OmniFall: A Unified Staged-to-Wild Benchmark for Human Fall Detection

David Schneider[*†] **Zdravko Marinov**[†] **Rafael Baur**[†] **Zeyun Zhong**[†] **Rodi Düger**[†]

Rainer Stiefelhagen[†]

Abstract

Current video-based fall detection research mostly relies on small, staged datasets with significant domain biases concerning background, lighting, and camera setup resulting in unknown real-world performance. We introduce *OmniFall*, unifying eight public fall detection datasets (∼14 h of recordings, ∼42 h of multiview data, 101 subjects, 29 camera views) under a consistent ten-class taxonomy with standardized evaluation protocols. Our benchmark provides complete video segmentation labels and enables fair cross-dataset comparison previously impossible with incompatible annotation schemes. For real-world evaluation, we curate *OOPS-Fall* from genuine accident videos and establish a staged-to-wild protocol measuring generalization from controlled to uncontrolled environments. Experiments with frozen pre-trained backbones such as I3D or VideoMAE reveal significant performance gaps between in-distribution and in-the-wild scenarios, highlighting critical challenges in developing robust fall detection systems.
OmniFall: https://huggingface.co/datasets/simplexsigil2/omnifall
Code: https://github.com/simplexsigil/omnifall-experiments

1 Introduction

Falls remain a leading cause of fatal and non-fatal injury in older adults, claiming about 684,000 lives globally and sending more than 37 million people to medical care each year [23]. Preventing death hinges less on detecting the fall itself and more on recognizing the sustained *fallen state*—when a person remains on the ground unable to summon help. Prolonged "long-lie" episodes increase one-year mortality significantly [19, 13], underscoring the importance of rapid detection.

Computer vision-based fall detection systems show tremendous potential for 24/7 monitoring in care environments, but face three critical limitations: (1) most systems focus on detecting brief fall events rather than persistent fallen states; (2) models generalize poorly across environments due to training on small, visually homogeneous datasets; and (3) no standardized benchmark exists to evaluate performance on genuine accidents in uncontrolled settings.

Existing research relies on datasets collected in single environments with constant conditions [15, 14, 24, 1, 6]. Even larger multi-camera collections [3, 21, 17] feature similar controlled settings. This lack of diversity leads to systems that perform well within their training domain but fail under novel conditions. Furthermore, incompatible annotation taxonomies hinder cross-dataset comparisons.

To address these challenges, we introduce *OmniFall*, a unified benchmark with two key contributions:

First, we create a densely annotated *Staged-Fall-Superset* unifying eight datasets under a consistent ten-class taxonomy distinguishing between transient actions (*fall*, *sit down*, *lie down*, *stand up*) and

[*]Corresponding author: david.schneider@kit.edu
[†]Karlsruhe Institute of Technology

Preprint. Under review. Please do not distribute!

static states (*fallen*, *sitting*, *lying*, *standing*). This enables detection of both fall events and medically critical fallen states across 14 hours of footage with 101 subjects and 29 camera views.

Second, we introduce *OOPS-Fall*, containing genuine accidents from uncontrolled environments, establishing the first rigorous staged-to-wild evaluation protocol for fall detection.

Our experiments with frozen backbones reveal significant performance gaps between laboratory settings and real-world conditions, emphasizing the need for comprehensive benchmarks like OmniFall to develop systems that work reliably where they matter most—in real-world settings where prompt medical intervention can save lives.

2 Related Work

Many fall detection datasets have been introduced in recent years [21, 17, 6, 1, 24, 3, 14, 2, 15, 4, 7, 8], reflecting the growing interest in automated fall detection. As part of our related work, we conducted a systematic search to identify publicly available video-based datasets to characterize the landscape of the fall detection field. We found that most existing datasets feature staged falls performed by healthy individuals in controlled settings, with no standardized benchmarks for evaluating how models trained on such data perform on real-world falls. In the following, we first review the datasets included in our OmniFall benchmark, then discuss additional datasets we identified but did not include.

Table 1: Overview of staged fall detection datasets in our OmniFall benchmark. A *recording* is a single event that is captured by multiple synchronized camera views, i.e., in multiple videos. *Recording duration* is the total length of all events, while *total duration* is the combined length of all videos. **Rec**: Recordings, **h**: hours, **m**: minutes, **Syn**: synchronized camera views, **Sub**: subjects

Dataset	Rec	Rec Duration	Total Duration	Views	Syn.	Resolution	FPS	Sub
CMDFall [21]	55	7h 25m	27h 59m	7	✓	640×480	20	50
UP Fall [17]	561	4h 37m	9h 15m	2	✓	640×480	18	17
Le2i [6]	222	51m	51m	6	✗	320×240	25	9
GMDCSA [1]	160	21m	21m	3	✗	1280×720	30	4
EDF [24]	5	14m	28m	2	✓	320×240	30	5
OCCU [24]	5	14m	28m	2	✗	320×240	30	5
MCFD [3]	24	18m	2h 25m	8	✓	720×480	≤30	1
CAUCA Fall [14]	100	16m	16m	1		720×480	23	10

