available video anomaly datasets fall into this category. Some of these datasets only possess safety-critical anomalies such as blast, fight, robbery, gunshot, etc., while others possess non-safety-critical anomalies such as walking with a cart in a pedestrian area, throwing papers in the air, walking on grass, etc. Further, some datasets have both types of anomalies. Thus, the public monitoring category is further categorized into ‘security-threat’, ‘non-security-threat’, and ‘both’ sub-categories, based on the criticalness of present anomalies.
TABLE I

| Dataset               | Continuity | Total Duration | Total No. of Frames | Total No. of Videos | Resolution | FPS | Camera motion |
|-----------------------|------------|----------------|---------------------|---------------------|------------|-----|---------------|
| i-Lids (2007)         | no         | ≈24 min        | 35000               | 7                   | 72x576     | 25  | none          |
| QMUL (2008)           | no         | 22 min         | 34000               | 104(N)+8(A)         | 360x288    | 25  | none          |
| Canoe (2008)          | yes        | 34 s           | 1050                | 1                   | 320 x 240  | 30  | none          |
| Subway Entrance (2008)| yes        | 96 m 9 s       | 144225              | 1                   | 512x384    | 25  | none          |
| Subway Exit (2008)    | yes        | 43 m 16 s      | 64900               | 1                   | 512x384    | 25  | none          |
| UMN (2009)            | no         | 4 min 17 sec   | 7710                | 11 (available as 1)| 320x240    | 30  | none          |
| Web (2009)            | no         | 7 min 35 sec   | 11,962              | 12(N)+8(A)          | multiple   | multiple slight jerks |
| PETS2009 (2009)       | no         | ≈1-2 hrs       | 42182               | 59                  | 720x576    | 7   | none          |
| U-Turn (2009)         | no         | ≈20 min        | 25182               | 8                   | 360 x 240  | multiple | none          |
| Idiap (2009)          | yes        | 44.13 min      | 66324               | 1                   | 288x360    | 25  | none          |
| UCSD Ped1 (2010)      | no         | ≈5-7 min       | 14000               | 34(N)+36(AN)        | 238x158    | -   | none          |
| UCSD Ped2 (2010)      | no         | ≈2-3 min       | 4500                | 16(N)+14(AN)        | 360x240    | -   | none          |
| Train (2010)          | yes        | 12 min         | 19218               | 1                   | 288x386    | 25  | moving        |
| Belleview (2010)      | no         | 4 min 51 sec   | 2918                | 1                   | 320 x 240  | 10  | jitter        |
| Boat-Sea (2010)       | yes        | 1 min 56 sec   | 450                 | 1                   | 720x576    | 19  | none          |
| Boat-River (2010)     | yes        | 1 min 8 sec    | 250                 | 1                   | 704x576    | 5   | none          |
| Caoufleau (2010)      | yes        | 54 sec         | 1629                | 1                   | 320x240    | 29.97 | none          |
| Rodriguez’s           | no         | 10 hrs 24 min  | -                   | 520                 | 720x480    | -   | none          |
| Hockey (2011)         | no         | 27 min         | 50000               | 1000                | 720x576    | -   | none          |
| Movie (2011)          | no         | 6 min          | -                   | 200                 | multiple   | multiple in some |
| UCF Crowd (2012)      | no         | ≈11 min        | 16320               | 38                  | multiple   | multiple in some |
| WeddingFloors (2012)  | no         | 14 min 6 sec   | 22,156              | 246                 | 320x240    | multiple | significant |
| Grand Central Station (2012)| no | 33 min 20 sec | 50010               | 1                   | 480x720    | 25  | none          |
| AGORASET (2012)       | no         | ≈20 min        | >33641              | 23                  | 640x480    | 30  | none          |
| Meta-tracking (2013)  | no         | –              | >4000               | 12                  | multiple   | multiple in some |
| Avenue (2013)         | no         | 20 min 26 sec  | 30652               | 16(N)+21(AN)        | 640x360    | 25  | slight in few |
| ARENA (2014)          | no         | –              | –                   | 22                  | 1280x960   | 30  | none          |
| PFWFD (2015)          | yes        | 1 hour         | 5000                | 1                   | 1920x1080  | 1.25 | none          |
| RoadID (2015)         | no         | <2 hrs         | -                   | 30                  | 1280x720   | multiple | in some |
| MED (2016)            | no         | ≈24-25 min     | 43,626              | 31                  | 554x225    | 31  | none          |
| ShanghaiTech (2017)   | no         | 3-3.5 hrs      | 317398              | 330(N)+107(AN)      | 856x480    | 24  | none          |
| LV (2017)             | no         | 3.93 hrs       | -                   | 30                  | multiple   | multiple in some |
| UCF-Crime (2018)      | no         | ≈128 hrs       | 13 million          | 950(N)+950(A)       | 320x240    | 30  | slight in few |
| IITH Accidents (2018) | no         | –              | 128001              | –                   | multiple   | multiple in some |
| CCTV-Fights (2019)    | no         | ≈10 hrs        | 1000                | –                   | multiple   | multiple in some |
| Street Scene (2020)   | no         | 3-4 h          | 203257              | 46(N)+35(AN)        | 120x240    | 15  | none          |
| ADDC (2020)           | yes        | 24 hrs         | 259127              | 4                   | 2048x1536  | 3   | none          |
| HTA (2020)            | no         | ≈4 hrs         | 0.4 million         | 286(N)+107(AN)      | 1280x720   | 30  | moving        |

Some datasets can be used for both traffic as well as public monitoring as they possess anomalies for both categories. We list such datasets under the miscellaneous category. The datasets in each category are listed in order of the count of distinct safety-critical anomalies present in them. A dataset that offers a larger number of distinct safety-critical anomalies is mentioned first.

