Advancing Human Fall Detection with TsetFall: A Comprehensive, Fine-Grained, and Publicly Available Dataset

Eduardo Façanha Dutra
PPGIA
Universidade de Fortaleza
Fortaleza-CE, Brazil
eduardo_dut@edu.unifor.br

Thiago Rodrigo Carvalho de Oliveira
PPGIA
Universidade de Fortaleza
Fortaleza-CE, Brazil
thiago.rodrigo@edu.unifor.br

Maria Andréia Formico Rodrigues
PPGIA
Universidade de Fortaleza
Fortaleza-CE, Brazil
andreia.formico@gmail.com

***Abstract*—Human fall detection research remains an underdeveloped area, particularly with respect to utilizing dynamic images for analysis. Existing fall detection datasets often suffer from size, diversity, and representativeness limitations, as they generally comprise a small number of videos featuring a limited range of actions, camera perspectives, and lighting conditions. Moreover, these datasets typically lack the complexity of real-world fall scenarios, as they often exclude distracting objects and offer only one or two camera angles. To address these challenges, this work introduces TsetFall, a comprehensive, fine-grained, and publicly available dataset generated for fall detection research. Employing a novel AI technique devised by the present authors, the dataset annotation process was notably expedited. By bridging the current gaps in available resources, the TsetFall dataset serves as a valuable benchmark for advancing the field and tackling the multifaceted nature of human fall detection.**

***Index Terms*—fall detection, dataset, fine-grained, comprehensive annotation, publicly available, AI-assisted annotation**

I. Introduction

The unresolved problem of human fall detection could be addressed by developing reliable and representative fine-grained annotated datasets of video sequences, an approach not yet explored in previous work. This strategy becomes even more significant in light of the global population aging – the United Nations anticipates a near doubling of the population aged 65 and older, reaching 1.5 billion by 2050 [1]. Falls are the second leading cause of accidental or unintentional injury deaths worldwide, according to the World Health Organization [2]. Globally, an estimated 684,000 fatal falls occur annually, translating to an average of approximately 1,874 falls per day.

To effectively tackle the challenges in fall detection, creating comprehensive datasets becomes a pressing need. These datasets should encompass diverse scenarios, environmental conditions, and participant demographics to ensure that the developed fall detection systems are versatile and accurate. Moreover, comprehensive fine-grained annotations are essential for training machine learning algorithms capable of detecting real-time falls, even in complex situations in the wild.

978-1-6654-8439-8/22/$31.00 ©2023 IEEE

By addressing these challenges and generating high-quality datasets, researchers can drive advancements in fall detection technology and contribute to the prevention and mitigation of fall-related injuries and fatalities globally. However, there remains a significant dearth of benchmark datasets in the field that is based on RGB video imagery for application in fall detection. Existing datasets often suffer from size, diversity, and representativeness limitations, as they generally comprise a small number of videos featuring a limited range of actions, camera perspectives, and lighting conditions. Moreover, these datasets typically lack the complexity of real-world fall scenarios, as they often exclude distracting objects in the recorded scene and offer only one or two camera angles. Usually, datasets contain a limited set of annotations. Essential information, such as the bounding geometries of objects of interest, which demarcate their location within the environment [3], [4], is typically either not included or inadequately represented in these datasets. This information scarcity adversely impacts the performance of fall detection systems that rely on them, yielding less accurate results.

This work introduces TsetFall, a comprehensive, fine-grained dataset annotated with the assistance of artificial intelligence (AI) and human supervision. This resource is publicly available [5] and has been meticulously developed by the Visualization and Interaction Research Group, affiliated with the PPGIA department at the Universidade de Fortaleza, Brazil. By helping to bridge the current gaps in available resources, the TsetFall dataset serves as a valuable benchmark for advancing the field and tackling the multifaceted nature of human fall detection. TsetFall contains high-resolution RGB and infrared images from detailed video sequences, capturing various actions, postures, and falls filmed from four cameras with different viewpoints. The dataset's inclusion of the object's location and the class representing its action eases the training and evaluation of algorithms working with image segmentation, such as object detection algorithms. The variety of camera perspectives offered also allows for the study of more robust solutions to achieve generalization across diverse fall scenarios. The synergy between AI and human supervision has proven successful across various application domains [6]. Subsequently, this paper presents an AI-assisted annotation technique that markedly improves information extraction, streamlines the annotation process, and minimizes human error.

II. Dataset Description

This section describes the TsetFall dataset, including the subject performing the activities in the video sequences, data collection process, video capture configuration, sample size, and data sources.

A. Participant

The video sequences feature a healthy 31-year-old male Computer Science postgraduate from the Universidade de Fortaleza, Brazil, who also set up the filming environment and granted permission for public sharing of his footage. Two pets, a kitten and a German Shepherd, serendipitously participated. The selection of a younger individual for fall scenario enactment was safety-driven, minimizing potential injury risks. A small room equipped with four corner cameras provided a controlled setting with comprehensive coverage and diverse angles for accurate analysis.