Fall Detection Datasets in OmniFall. To collect existing public fall detection datasets from prior work, we conducted a systematic search in accordance with PRISMA guidelines (5)-(9)[3]. We queried PubMed and Google Scholar using the keywords: [fall] and [detection] and ([survey] or [review]). Duplicates were removed, including pre-prints later published as peer-reviewed articles. We then manually screened titles and abstracts for relevance. Surveys were included if they met the following criteria: 1) full text available in English; 2) peer-reviewed; and 3) listed public RGB-based video fall detection datasets. Of the 98 surveys retrieved, only 9 met all inclusion criteria. The public datasets (1)-(8) extracted from these 9 surveys and included in OmniFall are listed in Table 1.

(1) The **CMDFall** dataset [21] is a multimodal, multiview collection acquired from seven synchronized Kinect sensors (providing color and depth data) together with two accelerometers. It comprises recordings of 50 subjects performing eight distinct fall types alongside twelve non-fall activities, supporting cross-view and multimodal analysis. (2) **UP Fall** [17] offers a multimodal setup from 17 subjects in a laboratory environment. It includes vision data from a pair of synchronized cameras (640×480 at 18 fps), inertial measurements from accelerometers and gyroscopes worn at different body locations, and context sensors, covering six activities of daily living (ADL) and five fall types. (3) The **Le2i** dataset [6] contains 143 staged falls and 79 ADL actions recorded from a single fixed camera at 640×480 pixels and 25 fps across four environments and six views, stressing various real-world issues such as occlusions and lighting variations. (4) The **GMDCSA** dataset [1] delivers high temporal resolution data, capturing 81 fall and 79 ADL clips recorded with a stationary laptop

[3] `https://www.prisma-statement.org/prisma-2020-checklist`

Figure 1: Examples from selected fall detection datasets.

camera at 1280×720 pixels and 30fps in three distinct home settings. Its design emphasizes intra-class variability, with subjects altering clothing and backgrounds between clips. (5) The **EDF** and (6) **OCCU** datasets [24] were acquired in a controlled apartment environment using two Kinect sensors at 320×240 pixels and 30 fps. The EDF set focuses on 40 non-occluded falls by five subjects performing falls from eight directions and 30 non-fall actions, while OCCU presents a more challenging scenario in which 30 falls become partially occluded near the end, accompanied by a suite of 80 ADL actions. (7) **The Multiple Camera Fall (MCFD)** dataset [3] leverages eight inexpensive, synchronized IP cameras installed in a fixed indoor setting to capture 22 fall scenarios interleaved with pre-fall ADL sequences, facilitating 3D reconstructions either explicitly or implicitly. (8) The **CAUCA Fall** dataset [14] provides recordings from 10 subjects in an apartment setting under varying lighting conditions. Captured at 720×480 resolution and 23 fps, it includes both color and infrared modes (in low-light situations) with detailed frame-wise annotations that report not only fall versus non-fall labels but also spatial metrics such as the distance and angle with respect to the camera. **OOPS!** [11] is a 20.7 k-clip (>50 h) video dataset with labels marking when intentional actions turn accidental, supporting benchmarks in intention classification, failure-point localization and accident forecasting.

Other Datasets. Our systematic search identified several additional fall detection datasets, but we excluded them from OmniFall due to specific limitations. UR Fall [15] includes 30 falls and 40 ADLs recorded with two Kinect cameras. However, its limited size (≈6 minutes of footage) and lack of subject and camera annotations made it unsuitable. MUVIM [10] offers multimodal data (RGB, depth, infrared, thermal), which benefits privacy and low-light scenarios. Nonetheless, all videos are captured from a top-down view, which is incompatible with our benchmark's intended camera perspectives. FPDS [16] consists of static images (1072 fall, 1262 ADL) captured by a robot-mounted camera at 76 cm height. Since our benchmark requires video data, we excluded datasets composed solely of still frames. Finally, while High Quality FSD [4], FallFree [2], SIMPLE [8], and ACT4^2 [7] align well with our objectives, we could not include them due to missing usage licenses or the lack of explicit permission from the authors.