B. Scene-type

Some of the video anomaly datasets are recorded by placing the recording device at one place, while others collect video data from multiple places. Therefore, based on the dataset recording scene, there can be two categories of datasets: single-scene and multi-scene. They can be further categorized based on whether the scene is indoor or outdoor. A dataset recorded indoor or outdoor may help to evaluate the robustness or claim that the framework works in indoor and/or outdoor scenarios. We show the taxonomy for scene-based video datasets categorization via Fig 3. For datasets falling in the multi-scene category, we mention the exact number of locations in parenthesis along with the dataset if provided by the authors of the datasets. Also, the datasets in each subcategory are listed in order of crowd density present in the dataset. Some frameworks are specifically designed to handle crowded scenes, whereas others work only for low density of people; thus, the information of density in a dataset may help to choose suitable datasets for evaluations.

In the multi-scene category, the datasets mainly possess non-contextual anomalies, which means if an event is anomalous in one scene, it will be anomalous in other scenes as well. For example, fighting with hands is an anomaly for a boxing game scene as well as in a park scene. Therefore, datasets falling in the multi-scene category largely ignore the context. On the other hand, in the case of single-scene datasets, events are regarded as anomalies based on their frequency. An event that is rare for the specific scene is regarded anomaly for that scene only. If the aim is to develop a generic scene monitoring, then single-scene datasets are useful, whereas, for detecting specific anomaly events, multi-scene datasets are suitable.

C. Anomaly induction mode

Based on the induction mode of anomalies, the datasets can be categorized into four groups, viz., acted, natural, acted
Fig. 2. Categorization of video anomaly datasets based on area of application

Fig. 3. Categorization of video anomaly datasets based on scene-type

as well as natural, and artificial, as shown in Fig. 4. Since anomaly is a rare phenomenon, it will require a longer waiting time to get the anomalous samples/ events; consequently, researchers have also tried inducing anomalies manually with the help of actors or artificially with the help of computer software. However, the acted events add unrealistic flavor, and hence less complex anomaly events are generated. In Fig. 4, we also mention the distinct number of anomalies present in each dataset. Further, if the dataset has acted anomalies too, the number of actors involved in inducing anomalies are also reported in parenthesis.

Generally, for traffic surveillance-related datasets such as HTA [19], QMUL [24], Idiap [25], IITH Accidents [26], etc., getting natural anomaly samples is easier as the anomalies in these datasets, which include accidents, illegal turns, or other traffic rule violations, occur a bit frequently. Also, safety-critical events in public monitoring such as fights, blasts, robbery, gunshots are easier to collect from CCTV footage, movies, or YouTube. Hence, datasets such as UCF-Crime [15], LV [17], CCTV-Fights [27], Movie [28], etc., have natural anomalies. On the other hand, datasets having fewer safety-critical anomalies are built by inducing acted events too. Some of these include ARENA [13], Avenue [12], MED [29], ShanghaiTech [14], etc.

D. Labelling procedure

The existing video anomaly datasets have diversity in terms of the type of labeling too. We categorize them based on the available annotation type, as shown in Fig. 5. Some datasets give only clip-level annotation; if a clip is annotated anoma-
ous, then it is assumed that anomaly is present or spanned to all the frames in the clip. The assumptions hold true in most cases, especially for small databases, but it is not scalable to validate this assumption for larger datasets. Hence, some frames may be wrongly annotated. Frame-level annotations are more precise and accurate [30]. For some datasets, pixel-level annotations are also provided, which are helpful for models which aim to precisely localize the occurrence of an anomaly in the spatial domain along with the temporal domain. Pixel-wise annotations are too precise and hence difficult to do manually; therefore, some datasets such as Train [23], Belleview [23], Boat-Sea [23], BoatRiver [23], Airport-WrongDir [23], etc. first divide the frame into grids (a grid can have multiple pixels) and then annotate the grids that can be served for spatial anomaly localization. Some also provide the rectangle bounding boxes around the area of interest, i.e., anomalous object/person. Some datasets, especially for people tracking crowd datasets, e.g., PWPD [31], meta-tracking [32], etc., provide the trajectories and anomaly information on these.

### E. Dataset Overview

A brief introduction of all the publicly available video anomaly datasets is given through Tables II to IV. The datasets are listed in chronological order of their release year. Some examples of anomalies present and the recording location of the dataset are also mentioned. Information about dataset collection scene, i.e., whether it is a mall or an airport or a traffic junction, etc., may help to choose scene-specific dataset for the evaluation of scene-specific surveillance frameworks. The table also mentions the area of application which gives an idea of different applications where the dataset has been used for benchmarking. Visuals from the dataset, i.e., a sample normal and an abnormal frame from the datasets, are also shown.