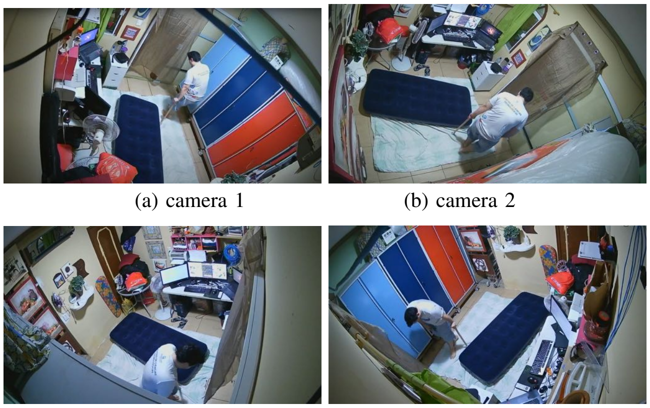
The script of the activities to be filmed was created and directed by one of the authors of this work. The filming took place during the COVID-19 lockdown period. Thus, the capture process was conducted via remote access software, enabling the recording process and control of the sequences' start and end. The recording of scenes was accomplished using Python-implemented software, leveraging the Real-Time Streaming Protocol (RTSP), a technology designed for real-time data transmission over the internet. The OpenCV library was also used to facilitate communication, visualization, and recording of images. Recordings were saved in the AVI format using the XVID codec, a video compression standard that strikes a favourable balance between image quality and file size. This software enabled efficient capture, storage, and display of images from the four cameras used.

B. Human Activity Recognition

In this study, emphasising fall detection, we have established the utilization of a dataset designed for Human Activity Recognition through RGB video images. This format aligns with the conventional approach adopted by Pose Estimation solutions [7], [8], [9].

C. Scenario Description

The scenario used to record the TsetFall dataset consists of a standard room, measuring $3.2 \times 2.8 \times 2.8m$, containing distractor objects that potentially collide with the person in the scene (Fig. 1). Four analogue Intelbras VHD 3220DG cameras with night vision were fixed in the room's four corners at $2.8m$, as shown in Fig. 2. The cameras were connected to a DVR (Intelbras Mhdx 1104) via coaxial cable, allowing simultaneous video recording through a script.


(a) camera 1

(b) camera 2

(c) camera 3

(d) camera 4

Fig. 1: A single scene example (walking with a cane) captured from four distinct camera perspectives, with assorted distractors and potentially colliding objects scattered everywhere.

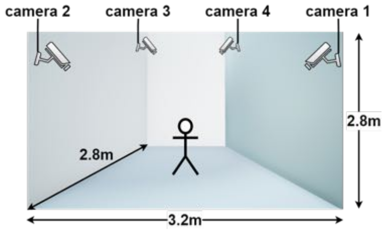


Fig. 2: Illustrative diagram of the room setup in which the TsetFall dataset was recorded.

To obtain a diverse dataset, the subject in the scene was given a script on how to act, directed by a researcher who was a collaborator on the project, with various instructions, such as performing a degraded walk to emulate people with limited mobility, interacting with objects in the scene, etc. In addition, some movements similar to falls were added to the actor's actions, such as lying on chairs, sitting down, and performing physical activities, primarily to evaluate the robustness of fall detection algorithms.

These instructions recorded fall activities and additional actions (36 activities in total) that could make the classification more complex (with movements such as squatting, lying down, falling, sitting, leaning forward, etc.). More specifically, the scenes were captured from four different viewpoints, with RGB images and night vision.

Upon situating the subject within the scene and acknowledging the extensive frame count per sequence, annotations were carried out at four-frame intervals. This process encompassed each of the four distinct camera viewpoints, initiated asynchronously at marginally offset times. The resulting annotation set comprised Axis-Aligned Bounding Boxes (AABBs), which represented the individual's screen coordinates relative to their position in the scene.

To finalize the remaining annotations, an Object Tracking technique [10] was employed alongside various Object Detection solutions [11], [12], as detailed below. Initially, all frames without human annotation were subjected to a detector. Given the numerous false detections in the noisy environment, the object tracking technique was utilized, using human annotations as anchors to propagate the object's location of interest across subsequent frames. Consequently, the intersection over the union criterion for bounding boxes between tracker and detector boxes determined the selection of true detected objects.

The Fall Detection classification set includes: *Confounding (C)*, *Fallen (FN)*, *Falling (FG)*, and *Not Fallen (NF)*. In particular, *C* encompasses movements that may resemble fall precursors, such as squatting, erratic walking, or intentional lying down. Leveraging the single-actor scenario, class annotations are assigned solely to the initial frame of occurrence. Action propagation to subsequent sequence frames links classes to bounding box annotations, integrating boxes labelled "person" into the Fall Detection class selection.