3 Dataset

Our benchmark unifies eight *staged fall* datasets and complements them with genuine falls from OOPS [11], creating a staged-to-wild evaluation pipeline. As visualized in Figure 3, the datasets vary significantly in size and composition. In our aggregated staged-fall-superset, the staged-to-wild domain gap comprises multiple distribution shifts (subject identity, camera viewpoint, background,

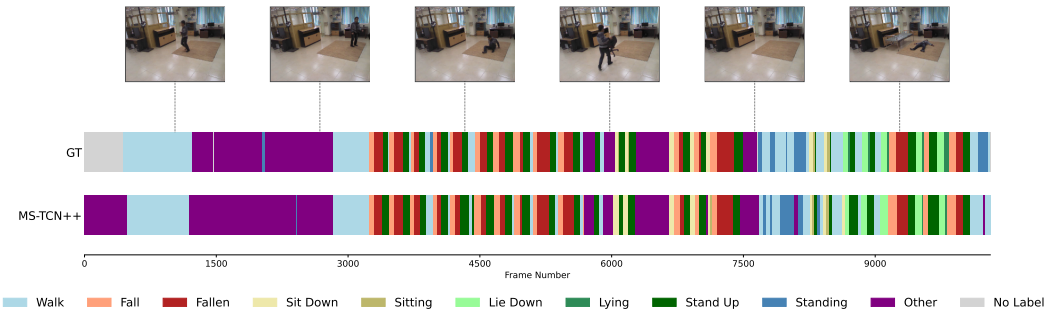

Figure 2: Our annotations and MS-TCN++ [12] action segmentation results.

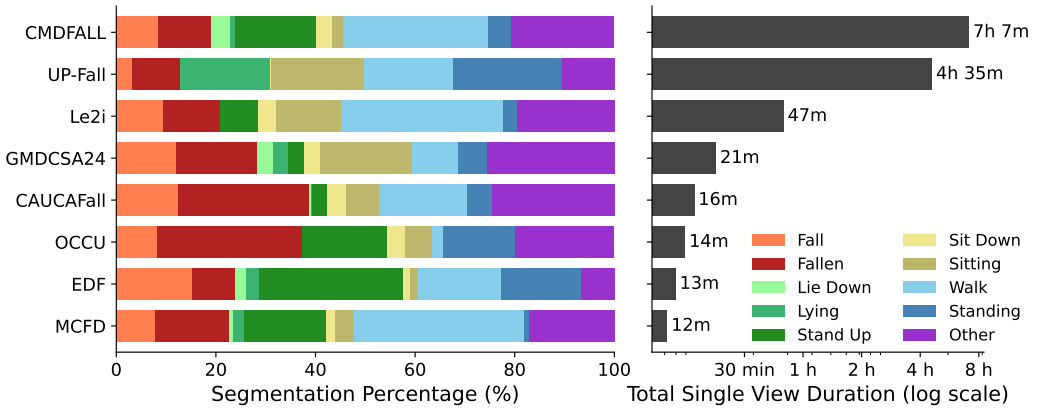

Figure 3: Segmented share of each label within datasets and total single view duration.

illumination, and safety props absent in real-world footage). To address these challenges, we create complementary *cross-subject* and *cross-view* splits for each constituent dataset where possible.

As depicted in Figure 3, our collection spans 14 hours of unique recordings. Large datasets (*CMDFall*: 7h 7m, *UP Fall*: 4h 35m) provide most training data, while *Le2i* (47m) offers mid-sized contribution with varied rooms. Smaller datasets (*GMDCSA24*, *CAUCAFall*, *OCCU*, *EDF*, *MCFD*: 12-21m each) contribute environmental diversity. Synchronized views are excempt from this calculation, *MCFD*'s eight views expand to 1.5 hours of footage across perspectives.

For real-world evaluation, we curate 1.3k *fall* / *no fall* segments from from the 20,000-video OOPS dataset [11] to create *OOPS-Fall* by filtering with original tags, then manually verifying and annotating fall and non-fall segments. This creates a challenging in-the-wild test set with authentic falls in uncontrolled conditions, enabling evaluation of generalization beyond laboratory constraints.

Annotations and Problem Formulation We cast fall detection as a multi-granular video classification problem using a consistent ten-class taxonomy: four transient actions (*fall*, *sit down*, *lie down*, *stand up*), four static states (*fallen*, *sitting*, *lying*, *standing*), *walk*, and *other*. This approach offers three advantages: (1) detecting medically critical *fallen* states even when the fall event is missed due to occlusions; (2) distinguishing genuine falls from similar activities like intentional lying down; and (3) supporting multiple detection granularities (10-class understanding, binary fall/fallen detection, unified fall/fallen alarming). We re-annotate all staged datasets with this taxonomy for cross-dataset compatibility, while for OOPS-Fall we mark fall and no fall segments.

Cross-Subject Splits. We implement dataset-specific subject splits, following original protocols where available: CMDFall (following published odd/even protocol: train: odd-IDs 1-49, val: IDs 2/8/18/42/48, test: remaining 20 even-IDs); CaucaFall (train: S1-S7, val: S10, test: S8-S9, with S6/S8 being nighttime recordings); EDF (train: Jianjun/Jinhui/Peter, val: Songsong, test: Zhong);