There are some other datasets that are not primarily developed for anomaly detection but other sub-tasks of scene monitoring such as detecting group formation, tracking people, counting people, human interaction/action classification, etc. We discuss them under the additional video dataset category in supplementary material attached with this survey, as they can also be utilized for generic scene surveillance or can be
| Dataset                  | Anomalies: collection scenario                                                                 | Applications                                                                 | Normal image example | Anomaly image example |
|-------------------------|------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------|----------------------|-----------------------|
| i-Lids [21] (2007)      | abandon bag, illegally parked vehicle: at railway station, road                               | abandon baggage detection [33], traffic surveillance [34]                   |                      |                       |
| QMUL [24] (2008)        | unusual traffic trajectory, rare behaviour of vehicles: at road (traffic junction)            | vehicle tracking [35], trajectory classification [36], anomaly detection [37]–[39] |                      |                       |
| Canoe [8] (2008)        | canoe: in river                                                                                | anomaly detection [23], [40]                                                |                      |                       |
| Subway Entrance [10](2008) | wrong direction: at subway entrance                                                          | Anomaly detection [41], [42], abnormal behavior modeling [43], [44]          |                      |                       |
| Subway Exit [10](2008)  | wrong direction: at subway exit                                                               | Anomaly detection [41], [42], abnormal behavior modeling [43], [44]          |                      |                       |
| UMN [9] (2009)          | run: in field, courtyard, hallway                                                             | anomaly detection [11], [41], abnormal crowd behaviour [45], [46], crowd aggregation detection [47], crowd escape detection [48] |                      |                       |
| Web [9] (2009)          | panic-escape, clashing/fighting: at multiple location                                          | anomaly detection [49]                                                      |                      |                       |
| PETS2009 [50] (2009)    | run, panic: in university                                                                     | tracking [51], [52], crowd profiling/counting [53], [54], crowd analysis [46], human detection [55], [56], person re-identification [57], crowd escape behaviour detection [48] |                      |                       |
| U-Turn [58] (2009)      | illegal u-turns, running, abandon baggage: at road (intersection), university                 | traffic based anomaly detection [59]–[61]                                 |                      |                       |
| Idiap [25] (2009)       | wrong road crossing, wrong vehicle parking, etc.: at road                                     | anomaly detection [38], [39], recurrent activity mining [62]                |                      |                       |
| USCD Ped1 [11] (2010)   | bikers, small carts, walking across walkways: at campus (walkway at UCSD)                    | crowd profiling/counting [53], [63], anomaly detection [64], [65], crowd density estimation [66] |                      |                       |
| USCD Ped2 [11] (2010)   | bikers, small carts, walking across walkways: at campus (walkway at UCSD)                    | anomaly detection [67], [68], crowd profiling/counting [53], action classification [69] |                      |                       |
TABLE III
CONTINUED FROM TABLE II: VIDEO ANOMALY DATASETS: PRIMARY INFORMATION

| Dataset               | Anomalies: collection scenario | Applications                                      | Normal image example | Anomaly image example |
|-----------------------|--------------------------------|---------------------------------------------------|----------------------|-----------------------|
| Train [23] (2010)     | people movement: inside train  | anomaly detection [40], [68], [70]                |                      |                       |
| Bellevue [23] (2010)  | illegal turns: at road (intersection) | anomaly detection [40], [70]                  |                      |                       |
| Boat-Sea [23] (2010)  | boat: in sea                   | anomaly detection [40], [71]                     |                      |                       |
| Boat-River [23] (2010) | boat: in river                 | anomaly detection [40], [44]                     |                      |                       |
| Caouflage [23] (2010) | person in Caouflage: in room   | anomaly detection [44]                           |                      |                       |
| Airport-WrongDir [23] (2010) | movement in wrong direction: at security check-point | anomaly detection [44], [72]             |                      |                       |
| Rodriguez’s [20] (2011) | anomalous trajectory: from multiple scenarios | crowd profiling/counting [54], crowd saliency detection [73], tracking [52], crowd analysis [74], crowd segmentation [75] |                      |                       |
| Hockey [28] (2011)    | fight: at ice hockey game      | violence detection [76]–[78]                     | available upon request |                       |
| Movie [28] (2011)     | fight: in movie clips          | violence detection [77]–[80]                     | available upon request |                       |
| UCF Crowd [81] (2012) | anomalous trajectory: from multiple scenarios | abnormal crowd detection [82], crowd profiling/counting [54], crowd saliency detection [73], crowd segmentation [83] |                      |                       |
| Violent-Flows [84] (2012) | crowd violence: at multiple scene | abnormal crowd behaviour detection [45], violence detection [77], [79] |                      |                       |
| Grand Central Station [85] (2012) | rare walking pattern: at terminal station | pedestrian trajectory prediction [86], tracking [35], [87], pedestrian behavior modeling [88], crowd counting [63], [89], crowd behavior analysis [90], [91], human re-identification [92] | trajectory based |                       |
| AGORASET [93] (2012)  | evacuation, dispersion: at multiple scene | panic behaviour detection [94], crowd flow tracking [95], crowd motion classification [96], crowd behaviour analysis [97] |                      |                       |
## TABLE IV