D. Falls and Activities of Daily Living in the Video Sequences

The recorded indoor video sequences and corresponding falls and Activities of Daily Living (ADLs) performed were as follows. The falls were modelled across a spectrum of risk levels, ranging from mild to more severe incidents, while employing an inflatable mattress to attenuate the impact of the falls for safety purposes. The thumbnails of each of the 36 video sequences are shown in Fig. 3. Each image is numbered in accordance with the recorded activities.

1. Walking and interacting with objects.
2. Fainting.
3. Lying on three chairs.
4. Sitting in a chair with a kitten by the side.
5. Sitting on the floor.
6. Walking with a cane, followed by a fall.
7. Walking with a walker, followed by a fall.
8. Falling backwards.
9. Falling forward.
10. Walking and interacting with objects, with a coat hanging on a chair.
11. Sitting on a chair located next to another chair with a coat hanging on it.
12. Falling forward in the dark.
13. Handling a button-up shirt on a hanger in the dark.
14. Jumping jacks exercise in the dark.
15. Front support exercise in the dark.
16. Abdominal exercise in the dark.
17. Push-ups with weights exercise in the dark.
18. Squats exercise in the dark.
19. Falling backwards while holding a mannequin torso in the dark.
20. Walking to the centre of the room, squatting down, standing up, and exiting, all in the dark.
21. Walking crouched in the dark.
22. Walking with a towel turban in the dark.
23. Sit-stand with difficulty in a chair.
24. Sit-stand with difficulty in a chair in the dark.
25. Falling backwards while holding a mannequin torso.
26. Handling a button-up shirt on a hanger.
27. Jumping jacks exercise.
28. Front support exercise.
29. Abdominal exercise.
30. Push-ups with weights exercise.
31. Squats exercise.
32. Walking to the centre of the room and squatting down.
33. Walking crouched.
34. Walking with a towel turban.
35. Struggling to walk with hands behind the back, then sitting, standing, and leaving the room.
36. Walking and interacting with a dog.