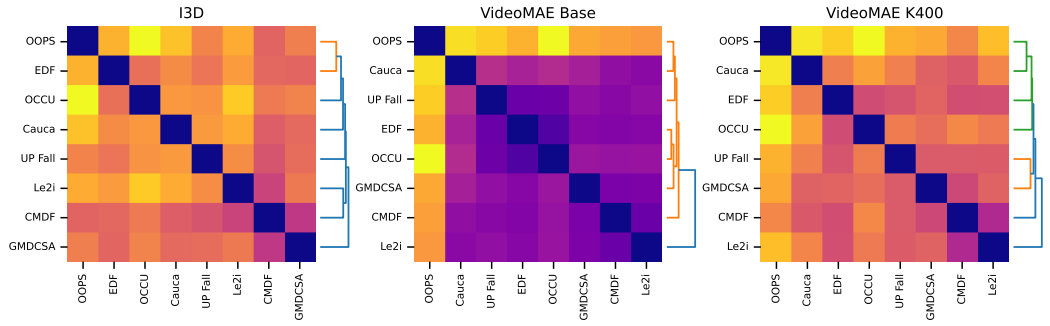

Figure 4: Pairwise Fréchet Video Distances calculated on features from I3D, VideoMAE, and VideoMAE pretrained on Kinetics400, visualised as heat-maps. Darker colors indicate higher similarity (smaller distances) between datasets. Numerical values are listed in the supplementary.

OCCU (train: Jiayan/Wangde/Yangyi, val: Zhangbo, test: Zhouhui); GMDCSA (train: subjects 1/3, val: subject 2, test: subject 4); Le2i (train: subjects 0/1/3/4/6/8, val: subject 5, test: subjects 2/7); UP Fall (train: IDs 1/3/5-10/12-14/17, val: ID 11, test: IDs 2/4/15/16). MCFD's single subject is used only for training in cross-subject evaluation.

Cross-View Splits. For multi-view datasets, we create camera-based splits: CMDFall (train: camera 1, val: camera 5, test: cameras 2/3/4/6); EDF and OCCU (each using train/val: view 1, test: view 2); GMDCSA (train: subjects 2/3 location with identical room/sensor, val: subject 1 location, test: subject 4 location); Le2i (train: Coffee-room and Lecture-room views, val: Office view, test: Home-room views); MCFD (train: camera 1, val: camera 3, test: cameras 2/4/5/6/7/8); UP Fall (train/val: camera 2, test: camera 1). Single-view CaucaFall participates only in training for cross-view evaluation.

Staged-to-Wild Evaluation. We concatenate all staged dataset training partitions to form a single source domain and train each model on this union (for either CS or CV). We then evaluate out-of-distribution (OOD) generalization to the *OOPS-Fall* test set, which presents varied backgrounds, lighting conditions, camera qualities, and naturally occurring falls without safety equipment. This approach quantifies model robustness to real-world conditions and assesses deployment potential in environments where falls are unplanned and uncontrolled.

4 Experiments

Backbone networks We decouple representation learning from classification by using pre-trained frozen backbones as feature extractors, allowing systematic comparison of visual representations across domains. In the main paper, we focus on three representative backbones: I3D (Kinetics-400 pre-trained), VideoMAE-Pre (MAE pre-training only), and VideoMAE-K400 (with Kinetics-400 fine-tuning). Additional backbone results and feature settings are listed in the supplementary.

4.1 Domain–shift analysis

To understand generalization challenges between datasets, we analyze domain shifts using 5000 randomly sampled pooled features per dataset, each consisting of averaged 18-feature sequences.

Dataset similarity metrics. Figure 4 shows pairwise Fréchet Video Distances (FVD) between datasets. The versatile CMDFall dataset demonstrates high similarity (darker colors) to most other datasets, highlighting its versatility as a training source. As expected, OOPS-Fall shows substantial dissimilarity to all staged datasets, corresponding to the OOD performance gaps in Section 4.2.

Feature space visualization. Figure 5 presents T-SNE [22] and H-NNE [18] embeddings for our three backbones. Features cluster by dataset rather than by action class across all networks, demonstrating significant domain gaps. I3D features exhibit more visible intra-domain activity clustering compared to transformer-based backbones, suggesting a simpler, more action-focused