**CONTINUED FROM TABLE III: VIDEO ANOMALY DATASETS: PRIMARY INFORMATION**

| Dataset         | Anomalies: collection scenario                      | Applications                                                                 | Normal image example | Anomaly image example |
|-----------------|-----------------------------------------------------|------------------------------------------------------------------------------|----------------------|-----------------------|
| Meta-tracking   | anomalous trajectory: at multiple location          | tracking [98], [99], anomalous walking pattern detection [98] measuring collectiveness [99], crowd segmentation [100] |                      |                      |
| Avenue [12]     | loitering, running, throwing objects, new object: in campus (avenue) | anomaly detection [101], [102]                                               |                      |                      |
| ARENA [13]      | abnormal behaviour, threats: in campus (university of Reading) | abnormal activity/behaviour detection [103]–[105], group walking event detection [106], Human Full-Body/Body-Parts detection and tracking [107] |                      |                      |
| PWPD [31]       | anomalous trajectory, abnormal crowd: at terminal station | trajectory prediction [108], anomaly detection [108], pedestrian speed detection [109], crowd behaviour analysis [110] |                      |                      |
| RE-DID [111]    | fight: from multiple vehicle dash-cams             | fight detection [111]                                                        |                      |                      |
| MED [29]        | Panic, fight, congestion, obstacle, neutral: at campus (walkway) | anomaly detection [112], panic detection [113]                              |                      |                      |
| ShanghaiTech [14] | bicycle, small carts, fight, etc.: at campus (ShanghaiTech) | anomaly detection [101], [114], action classification [69], crowd counting [115] |                      |                      |
| LV [17]         | realistic security threats: at multiple location    | anomaly detection [116]–[118], panic detection [119]                        |                      |                      |
| UCF-Crime [15]  | abuse, arrest, assault, accident, burglary, etc.: at multiple location | anomaly detection [120], [121], action classification [69]                  |                      |                      |
| IITH Accidents  | accidents: at road (intersection, junction)         | road accident detection [26]                                                |                      |                      |
| CCTV-Fights [27] | fight: at multiple location                         | fight detection [122]                                                        |                      |                      |
| Street Scene [30] | jaywalking across road, pedestrians loitering, u-turns: on road (two-lane street) | anomaly detection [123], road surveillance [124]                           |                      |                      |
| ADOC [18]       | walking with balloons/dog, person on vehicle, crowd gathering, etc.: at campus (walkway) | anomaly detection [123]                                                      |                      |                      |
| HTA [19]        | accident, speeding vehicle, close merge: at multiple location | anomalous motion detection [19]                                             |                      |                      |
useful in creating some giant anomaly datasets.

### III. AUDIO ANOMALY DATASETS

Audio anomaly detection is gaining a lot of attention in many applications in various fields, such as traffic surveillance, industries, music, medicines, etc. Unlike for video, not many audio anomaly datasets are available for study. A few datasets have been created based on applications and further provided publicly. In this section, we describe and discuss the features of publicly available audio datasets. Table V gives the specifications of the audio samples collected for different datasets. The specifications include time duration of each dataset, number of samples collected, and sampling rate for audio-digitization. It can be observed from the table that the duration of audio for different datasets varies from a few seconds to a number of hours. AudioSet soundscape dataset has the longest duration of audios, followed by ToyADMOS2 [125] and ToyADMOS [126] datasets. A very small recording of 550 seconds has been collected for the LIFE DYNAMAP project [127].

Further, from the table, it can be seen that the sampling rate for different audio datasets varies between the range of 4kHz to 48 kHz. The sampling rate of 48 kHz has been mostly considered for audio digitization in various datasets. Also, for ICBHI 2017 [128], there is no fixed value of sampling rate for the data; rather, it varies between the range of 4 to 44.1 kHz. Thus, different sampling rates have been used as per the requirement of the tasks.

Dataset classification based on various attribute as mentioned in Fig. 1 is discussed in this section. Unlike in video datasets, since only segment level labelling is followed for the audio datasets, hence, a separate classification of audio datasets based on labelling has not been provided.

#### A. Area of application

Audio anomaly datasets cover a variety of applications such as monitoring of traffic, machine, public, patient, etc. The area of application based classification is shown in Fig. 6. Majority of the audio anomaly datasets have been developed for machine monitoring, whereas a few datasets are available for monitoring traffic and public. Heath supervision is also made possible using ICBHI [128] dataset by monitoring respiratory patterns in the patients.

#### B. Scene-type

Scene based categorization for audio datasets is shown via Fig. 7. Owing to different recording locations for each dataset, the number of multi-scene datasets is much higher and more easily available as compared to those collected from one location, such as ICBHI [128] and SMD [129]. The single-scene datasets have been collected in indoor scenarios, whereas multi-scene datasets have been created at various indoor and outdoor locations such as industry, road, forests, etc.

#### C. Anomaly induction mode

Fig. 8 shows anomaly induction mode based categorization of datasets. The count of distinct anomalies present in the dataset is also listed along with. As it can be seen from the
### TABLE V

**SPECIFICATIONS OF AUDIO ANOMALY DATASETS**

| Dataset                                                                 | Continuity | Total duration | No. of samples | Sampling rate |
|------------------------------------------------------------------------|------------|----------------|----------------|---------------|
| MIVIA (2014)                                                           | no         | 3580 seconds   | -              | 32 kHz        |
| AudioSet Soundscape: Ithaca, New York                                   | no         | #1: 797 hrs    | -              | 48 kHz        |
|                                                                       |            | #2: 638 hrs    |                | 48 kHz        |
| AudioSet Soundscape: Sabah, Malaysia                                    | no         | Tuscan: 27 hrs 40 min | -  | 44.1 kHz  |
|                                                                       |            | Audiomoth: 784 hrs |     | 16 kHz     |
| AudioSet Soundscape: New Zealand (2015)                                 | no         | 240 hrs        | -              | 32 kHz        |
| AudioSet Soundscape: Sulawesi                                          | no         | 64 hrs         | -              | 48 kHz        |
| AudioSet Soundscape: Republic of Congo                                  | no         | 238 hrs 20 min | -              | 8 kHz         |
| DCASE 2017 (2017)                                                       | no         | 39 min         | 1170           | 44.1 kHz      |
| ICBHI 2017 (2017)                                                       | no         | > 5.5 hrs      | 2048           | 4-44.1 kHz    |
| SMD (2018)                                                              | no         | ≈ 1 hour       |                | 16.3 kHz      |
| MIMII DUE (2019)                                                        | no         | >420000 sec    | 32157          | 16 kHz        |
| ToyADmos (2019)                                                         | no         | ≈ 540 hrs      | > 12000        | 48 kHz        |
| LIFE DYNAMAP (2020)                                                     | no         | 550 sec        |                | 48 kHz        |
| ToyADMos2 (2021)                                                        | no         | 604 hrs        | ≈264k          | 48 kHz        |
| DCASE 2021 (2021)                                                       | no         | ≈82 hours      | 29463          | 16 kHz        |