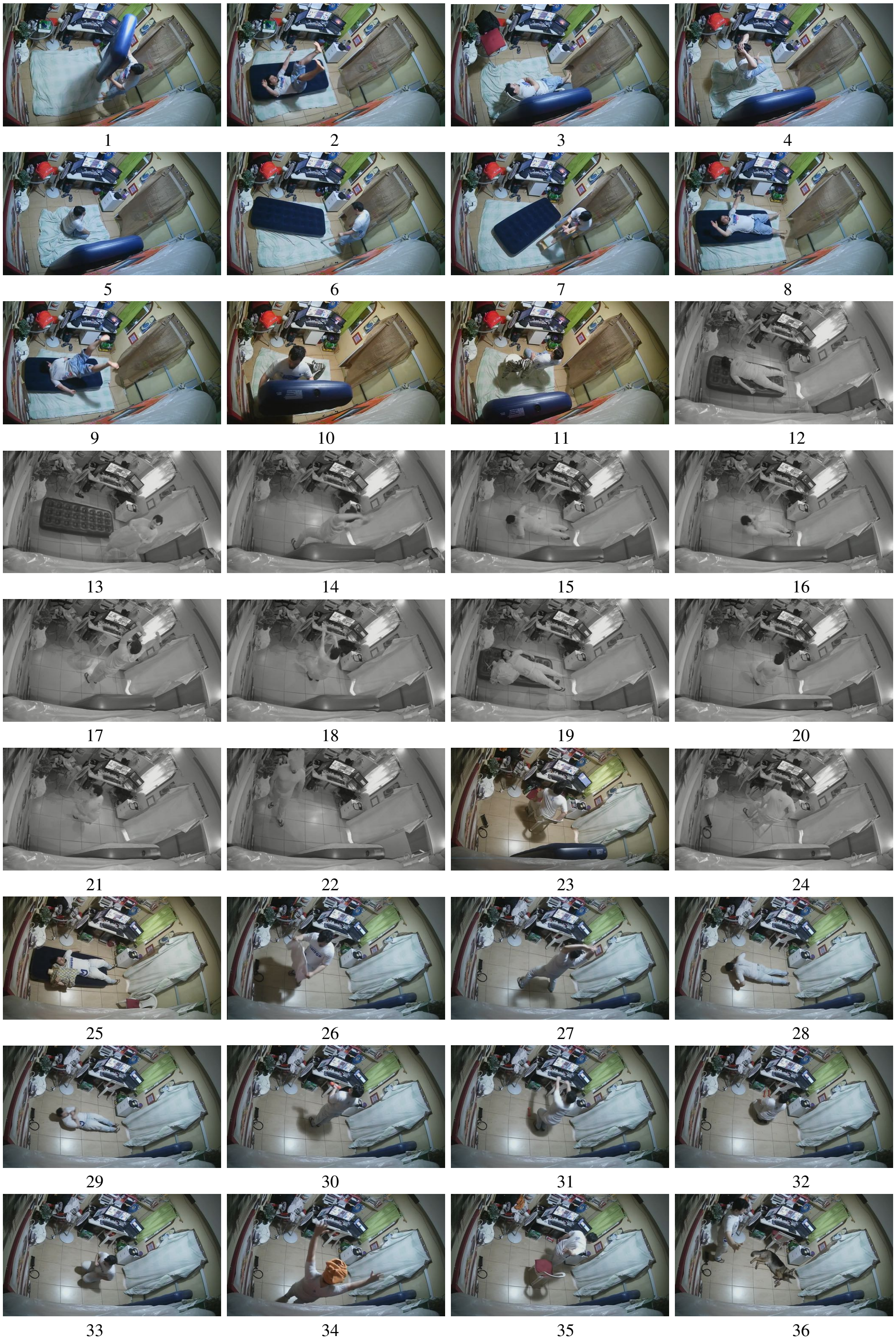
III. Data Preprocessing and Analysis

Usually, the development process of a fall detection solution begins with annotating the dataset to extract the necessary information. Specifically, for a visual fall detection solution, the location of the object of interest in each frame and its respective class according to a predefined set of categories must be obtained from the dataset. Within this context, annotating an extensive dataset like the TsetFall dataset can be laborious and prone to human errors. As a result, our TsetFall dataset was annotated with bounding boxes (every 4 frames) and class labels (C, FN, FG, NF) to extract the necessary information for a visual fall detection solution.

We synchronized each sequence in the dataset frame by frame from the viewpoints of the four cameras, removing excess frames at the beginning and end that did not correspond to other viewpoints, mainly from camera 1. This trimming did not affect frames with the actor, preserving all relevant information for analysis. The synchronization issue may have arisen from the recording script, which initiated recordings at closely spaced yet distinct time points. After synchronization, we observed that the annotations in the video mosaic did not consistently co-occur in most frames, resulting in some viewpoints being annotated while others were not. Additionally, since the class labels were associated with the bounding boxes, there was a lack of continuity in the class label annotations, making it challenging to measure classification results in unannotated frames.

Upon closer analysis of our generated dataset, we noticed that TsetFall has some characteristics that can reduce the need for human annotation, such as only one person in the captured images. This individual enters the scene, performs actions, and leaves the room or falls and remains on the floor for seconds. This allows the class annotation to be performed by observing when the individual's state changes. Furthermore, our careful observations revealed that the presence of only one individual in the scene enables a visual tracking algorithm to estimate the approximate location in frames that have not undergone manual annotation.

Following that, we have established a class set to depict the action state of the person in the scene, including C, FN, FG, NF. This class set may be augmented to create a more sophisticated fall detection solution. However, the labels FN and NF are sufficient for detecting the individual's fall status. Additionally, to enable a more comprehensive



1 2 3 4

5 6 7 8

9 10 11 12

13 14 15 16

17 18 19 20

21 22 23 24

25 26 27 28

29 30 31 32

33 34 35 36

Fig. 3: TsetFall with 36 video sequences featuring human activities (including frontal, lateral, and back falls) by one of the members of the Visualization and Interaction Research Group from PPGIA/UNIFOR, recorded using four cameras in a room with distractors and colliding objects. Captured by camera 2 (see Fig. 2 for reference). TsetFall is publicly available at [5].

comparative analysis of the outcomes, our current research has also included the C and FG labels in the class set.

Each video in the dataset is treated separately and segregated into frames. In this set of frames, the first frames in which one of the predefined labels can be assigned and the frame in which the actor leaves the scene are identified. We refer to these frames as reference frames. The human's location is manually annotated in these frames as an AABB.

After the initial annotation process carried out by a human, two highly intuitive techniques were used to complete the set of annotations. The first technique employs the principle of temporal coherence within a video sequence, indicating that consecutive frames from the same sequence are likely to feature the same object(s) consistently. Thus, label propagation was performed from reference frames to subsequent frames until the last frame of the current action.

The second technique involves using a combination of an Object Detection algorithm [11], [12] and a Visual Object Tracker [10], pre-trained on the COCO, TrackingNet, LaSOT, and GOT10k datasets [13], [14], [15], [16], into an AI-assisted technique, to obtain the location of the person in all frames. Starting from each reference frame, the object tracking algorithm was applied, and thus, the person's approximate location in the scene was propagated to subsequent frames. However, since the bounding box produced by the tracker does not perfectly fit the object in the scene, either by not completely enclosing the object or by encompassing too much of the background, a pre-trained Object Detector is applied to a dataset that includes humans, in this case, COCO [13].