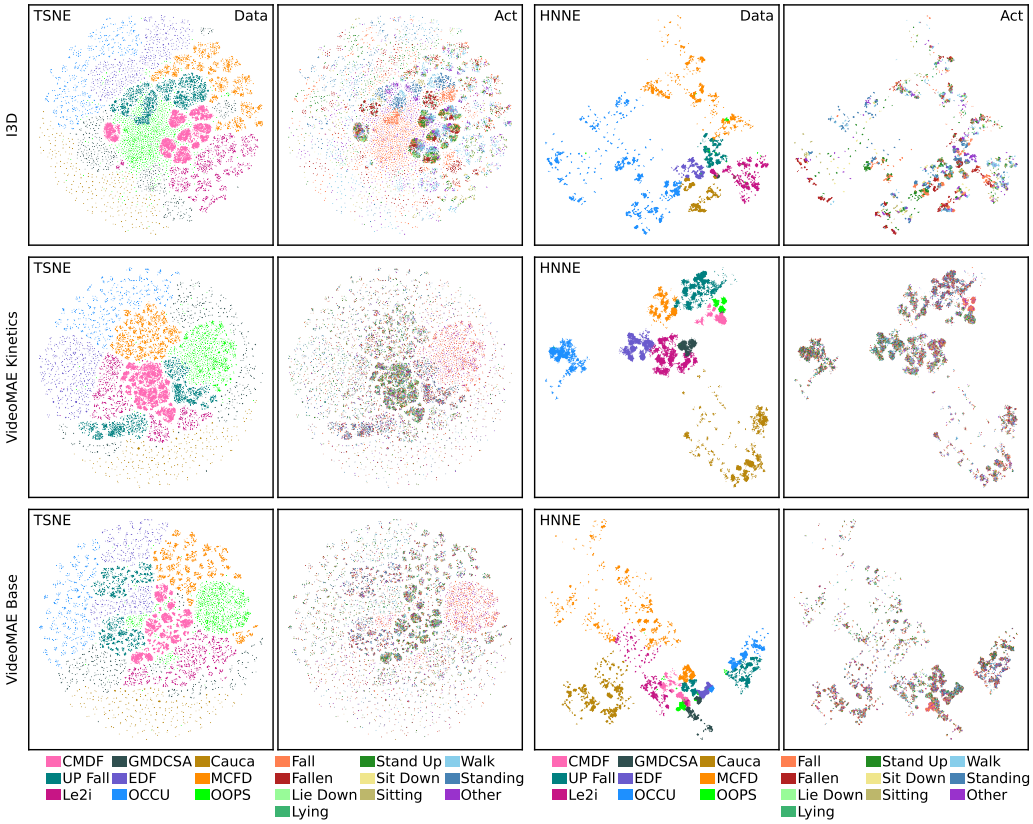

Figure 5: Low-dimensional clip embeddings, colour-coded by dataset and activity. Rows correspond to feature type, columns show T-SNE and H-NNE clustering. Full-size plots in supplementary.

latent space. This likely stems from I3D's sequential ImageNet pre-training followed by Kinetics fine-tuning versus the more complex self-supervised pre-training of transformer models, which capture broader visual characteristics. This observation helps explain I3D's superiority on out-of-distribution fall detection in Section 4.2, as its coarse representations may generalize better to unseen domains.

4.2 Action and fall classification

Classification Architecture and Training For classification, we use a lightweight temporal transformer with two encoder blocks ($d_{\text{model}} = 768$, 8 heads, 4× feed-forward expansion, dropout 0.3) that processes 18 input tokens with sinusoidal position encodings.

To handle dataset and class imbalance, we implement a two-level weighting strategy: (1) For dataset balancing, we weight each domain D_i by $w_i = \min(\frac{\max_j |D_j|}{|D_i|}, 20)$, using capped oversampling to raise smaller datasets (e.g., GMDCSA24, 459 segments) to contribute meaningfully without excessive duplication in order to mitigate the dominant influence of the larger ones (CMDFall, 40k segments). (2) We apply class-balanced weighting using effective sample numbers with label-smoothed ($\alpha = 0.1$) cross-entropy loss [9].

Training uses AdamW optimizer (lr=5×10^{-4} with cosine decay, weight decay 0.01, momentum 0.9/0.999, gradient clipping 1.0) for 150 epochs with batch size 256 across 4 GPUs, selecting the best validation model.

Experimental Setup We use predefined splits from Section 3, merging all staged datasets into a unified training pool while keeping OOPS-Fall for out-of-distribution testing. We perform separate training runs for cross-subject and cross-view evaluations, reporting balanced accuracy, overall accuracy, and macro-F1 for the 10-class task, plus sensitivity, specificity and F1-score for binary fall