**Fig. 6.** Categorization of audio anomaly datasets based on area of application

**Fig. 7.** Categorization of audio anomaly datasets based on scene-type

**Fig. 8.** Categorization of audio anomaly datasets based on anomaly induction mode

So far, we have discussed a number of publicly available datasets for video and audio anomaly detection applications individually. Since a course of actions consists of both audio

**D. Dataset Overview**

Table VI list all the publicly available audio anomaly datasets in chronological order of their release year. Some of the anomalies that have been generated are mentioned for every dataset in the table. The datasets have been used for applications in different fields, such as in industries for machine monitoring and inspection, road surveillance, etc. A dataset for anomaly detection in respiratory patterns in human bodies has also been created. Most of the datasets have been artificially generated by deliberately adding anomalies and made available to the public recently.

**IV. AUDIO-VISUAL DATASETS**

So far, we have discussed a number of publicly available datasets for video and audio anomaly detection applications individually. Since a course of actions consists of both audio...
TABLE VI
AUDIO ANOMALY DATASETS: PRIMARY INFORMATION

| Dataset | Anomalies: collection scenario | Application |
|---------|-------------------------------|-------------|
| MIVIA [130] (2014) | crashes and tire skidding: on road | road surveillance [131]–[133] |
| AudioSet Soundscape [134] (2015) | music, human speech, machine noise, etc.: from nature | anomaly detection from eco-acoustic data [134] |
| DCASE 2017 [135] (2017) | baby cry, glass break, gun shot: city | detection of rare sounds [136], [137] |
| ICBHI 2017 [128] (2017) | crackle, wheeze, crackle and wheeze: in human body | anomalous respiratory pattern detection in humans [138]–[140] |
| SMD [129] (2018) | non-greased line, B-line, C-line to B-line: from industry | machine monitoring in industries [129] |
| MIMII DUE [141] (2019) | wing damage, clogging, gear, contamination etc.: from industry | investigation and inspection of industrial machine |
| ToyADMOS [126] (2019) | deformed gears, over/under voltage, pulley, chipped wheel, axle, excessive tension, etc.: from industry | anomalous sound detection in miniature machines [142], [143] |
| LIFE DYNAMAP [127] (2020) | horns, church bells, birds, thunder etc.: on road | detection and removal of anomalous events for road traffic mapping [144], [145] |
| ToyADMOS2 [125] (2021) | bent shaft, melted gears, flat tire etc.: from industry | anomalous sound detection in miniature machines |
| DCASE 2021 [146] (2021) | wing damage, clogging, chipped wheel, axle etc.: from industry | detection of anomalous sounds during machine monitoring [147]–[149] |

TABLE VII
SPECIFICATIONS OF AUDIO-VISUAL ANOMALY DATASETS

| Dataset | Continuity | Total Duration | Total No. of Frames | Total No. of Videos | Video resolution | FPS (video) | Camera motion |
|---------|------------|----------------|---------------------|---------------------|------------------|------------|--------------|
| Human-human interaction (2014) | no | 32.24 min | - | 8 | - | - | none |
| VSD (2015) | no | 35 hrs 18 min | - | 25 | multiple | multiple | in some |
| EMOLY (2018) | no | - | - | 123 | multiple | - | in some |
| XD-Violence (2020) | no | 217 hrs | - | 4754 | multiple | multiple | none |
| BAREM (2021) | no | 6 hrs | ≈ 5,40,000 | 72 | - | 25 | none |

and video components dependent on each other, hence, audio-visual analysis can result in more accurate results compared to audio and video recordings being used individually [4], [150]. There are very few audio-visual datasets for anomaly detection. They are mainly developed for applications such as detection of anomalous expression, violence, stress, etc.

Table VII provides details about the features and specifications of the audio-visual clips. The table enables a quick comparison of datasets based on their total frame count, resolution, duration, etc. Mainly, information about the total number of videos and total duration is made available by the authors. The datasets for violence detection records the largest length of 217 hours compared to others. This is because violence event is comparatively more often to occur and hence easily available in CCTV footage. The total number of videos available for XD-Violence [151] is maximum, i.e., 4,754, followed by the EMOLY [152] dataset consisting of 123 videos. Least number of videos have been provided in the Human-human interaction dataset, with a total duration of 32.24 minutes. Clips have been recorded at multiple resolutions and at multiple frame rates. However, the sampling rate of audios recorded in the databases has not been made available by the authors. Few clips from VSD [153] as well as XD-Violence [151] datasets consist of camera motion, thus making them challenging for further analysis.