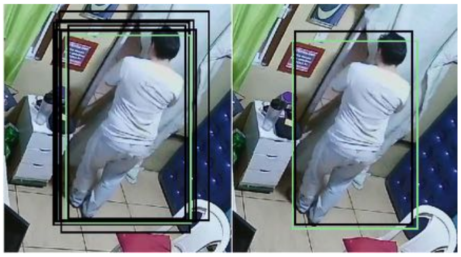


Fig. 4: (a) Overlap of the tracking bounding box (in green) with the object detection algorithm results, and (b) chosen detection (in black) to compose the annotation by the best intersection over union criterion.

Although only one person is in the scene, the object detector generates multiple bounding boxes as output. To choose the detection that best fits the person in the scene, the Jaccard Similarity Calculation [17], as shown in Eq. 1, is used to calculate the ratio between the intersection area and the union area of two bounding boxes A and B. This metric is calculated between the output obtained by the object tracker and the outputs obtained through the object detector. The correct detection is the one that achieves the highest metric among the candidate detections. Additionally, one or more detectors can be used at this stage to generate detections.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

Fig. 4 illustrates the annotation process we implemented and applied to the TsetFall dataset. Table I provides a comprehensive overview of the AI-assisted and human-supervised annotations performed on each of the 36 video sequences.

IV. Baseline method for Fall Detection Evaluation using TsetFall Dataset

This work's baseline fall detection method is based on object detection, specifically utilizing four pre-trained YOLOv8 architectures [18] from the COCO dataset [13]. We trained the models using the bounding box annotations from cameras 1, 2, and 3 of the TsetFall dataset. Camera 4 was reserved for validation. The C, FN, FG, and NF used classes were derived from initial movement annotations and propagated to the following frames. For the training phase, the dataset was confined to the human-annotated bounding box set.

The models were trained for 200 epochs, with an early stopping tolerance of 50 epochs for mAP improvement. Stochastic Gradient Descent [19] was employed as the optimizer, with an initial learning rate of 0.01. The performance of the base models is summarized in Table II. The training results are displayed in Table III. In addition to these findings, the models were subjected to a more rigorous assessment using the 4^{th} camera from the complete annotation set, derived from the data extraction technique that integrates object detection and tracking. These results are in Table IV.

The object detector results were adapted to simplify evaluation and enable future comparisons with other algorithms. To achieve this, the task understanding was streamlined. First, we interpreted the FN and FG classes as *Fall* and the NF and C classes as *NoFall*. The new class set represents the presence or absence of a fallen individual within a frame, rather than representing an action. This expands the fall detection problem, encapsulating it in a more comprehensive manner. Then, for each frame with detected objects from the original classes, we assigned a *Fall* class if at least one object in the scene belonged to that class and *NoFall* if all objects were in the *NoFall* class. For frames without detected objects, the *NoFall* class was assigned. After this simplification, the results shown in Table V were obtained, with the geometric mean values calculated using $Geometric\ Mean = \sqrt{Sensitivity \times Specificity}$.

V. Comparison to Prior Work

In this section, we will provide an overview of existing fall detection datasets [24], [25], [26], [5], based on RGB video sequences, examining their advantages and shortcomings in terms of their applicability and effectiveness in developing fall detection systems.

The Multicam Fall dataset [21] comprises 24 sequences from 8 distinct angles, with 22 focused on falls. Created in a laboratory setting with distractor objects, it features different actors and a balanced proportion of ADLs and falls. Annotations specify the start and end times of the primary

TABLE I: Stratification of object actions based on 36 video takes and four camera perspectives using AI-assisted and human-supervised annotations. Concisely, the class labels are represented as *Confounding* (C), *Fallen* (FN), *Falling* (FG), and *Not Fallen* (NF).

Video Sequence	**Camera #1**				**Camera #2**				**Camera #3**				**Camera #4**				**Overall Frames**
	C	**FN**	**FG**	**NF**	**C**	**FN**	**FG**	**NF**	**C**	**FN**	**FG**	**NF**	**C**	**FN**	**FG**	**NF**	
1				1235				1230				1189				1240	4894
2		137	32	84		137	32	78		137	32	72		137	32	80	990
3	180			49	180			44	180			36	180			46	895
4				430				418				366				383	1597
5	364			62	364			60	364			52	364			55	1685
6		72	20	620		72	23	612		72	23	574		72	23	612	2795
7		150	28	332		150	28	313		150	28	284		150	28	316	1957
8		153	19	94		153	20	89		153	20	80		153	20	92	1046
9		72	21	25		72	21	20		72	21	14		72	21	21	452
10				309				304				272				283	1168
11	77			51	77			49	77			40	77			46	494
12		54	20	58		54	21	48		54	17	45		54	22	46	493
13				229				213				105				206	753
14				238				233				235				231	937
15	134			155	131			134	134			133	140			131	1092
16	232			170	229			153	232			148	226			152	1542
17				544				527				509				519	2099
18	82			170	82			155	82			154	81			151	957
19		59	21	325		59	24	314		59	21	318		59	23	316	1598
20	201			109	201			94	201			99	201			91	1197
21	212			100	189			90	206			79	188			80	1144
22				330				287				296				300	1213
23	420			329	420			299	420			272	420			296	2876
24	390			372	390			346	390			329	390			314	2921
25		71	30	249		71	30	242		72	30	235		71	30	238	1369
26				323				305				288				304	1220
27				251				245				239				241	976
28	181			180	181			167	181			151	181			161	1383
29	304			142	304			136	304			130	304			136	1760
30	55			601	55			590	55			577	54			583	2570
31	167			183	167			172	167			168	167			166	1357
32	221			42	221			38	221			36	221			41	1041
33	284			109	283			93	279			80	284			86	1498
34				295				280				256				278	1109
35	488			294	486			271	482			275	468			257	3021
36				648				618				608				618	2492
Total of Frames	3992	768	191	9737	3960	768	199	9267	3975	769	192	8744	3946	768	199	9116	56591