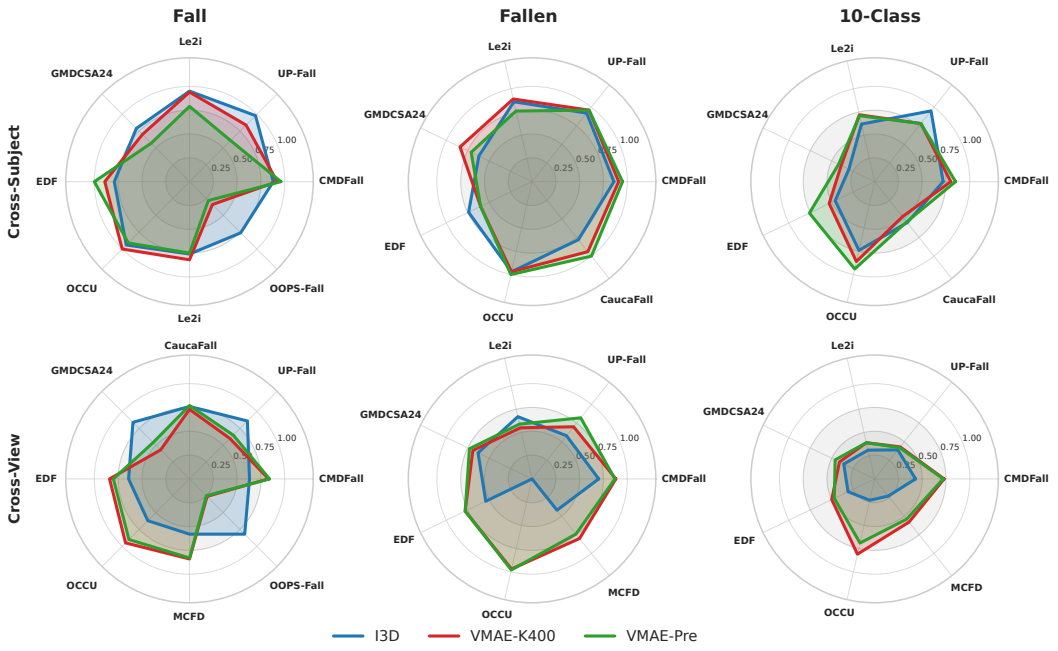

Figure 6: F1 scores on the classes *fall* and *fallen* as well as macro F1 scores on all classes.

detection subtasks. Tables 2–3 provide detailed metrics, while Figure 6 visualizes key F1-scores. We mark best values accross the three backbone tables in bold formatting.

I3D (Table 2) delivers mid-range in-distribution performance for the 10-class task (overall balanced accuracy: CS 0.72, CV 0.44), yet is the *only* model that maintains high binary fall detection sensitivity on OOPS-Fall (Se: 0.68 CS, 0.82 CV). Its binary fall specificity remains >0.98 across all domains, indicating that I3D captures generic motion patterns while avoiding over-reliance on appearance cues.

VideoMAE-K400 (Table 3), pre-trained with masked autoencoding then fine-tuned on Kinetics-400, improves the overall 10-class balanced accuracy by $+6\%$ CS and $+25\%$ CV relative to I3D. This pre-training approach narrows the cross-subject to cross-view gap on CMDFall (10-class BAcc: 0.79→0.71) and achieves strong performance on OCCU (CV 10-class BAcc: 0.85). However, it performs poorly on binary fall detection for OOPS-Fall (Se: 0.21 CS, 0.15 CV), despite maintaining high specificity (≥ 0.96), suggesting a overfitting towards the training distribution.

VMAE-K400-Pre (table in supplementary) uses only self-supervised pre-training without Kinetics fine-tuning. This model achieves the best overall cross-subject 10-class performance (BAcc: 0.83) and nearly matches the Kinetics-tuned variant for cross-view (BAcc: 0.68). However, for binary fall detection on OOPS-Fall, it shows limited sensitivity (Se: 0.17 CS, 0.14 CV), comparable to the Kinetics-tuned version but significantly behind I3D.

Cross-subject vs. cross-view. Masked auto encoder pre-training excels at subject generalisation for binary fall detection: VMAE-K400-Pre yields the highest cross-subject sensitivities on multiple datasets (CMDFall: 0.97, EDF: 1.00, CaucaFall: 1.00). Conversely, Kinetics-400 action fine-tuning introduces pronounced view invariance: VideoMAE-K400 achieves the highest cross-view binary fall sensitivities (CMDFall: 0.88, OCCU: 0.93) and improves 10-class balanced accuracy across nearly all datasets. These patterns suggest supervised signals encourage alignment across camera geometries. The training corpus of the cross-subject split is significantly larger than for cross-view.

Effect of dataset size. Performance trends correlate with training-set representation. On the large, domain-diverse CMDFall dataset, all backbones achieve similar cross-subject 10-class F1 scores (0.80–0.85) but diverge for cross-view evaluation, where the Kinetics-supervised transformer gains approximately 10 percentage points in balanced accuracy. For smaller datasets (GMDCSA24, EDF, MCFD), transformer models outperform I3D by up to 15 percentage points in 10-class balanced

Table 2: **I3D** [5] on cross-subject and cross-view. **B**alanced **Acc**uracy, **Se**nsitivity, **Sp**ecificity.