Also, it can be observed from the table that none of the datasets is collected in continuation and hence are not suitable for evaluation of adaptive anomaly detection frameworks. On the same categorization basis as for audio datasets, we present and discuss the categorization of audio-visual datasets through Fig. 9 to Fig. 11. Unlike in video datasets, since only frame level labelling is followed for the audio-visual datasets, hence, a separate class of audio-visual datasets based on labelling has not been provided. Since all the datasets possess frame-level annotations, hence, anomaly localization is not yet explored in audio-visual datasets.

A. Area of application
The audio-visual datasets have been designed and made available to the users recently. They all fall into public monitoring applications. The categorization is shown in Fig. 9. Different sub-applications of public monitoring, such as anom-
lous expression detection, violence detection, detection of stress in a particular situation, etc., have been covered for these datasets. Based on whether the anomalies possess security threats or not, they are further categorized into two groups viz., security-threat and non security-threat. VSD [153] and XD-Violence [151] datasets have security threat-based anomalies, and the rest datasets do not possess any security threat-based anomalies.

**C. Anomaly induction mode**

Categorization based on anomaly induction mode, i.e., natural or acted, is shown via Fig. 11. All the datasets contain some acted anomalies too. The number of actors in the datasets is also mentioned in parenthesis. Each dataset has only one type of distinct anomaly except for the violence detection datasets, i.e., VSD [153] and XD-Violence [151]. VSD [153] consists of 8 distinct anomalies, while XD-Violence [151] contains 6 distinct anomalies. Some of the example anomalies in audio-visual datasets include fights, stress, explosions, abuse, car accident, shooting, frustration, etc.

**D. Dataset Overview**

For a quick analysis of all the publicly available audio-visual anomaly datasets, we list them in increasing order of their release year via Table VIII. The table has other information, viz., where it was collected, what anomalies are in them, applications where it has been used for benchmarking, a normal sample image, and an abnormal sample image. These datasets mainly focus on the specific type of anomaly. VSD [153] dataset, XD-Violence [151] dataset, etc., consist of only fight events, whereas EMOLY [152] dataset has an anomalous mood of a person as an anomaly. Human-human interaction dataset has agitation behavior as an anomaly. Apart from having one class of anomaly, the datasets are mainly recorded in a controlled environment (except for fight anomaly, which is easy to collect) by some human actors. The existing audio-visual dataset is good for application-specific anomaly detection; however, there is a lack of audio-visual datasets for generic scene surveillance. Some researchers have tried collecting such datasets, but they are not released publicly due to privacy or legal issues.

**V. DISCUSSION: TOWARDS THE FUTURE**

During the last decade, there has been a drastic shift from datasets with less number of anomaly samples and total duration to those with diverse range of anomalies and gigantic volume. It may be observed that the availability of datasets with crime-specific anomalies has facilitated the development of automated surveillance frameworks for crime detection. For videos, the category spans detection of the explosion, abuse, panic escape, assault, accident, violence, fight, etc. In the case of audio, they span as detection of the gunshot, shout, etc. Further, for the audio-visual dataset, the categories span as detection of riots, fight, and violence.
### TABLE VIII

**AUDIO-VISUAL ANOMALY DATASETS: PRIMARY INFORMATION**

| Dataset                  | Anomalies: collection scenario | Application                                                                 | Anomaly image example | Normal image example |
|--------------------------|--------------------------------|-----------------------------------------------------------------------------|-----------------------|---------------------|
| Human-human interaction [154] (2014) | stress: at help-desk          | Detection of stressful situations at a help-desk [156]                      | ![Image]              | ![Image]            |
| VSD [153] (2015)         | fights, fire, gunshot, cold weapons, car chases, gory, explosion, screams: at multiple location | violence detection [80], [157], [158]                                     | ![Image]              | ![Image]            |
| EMOLY [152] (2018)       | anomalous expression: in lab  | abnormal expression detection [152]                                         | ![Image]              | ![Image]            |
| XD-Violence [151] (2020) | abuse, car accident explosion, fight, riot, shoot: at multiple location | violence detection [159], anomaly detection [160]                         | ![Image]              | ![Image]            |
| BAREM [155] (2021)       | frustration: on e-service platform | Behaviour Analysis for Reverse Efficient Modeling [155]                      | Available upon request | ![Image]            |

By analyzing the applications and anomalies present in the datasets discussed across Sections II to IV, we can see that they have mainly specific types of anomalies. There are a very few datasets of heterogeneous nature. This is almost zero in the audio-visual category. Thus, existing datasets are useful for specific anomaly detection, particularly crime-oriented anomaly detection, traffic rule violation, etc. However, for future applications like smart city surveillance we need generic scene monitoring where we have to raise an alarm for interesting events, which may or may not be crime-oriented. This shows a strong need to develop datasets having diverse ranges of anomalous samples. Also, the datasets in the audio-visual category mainly contain acted anomalies, which limits its usefulness. Even if some acted situation/events need to be added, it should be in such a manner that it appears natural and contains the necessary amount of variations mimicking real life.

Apart from this, all the existing multimedia datasets lack anomalies with concept drift [6]. If an event/object is regarded anomaly in a dataset, it is always regarded as an anomaly for that scene, no matter how frequently it may occur in the distant future. This is due to the fact that datasets are too short to contain this effect. We should pay attention that the datasets which are more than a duration of 5-10 hours are not suitable here because the samples are collected from different time-stamp, location, and have only specific anomalies in rare amounts. Thus, interclass shift, i.e., abnormal to normal class and vice-versa, is not observed. There are attempts to record long untrimmed footage at one place, e.g., QMUL [24], ADOC [18], etc. However, the authors do not attempt to provide annotations in accordance with concept drift.