TABLE II: Pre-trained YOLOv8 model performance on the COCO dataset.

Model	**Image Size (pixels)**	**mAP val (%)**	**Net. Parameters (millions)**	**FLOPs (billions)**
YOLOV8l	640	52.9	43.7	165.2
YOLOV8m	640	50.2	25.9	78.9
YOLOV8s	640	44.9	11.2	28.6
YOLOV8n	640	37.3	3.2	8.7

TABLE III: YOLOv8 model performance on the human-annotated TsetFall subset, limited to camera 4.

Model	**mAP IOU [0.5:0.95]**	**mAP IOU 0.5**	**mAP IOU 0.75**	**AR** [20] **IOU [0.5:0.95] maxDets 10**
YOLOV8l	63.89	88.54	68.88	71.79
YOLOV8m	60.53	83.51	67.49	67.44
YOLOV8s	62.95	85.04	71.76	69.44
YOLOV8n	**65.61**	**89.97**	**74.39**	**72.87**

TABLE IV: YOLOv8 model performance on the complete annotation set of TsetFall, confined to camera 4.

Model	**mAP IOU [0.5:0.95]**	**mAP IOU 0.5**	**mAP IOU 0.75**	**AR** [20] **IOU [0.5:0.95] maxDets 10**
YOLOV8l	**68.17**	85.59	81.27	**74.29**
YOLOV8m	62.37	80.04	74.82	68.19
YOLOV8s	63.82	81.31	77.2	70.23
YOLOV8n	67.58	**85.61**	**81.45**	73.56

TABLE V: Sensitivity, Specificity, and Geometric Mean values for fall detection models.

Model	**Sensitivity**	**Specificity**	**Geometric Mean**
YOLOV8l	**98.96**	**99.94**	**99.45**
YOLOV8m	91.44	99.72	95.49
YOLOV8s	98.64	99.78	99.21
YOLOV8n	97.08	99.93	98.49

TABLE VI: Comparison of RGB datasets for human fall detection

	Multicam Fall [21]	**UR Fall Detection** [22]	**UMA Fall** [23]	**TsetFall** [5]
Dataset size	192 videos	140 videos	11 videos	144 videos
Environment type	Indoor	Indoor	Indoor and outdoor	Indoor
Controlled laboratory environment	Yes	Yes	Yes	Yes
Camera viewpoints	Multi (8)	Multi (2)	Single	Multi (4)
Camera positioning	4 corners and 4 mid-top edges	1 top and 1 lateral	Lateral	4 corners
Resolution	720x480	640x480	854x480	1920x1080
Data acquisition technology	RGB cameras	Kinect sensors	RGB cameras and wearables	RGB and infrared cameras
# of falls (per viewpoint, total)	(22, 176)	(30, 60)	(3,3)	(8, 32)
# of ADLs (per viewpoint, total)	(2, 16)	(40, 80)	(8, 8)	(28, 112)
Fall types	Forward, backward, lateral, others	Forward, backward, lateral	Forward, backward, lateral	Forward, backward, lateral, others
ADL types	Walking, sitting, standing, bending, others	Walking, sitting, standing, bending, others	Walking, sitting, standing, bending, exercising, others	Walking, sitting, standing, exercising, others
Annotation type	Human	Human	N/A	AI-assisted and human-supervised
Annotation level	Frame	Frame	Video Label Only	Object
Annotated features	Start and end times of falls and ADLs	Performing action per frame	N/A	Performing action and subject locations
Limitations	Lack of subject locations, artificial actors behaviour, limited gait and illumination diversity, indoor scenes (one single location)	Lack of subject location, artificial actor behaviour, limited gait and illumination diversity, indoor scenes, low video resolution, limited camera viewpoints	Lack of subject locations, artificial actors behaviour, rare fall events, single camera angle	Single actor scenes, indoor scenes (one single location)
Advantages	Multiple camera angles, multiple subjects, assorted objects	Assorted falls and indoor environments, unique top-view images, commonly used in fall detection studies and publications	Assorted ADLs, outdoor scenes, assorted lighting conditions	Multiple camera angles, assorted objects, assorted ADLs and falls, comprehensive annotation, infrared images

actor's actions in the main camera view. However, the lack of the actor's location complicates feature extraction using computer vision techniques, especially with multiple actors. Additionally, when trained on this dataset, artificial movements and unvaried gait patterns may impact algorithm generalization capabilities.