Dataset		10-class			Fall Δ			Fallen Δ			Fall ∪ Fallen Δ		
		BAcc	Acc	F1	Se	Sp	F1	Se	Sp	F1	Se	Sp	F1
CMDFall [21]	CS	0.72	0.77	0.72	0.84	0.99	0.88	0.84	0.99	0.86	0.88	0.98	0.91
	CV	0.45	0.56	0.43	**0.83**	0.87	0.63	0.64	0.98	0.70	0.80	0.85	0.71
UP-Fall [17]	CS	**0.96**	**0.96**	**0.95**	**0.97**	**1.00**	**0.98**	0.89	0.99	0.92	**0.94**	**0.99**	**0.96**
	CV	0.41	**0.69**	0.39	**0.94**	**0.93**	**0.86**	0.51	**0.94**	0.58	0.75	**0.84**	**0.75**
Le2i [6]	CS	0.61	0.73	0.62	0.95	**0.99**	**0.95**	**0.86**	0.98	**0.86**	0.95	**0.99**	**0.95**
	CV	0.37	0.50	0.31	0.89	**0.93**	0.76	**0.74**	0.94	**0.67**	0.84	0.86	0.73
GMDCSA24 [1]	CS	0.38	**0.52**	0.30	**0.76**	**0.96**	**0.79**	**0.88**	0.79	0.62	0.85	0.69	0.72
	CV	0.40	**0.60**	0.36	**0.76**	**0.99**	**0.84**	**0.76**	0.86	0.63	**0.79**	0.81	**0.75**
EDF [24]	CS	0.44	0.58	0.46	0.94	0.94	0.79	**0.65**	0.98	**0.74**	**0.78**	0.90	0.77
	CV	0.39	0.42	0.31	0.52	**0.98**	0.64	0.62	0.91	0.54	0.66	0.90	0.69
OCCU [24]	CS	0.75	0.86	0.74	0.94	0.99	0.94	0.94	**1.00**	0.97	0.94	**0.99**	0.95
	CV	0.30	0.38	0.23	**0.95**	0.76	0.62	0.00	**1.00**	0.00	0.55	0.73	0.53
CaucaFall [14]	CS	0.56	0.70	0.55	0.89	**0.89**	**0.76**	0.78	0.95	0.78	0.94	**0.86**	**0.87**
MCFD [3]	CV	0.24	0.41	0.23	0.77	0.80	0.58	0.51	0.86	0.42	0.81	0.64	0.63
Overall	CS	0.72	0.76	0.72	**0.82**	0.98	**0.86**	0.85	0.98	0.86	0.86	0.98	0.89
	CV	0.44	0.56	0.43	**0.82**	0.87	0.65	0.62	**0.97**	0.67	0.80	0.83	0.71
Oops-Fall	CS				**0.68**	0.79	**0.76**						
	CV				**0.82**	0.63	**0.82**						

accuracy when sufficient labeled context is available (particularly in cross-view settings for MCFD), demonstrating how their richer semantic representations excel when overfitting risk is reduced.

Take-aways. Our experimental results reveal interesting differences between the frozen backbones:

(1) The frozen I3D backbone shows modest in-distribution performance but exhibits surprisingly strong generalization to OOPS-Fall, suggesting that its simpler feature space may capture more fundamental motion patterns relevant to falls.

(2) The frozen VideoMAE backbone (with Kinetics-400 fine-tuning during pre-training) achieves higher in-distribution accuracy and view invariance, particularly on datasets with camera diversity.

(3) The frozen VideoMAE backbone with only self-supervised pre-training performs exceptionally well on cross-subject generalization but struggles with the domain shift to in-the-wild videos.

4.3 Timeline segmentation

We train MS-TCN++ [12] for action segmentation on the cross-subject split of all staged datasets (downsampled to 10 fps) with I3D [5] features, excluding OOPS-Fall for out-of-distribution testing. Table 4 reports standard metrics: segmental F1 scores at IoU thresholds of 10%, 25%, and 50%, normalized edit distance, and frame-wise accuracy. The model achieves strong performance on in-distribution data (>90% on all metrics for UP-Fall [17] and GMDCSA24 [1]), but on OOPS-Fall, it maintains high frame accuracy with lower F1 scores, suggesting over-segmentation on out-of-distribution data. Figure 2 shows a qualitative example of segmentation results.

5 Conclusion

We introduce OmniFall, a unified benchmark addressing key limitations in fall detection research by integrating eight public datasets with consistent ten-class annotations that distinguish between transient actions and static states. Our standardized cross-subject/cross-view splits and OOPS-Fall facilitate the first comprehensive evaluation of generalization capabilities in this domain.

Table 3: **VMAE-K400** [20] on cross-subject and cross-view. **B**al. **Acc**uracy, **Se**nsitivity, **Sp**ecificity.