### VI. CONCLUSION

This paper presents a survey of multimedia datasets for anomaly detection to researchers working towards automated surveillance. The structured comparison of datasets on various attributes also helps to understand datasets better. There are a large number of short-length and giant video datasets available. Some are developed for generic scene surveillance, whereas other are specific anomaly datasets. Datasets for heterogeneous anomaly are far less compared to specific anomaly datasets. In case of audio, datasets are mainly developed for machine surveillance such as defect detection, fault detection, etc. Generally, surveillance using audio alone in outdoor scenarios is not efficient due to the presence of multiple auditory signals superimposed together. However, when analyzed together with video, they can offer crucial and complementary information about the target scene. However, the datasets developed towards audio-visual surveillance are far less compared to that for audio or video. There are a few audio-visual datasets for a specific action, such as fight and agitation detection. A recently released dataset, viz., the EMOLY dataset, has used only the upper body of individuals and their speech information to facilitate abnormal behavior detection. However, there is a
strong need to develop more audio-visual datasets for generic scene surveillance. We believe the survey presented in this article will help the prospective researchers who intend to contribute datasets or research in this field.

REFERENCES

[1] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” ACM computing surveys, vol. 41, no. 3, pp. 1–58, 2009.

[2] N. Jaafar and Z. Lachiri, “Audio-visual fusion for aggression detection using deep neural networks,” in 2019 International Conference on Control, Automation and Diagnosis (ICCAD). IEEE, 2019, pp. 1–5.

[3] A.-U. Rehman, H. S. Ullah, H. Farooq, M. S. Khan, T. Mahmood, and H. O. A. Khan, “Multi-modal anomaly detection by using audio and visual cues,” IEEE Access, vol. 9, pp. 30587–30603, 2021.

[4] P. Kumar, “Situational anomaly detection in multimedia data under concept drift,” in Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 2960–2973.

[5] O. Elharrouss, N. Almadaed, and S. Al-Maadeed, “A review of video surveillance systems,” Journal of Visual Communication and Image Representation, vol. 77, p. 103116, 2021.

[6] P. Kumar and M. Saini, “Multivariate adaptive Gaussian mixture for scene level anomaly modeling,” in BigMM. New Delhi, India: IEEE, 2020, pp. 54–62.

[7] N. Patil and P. K. Biswas, “A survey of video datasets for anomaly detection in automated surveillance,” in ISED. IEEE, 2016, pp. 43–48.

[8] P.-M. Jodoin, J. Konrad, and V. Saligrama, “Modeling background activity for behavior subtraction,” in ICASSP. IEEE, 2008, pp. 1–10.

[9] R. Mehran, A. Oyma, and M. Shah, “Abnormal crowd behavior detection using social force model,” in CVPR. IEEE, 2009.

[10] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, “Robust real-time unusual event detection using multiple fixed-location monitors,” Transactions on pattern analysis and machine intelligence, vol. 30, 2008.

[11] V. Mahadevan, W. Li, V. Bhadodia, and N. Vasconcelos, “Anomaly detection in crowded scenes,” in CVPR. IEEE, 2010, pp. 1975–1981.

[12] C. Lu, J. Shi, and J. Jia, “Abnormal event detection at 150 fps in matlab,” in ICCV. IEEE, 2013.

[13] ARENA, “Dataset, pets,” 2014. [Online]. Available: http://www.cvg.reading.ac.uk/PETS2014/a.html

[14] W. Luo, W. Liu, and S. Gao, “A revisit of sparse coding based anomaly detection in stacked rnn framework,” in ICCC. IEEE, 2017, pp. 341–349.

[15] W. Sultani, C. Chen, and M. Shah, “Real-world anomaly detection in surveillance videos,” in CVPR. IEEE, 2018.

[16] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. Aggarwal, H. Lee, L. Davis et al., “A large-scale benchmark dataset for event recognition in surveillance video,” in CVPR. IEEE, 2011, pp. 3153–3160.

[17] R. Leyva, V. Sanchez, and C.-T. Li, “The lv dataset: A realistic surveillance video dataset for abnormal event detection,” in IJWBF. IEEE, 2017, pp. 1–6.

[18] M. Pranav, L. Zhenggang et al., “A day on campus—an anomaly detection dataset for events in a single camera,” in ACCV, 2020.

[19] H. Singh, E. M. Hand, and K. Alexi, “Anomalous motion detection on highway using deep learning,” in ICIP. IEEE, 2020, pp. 1901–1905.

[20] M. Rodriguez, J. Sivic, I. Laptev, and J.-Y. Audibert, “Data-driven activity for behavior subtraction,” in CVPR. IEEE, 2008, pp. 1575–1580.

[21] NVIDIA, “Ai city,” 2021. [Online]. Available: http://www.eecs.qmul.ac.uk/~andrea/avss2007

[22] P. Kumar and M. Saini, “Multivariate adaptive Gaussian mixture for scene level anomaly modeling,” in BigMM. New Delhi, India: IEEE, 2020, pp. 54–62.

[23] N. Patil and P. K. Biswas, “A survey of video datasets for anomaly detection in automated surveillance,” in ISED. IEEE, 2016, pp. 43–48.

[24] O. Elharrouss, N. Almadaed, and S. Al-Maadeed, “A review of video surveillance systems,” Journal of Visual Communication and Image Representation, vol. 77, p. 103116, 2021.