The UR Fall Detection dataset [22], composed of 70 videos from lateral and overhead perspectives, features diverse daily activities and falls in primarily indoor settings. It provides detailed frame-by-frame annotations, including executed action and depth-related data. However, its low image quality, lack of actor location annotations, controlled lab setting, and limited viewing angles constrain accurate evaluation of modern computer vision methods and the exploration of solution generalization in related studies.

The UMAFall dataset [23] includes 3 fall videos and 8 ADLs from a single viewpoint, featuring indoor and outdoor scenarios with varying lighting conditions and a diverse ADL collection. However, it lacks detailed annotations, only indicating the generic action type in each sequence. Primarily focused on wearables for fall detection, its utility for evaluating computer vision algorithms is limited.

The examined datasets [21], [22], and [23] focus on fall detection but lack necessary updates for training and evaluating computer vision-based deep learning algorithms. Despite varying annotation quality, none supply specific data on the object of interest, which would assist algorithms in extracting relevant scene information and excluding non-essential elements for fall detection.

Conversely, TsetFall dataset [5] stands out for its quality and wealth of information. Unlike other fall detection datasets discussed in this section, it provides comprehensive annotations for all video frames. The dataset's inclusion of the object's location and the class representing its action streamlines the training and evaluation of algorithms for image segmentation, such as object detection algorithms. Additionally, the inclusion of videos from 4 cameras with varying viewpoints allows for an in-depth and diverse assessment of the actor's movements. The range of camera perspectives allows for studying robust solutions to achieve generalization across diverse fall scenarios. Specifically, it is possible to use images from one camera viewpoint as a validation set, evaluating the limitations of a fall detection solution developed using the other 3 viewpoints. This is important in applications involving fall detection in complex and real-world environments, including in the wild, where scene conditions like camera positioning, lighting, image quality, and occlusion may challenge solutions developed in controlled laboratory environments or well-behaved scenarios. Table VI summarizes the comparison of RGB datasets for human fall detection discussed in this section.

VI. Applications and Future Directions

The comprehensive TsetFall dataset, featuring fine-grained annotations, high-resolution RGB and infrared images from four camera perspectives, has extensive applications across sectors like healthcare, elderly care, and sports. It can aid in creating automated fall detection systems for healthcare, improving monitoring and injury prevention in elderly care, and identifying fall-causing factors in sports, enhancing safety and performance. While the TsetFall dataset already provides a substantial resource for fall detection across diverse contexts, further enhancements could include additional annotations like poses, allowing for more advanced machine-learning models incorporating body dynamics. The introduction of outdoor scenes and multiple subjects could generate more realistic, varied fall detection scenarios, contributing to a richer dataset and promoting research advancements in multi-fall detection in real-world conditions.

VII. Concluding Remarks

TsetFall is a novel open dataset presented in this work for fall detection, offering unique, consistent object location annotations across all frames, facilitating multi-perspective scene analysis akin to surveillance systems. Crafted with attention to detail, TsetFall includes complex video sequences with occlusions, varied objects, and humanoid representations, like mannequin busts. This complexity enables more realistic and challenging conditions for model evaluation, enhancing the reliability, stability, and accuracy of results, and making the dataset a valuable resource for future research.

This research also introduces an object detection-based fall detection method, a less-charted approach in the field. The object detection support facilitates the evaluation of the algorithm's capacity to identify the object of interest throughout the dataset. The AI-assisted annotation process enabled a task that would have been challenging for a smaller developer team. Beyond this paper, the comprehensive annotations in the TsetFall dataset allow the evaluation of tracking algorithms. AI-assisted annotations simplify method evaluation, paving the way for future innovation. By adapting these results to a broader fall detection interpretation, comparisons with other dataset solutions are possible, fostering a unified benchmark to propel the field forward.

Acknowledgments

The authors would like to thank the Universidade de Fortaleza (UNIFOR) for its financial support (Ed. DPDI 79/2020).

References

[1] United Nations' Department of Economic and Social Affairs, "World population ageing report," 2019. [Online]. Available: https://www.un.org/en/development/desa/population/publications/pdf/ageing/WorldPopulationAgeing2019-Highlights.pdf

[2] World Health Organization, "Falls," 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/falls

[3] Z. Zou, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *arXiv preprint arXiv:1905.05055*, 2019.