[25] P. Kumar and M. Saini, “Multivariate adaptive Gaussian mixture for scene level anomaly modeling,” in BigMM. New Delhi, India: IEEE, 2020, pp. 54–62.

[26] N. Patil and P. K. Biswas, “A survey of video datasets for anomaly detection in automated surveillance,” in ISED. IEEE, 2016, pp. 43–48.

[27] P.-M. Jodoin, J. Konrad, and V. Saligrama, “Modeling background activity for behavior subtraction,” in ICASSP. IEEE, 2008, pp. 1–10.

[28] R. Mehran, A. Oyma, and M. Shah, “Abnormal crowd behavior detection using social force model,” in CVPR. IEEE, 2009.

[29] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, “Robust real-time unusual event detection using multiple fixed-location monitors,” Transactions on pattern analysis and machine intelligence, vol. 30, 2008.

[30] V. Mahadevan, W. Li, V. Bhadodia, and N. Vasconcelos, “Anomaly detection in crowded scenes,” in CVPR. IEEE, 2010, pp. 1975–1981.

[31] C. Lu, J. Shi, and J. Jia, “Abnormal event detection at 150 fps in matlab,” in ICCV. IEEE, 2013.

[32] ARENA, “Dataset, pets,” 2014. [Online]. Available: http://www.cvg.reading.ac.uk/PETS2014/a.html

[33] W. Luo, W. Liu, and S. Gao, “A revisit of sparse coding based anomaly detection in stacked rnn framework,” in ICCC. IEEE, 2017, pp. 341–349.

[34] W. Sultani, C. Chen, and M. Shah, “Real-world anomaly detection in surveillance videos,” in CVPR. IEEE, 2018.

[35] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. Aggarwal, H. Lee, L. Davis et al., “A large-scale benchmark dataset for event recognition in surveillance video,” in CVPR. IEEE, 2011, pp. 3153–3160.

[36] R. Leyva, V. Sanchez, and C.-T. Li, “The lv dataset: A realistic surveillance video dataset for abnormal event detection,” in IJWBF. IEEE, 2017, pp. 1–6.

[37] M. Pranav, L. Zhenggang et al., “A day on campus—an anomaly detection dataset for events in a single camera,” in ACCV, 2020.

[38] H. Singh, E. M. Hand, and K. Alexi, “Anomalous motion detection on highway using deep learning,” in ICIP. IEEE, 2020, pp. 1901–1905.

[39] M. Rodriguez, J. Sivic, I. Laptev, and J.-Y. Audibert, “Data-driven activity for behavior subtraction,” in CVPR. IEEE, 2008, pp. 1575–1580.

[40] NVIDIA, “Ai city,” 2021. [Online]. Available: http://www.eecs.qmul.ac.uk/~andrea/avss2007

[41] P. Kumari, “Situational anomaly detection in multimedia data under concept drift,” in Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 2960–2973.

[42] O. Elharrouss, N. Almadaed, and S. Al-Maadeed, “A review of video surveillance systems,” Journal of Visual Communication and Image Representation, vol. 77, p. 103116, 2021.

[43] P. Kumar and M. Saini, “Multivariate adaptive Gaussian mixture for scene level anomaly modeling,” in BigMM. New Delhi, India: IEEE, 2020, pp. 54–62.

[44] N. Patil and P. K. Biswas, “A survey of video datasets for anomaly detection in automated surveillance,” in ISED. IEEE, 2016, pp. 43–48.
[152] C. Fayet, A. Delhay, D. Lolive, and P.-F. Marteau, “Emo&ly (emotion and anomaly): A new corpus for anomaly detection in an audiovisual stream with emotional context.” in LREC, 2018.

[153] C.-H. Demarty, C. Penet, M. Soleymani, and G. Gravier, “Vsd, a public dataset for the detection of violent scenes in movies: design, annotation, analysis and evaluation,” Multimedia Tools and Applications, vol. 74, no. 17, pp. 7379–7404, 2015.

[154] I. Lefter, G. J. Burghouts, and L. J. Rothkrantz, “An audio-visual dataset of human–human interactions in stressful situations,” Journal on Multimodal User Interfaces, vol. 8, no. 1, pp. 29–41, 2014.

[155] R. Belmonte, A. Aissaoui, S. Mihoubi, B. Allaert, J. Mennesson, I. M. Bilasco, and L. Goncalves, “Barem: A multimodal dataset of individuals interacting with an e-service platform,” in CBMI, 2021.

[156] I. Lefter, G. J. Burghouts, and L. J. Rothkrantz, “Recognizing stress using semantics and modulation of speech and gestures,” Transactions on Affective Computing, vol. 7, no. 2, pp. 162–175, 2015.

[157] B. M. Peixoto, B. Lavi, Z. Dias, and A. Rocha, “Harnessing high-level concepts, visual, and auditory features for violence detection in videos,” Journal of Visual Communication and Image Representation, p. 103174, 2021.

[158] X. Li, Y. Huo, Q. Jin, and J. Xu, “Detecting violence in video using subclasses,” in ACM MM. ACM, 2016, pp. 586–590.

[159] W.-F. Pang, Q.-H. He, Y.-j. Hu, and Y.-X. Li, “Violence detection in videos based on fusing visual and audio information,” in ICASSP. IEEE, 2021, pp. 2260–2264.

[160] P. Wu and J. Liu, “Learning causal temporal relation and feature discrimination for anomaly detection,” Transactions on Image Processing, vol. 30, pp. 3513–3527, 2021.