[4] Y. R. Serpa, M. B. Nogueira, P. P. M. Neto, and M. A. F. Rodrigues, "Evaluating pose estimation as a solution to the fall detection problem," in *2020 IEEE 8^{th} International Conference on Serious Games and Applications for Health (SeGAH)*. IEEE, 2020, pp. 1–7.

[5] E. F. Dutra, T. R. C. de Oliveira, and M. A. F. Rodrigues, "TsetFall: A comprehensive, fine-grained, and publicly available dataset generated for human fall detection," 2023. [Online]. Available: https://github.com/eduardodut/TsetFall_dataset

[6] Y. R. Serpa and M. A. F. Rodrigues, "Human and machine collaboration for painting game assets with deep learning," *Entertainment Computing*, vol. 43, p. 100497, 2022.

[7] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "Vitpose: Simple vision transformer baselines for human pose estimation," *arXiv preprint arXiv:2204.12484*, 2022.

[8] Z. Geng, C. Wang, Y. Wei, Z. Liu, H. Li, and H. Hu, "Human pose as compositional tokens," *arXiv preprint arXiv:2303.11638*, 2023.

[9] Y. R. Serpa, L. A. Pires, and M. A. F. Rodrigues, "Milestones and new frontiers in deep learning," in *the 32^{nd} SIBGRAPI Conf. on Graphics, Patterns and Images Tutorials (SIBGRAPI-T)*, 2019, pp. 22–35.

[10] X. Chen, B. Yan, J. Zhu, D. Wang, X. Yang, and H. Lu, "Transformer tracking," in *CVPR*, 2021.

[11] S. Marcel and Y. Rodriguez, "Torchvision the machine-vision package of Torch," in *Proceedings of the 18^{th} ACM International Conference on Multimedia*, 2010, pp. 1485–1488.

[12] G. Jocher, A. Chaurasia, A. Stoken, and et al., "ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation," Nov. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.7347926

[13] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *European Conference on Computer Vision*. Springer, 2014, pp. 740–755.

[14] M. Muller, A. Bibi, S. Giancola, S. Alsubaihi, and B. Ghanem, "Trackingnet: A large-scale dataset and benchmark for object tracking in the wild," in *Proc. of the 2018 ECCV*, 2018, pp. 300–317.

[15] H. Fan, L. Lin, F. Yang, P. Chu, G. Deng, S. Yu, H. Bai, Y. Xu, C. Liao, and H. Ling, "Lasot: A high-quality benchmark for large-scale single object tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5374–5383.

[16] L. Huang, X. Zhao, and K. Huang, "Got-10k: A large high-diversity benchmark for generic object tracking in the wild," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[17] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 658–666.

[18] G. Jocher, A. Chaurasia, and J. Qiu, "YOLO by Ultralytics," Jan. 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[19] L. Ljung, G. Pflug, and H. Walk, *Stochastic approximation and optimization of random systems*. Birkhäuser, 2012, vol. 17.

[20] T.-Y. Lin, M. Maire, S. Belongie, and et al., "Microsoft COCO: Common Objects in COntext," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland*. Springer, 2014, pp. 740–755.

[21] E. Auvinet, C. Rougier, J. Meunier, A. St-Arnaud, and J. Rousseau, "Multiple cameras fall dataset," *DIRO-Université de Montréal, Tech. Rep*, vol. 1350, p. 24, 2010. [Online]. Available: http://www.iro.umontreal.ca/∼labimage/Dataset/

[22] M. Kepski and B. Kwolek, "Embedded system for fall detection using body-worn accelerometer and depth sensor," in *2015 IEEE 8^{th} IDAACS*, vol. 2. IEEE, 2015, pp. 755–759. [Online]. Available: http://fenix.ur.edu.pl

[23] E. Casilari, J. A. Santoyo-Ramón, and J. M. Cano-García, "UMAFall: A multisensor Dataset for the Research on Automatic Fall Detection," *Procedia Computer Science*, vol. 110, pp. 32–39, 2017. [Online]. Available: https://figshare.com/articles/dataset/UMA_ADL_FALL_Dataset_zip/4214283

[24] E. Auvinet, C. Rougier, J. Meunier, A. St-Arnaud, and J. Rousseau, "Multiple cameras fall dataset," *DIRO-Université de Montréal, Tech. Rep*, vol. 1350, 2010.

[25] B. Kwolek and M. Kepski, "Human fall detection on embedded platform using depth maps and wireless accelerometer," *Computer Methods and Programs in Biomedicine*, vol. 117, no. 3, pp. 489–501, 2014.

[26] E. Casilari, J. A. Santoyo-Ramón, and J. M. Cano-García, "Analysis of a smartphone-based architecture with multiple mobility sensors for fall detection," *PLoS one*, vol. 11, no. 12, p. e0168069, 2016.