Dataset		10-class			Fall Δ			Fallen Δ			Fall ∪ Fallen Δ		
		BAcc	Acc	F1	Se	Sp	F1	Se	Sp	F1	Se	Sp	F1
CMDFall [21]	CS	0.79	0.83	0.80	0.93	0.99	0.92	0.87	**1.00**	0.91	0.93	0.99	0.95
	CV	**0.71**	**0.81**	**0.73**	0.88	0.96	**0.83**	**0.95**	0.97	**0.88**	**0.92**	**0.93**	**0.87**
UP-Fall [17]	CS	0.91	0.88	0.78	0.83	0.96	0.84	**0.95**	0.99	**0.96**	0.89	0.93	0.90
	CV	**0.44**	0.61	**0.43**	0.69	0.84	0.60	0.83	0.87	**0.70**	0.85	0.68	0.73
Le2i [6]	CS	**0.71**	**0.82**	**0.72**	**1.00**	0.98	0.94	0.81	**1.00**	0.89	**0.98**	**1.00**	**0.99**
	CV	**0.47**	**0.59**	**0.39**	0.97	0.89	0.73	0.48	**0.96**	0.55	**1.00**	**0.92**	**0.88**
GMDCSA24 [1]	CS	**0.42**	0.55	**0.39**	0.88	0.86	0.70	0.76	**0.99**	**0.84**	**0.91**	**0.85**	**0.84**
	CV	**0.45**	0.47	**0.41**	0.35	0.93	0.43	0.71	0.92	**0.69**	0.65	**0.88**	0.70
EDF [24]	CS	0.54	0.65	0.53	**1.00**	0.96	0.89	0.45	0.99	0.60	0.72	0.96	0.79
	CV	**0.55**	**0.69**	**0.50**	**0.79**	0.98	**0.84**	**0.97**	**0.93**	**0.78**	**0.97**	**0.94**	**0.91**
OCCU [24]	CS	0.89	0.89	0.86	**1.00**	**1.00**	**1.00**	**1.00**	0.99	0.97	**1.00**	**0.99**	**0.98**
	CV	**0.85**	**0.86**	**0.81**	0.93	0.99	**0.95**	0.97	0.99	0.97	**0.96**	**0.99**	**0.98**
CaucaFall [14]	CS	0.49	0.66	0.47	**1.00**	0.89	**0.82**	0.89	**1.00**	0.94	0.94	0.86	0.87
MCFD [3]	CV	**0.56**	**0.72**	**0.58**	**0.85**	**0.96**	**0.84**	**0.79**	**0.97**	**0.80**	**0.84**	**0.92**	**0.84**
Overall	CS	0.78	0.81	0.79	0.81	0.98	0.86	0.86	**1.00**	**0.91**	0.86	**0.99**	**0.91**
	CV	**0.69**	**0.78**	**0.70**	0.78	**0.96**	**0.78**	**0.93**	**0.97**	**0.86**	**0.87**	**0.92**	**0.84**
Oops-Fall	CS				0.21	**0.97**	0.34						
	CV				0.15	**0.97**	0.26						

Table 4: Timeline segmentation performance of MS-TCN++ [12] with I3D [5] features.

	F1@{10,25,50}			Edit	Acc
CMDFall [21]	72.81	69.07	56.66	65.27	73.79
UP-Fall [17]	93.77	93.77	91.89	95.29	98.53
Le2i [6]	72.35	68.20	56.68	73.69	67.03
GMDCSA24 [1]	94.87	94.87	94.87	96.40	99.19
EDF [24]	66.41	60.23	44.79	60.79	55.15
OCCU [24]	79.82	77.19	71.93	70.14	78.76
CaucaFall [14]	70.91	70.91	61.82	70.20	78.24
Overall	70.33	66.42	53.78	61.10	74.99
Oops-Fall	45.55	39.26	21.82	44.86	64.96

Our experiments reveal distinct generalization patterns across visual representations: I3D shows surprising robustness to in-the-wild fall detection despite modest in-distribution performance, while transformer-based models excel in controlled settings but struggle with domain shifts. These findings emphasize the importance of benchmark diversity and multi-dimensional evaluation protocols.

Limitations. Despite our efforts to create a diverse benchmark, several limitations remain: (1) our in-the-wild test set represents only a fraction of possible real-world fall scenarios; (2) most source datasets feature young subjects in controlled environments, creating demographic biases and (3) privacy concerns persist for deployment in care settings.

Societal Impact. While fall detection systems have clear potential for elder care and health monitoring, we acknowledge potential negative impacts: surveillance concerns, false alarms causing trust erosion, reduced human oversight, and performance disparities across demographics. Future work should address these challenges through domain adaptation, efficient deployment methods, privacy-preserving approaches, and inclusive data collection from underrepresented populations.

References

[1] Ekram Alam, Abu Sufian, Paramartha Dutta, Marco Leo, and Ibrahim A. Hameed. GMDCSA-24: A dataset for human fall detection in videos. 57:110892.

[2] Mona Saleh Alzahrani, Salma Kammoun Jarraya, Manar Ali Salamah, and Hanêne Ben-Abdallah. Fallfree: Multiple fall scenario dataset of cane users for monitoring applications using kinect. In *2017 13th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pages 327–333. IEEE, 2017.

[3] Edouard Auvinet, Caroline Rougier, Jean Meunier, Alain St-Arnaud, and Jacqueline Rousseau. Multiple cameras fall data set.

[4] Greet Baldewijns, Glen Debard, Gert Mertes, Bart Vanrumste, and Tom Croonenborghs. Bridging the gap between real-life data and simulated data by providing a highly realistic fall dataset for evaluating camera-based fall detection algorithms. *Healthcare technology letters*, 3(1):6–11, 2016.

[5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[6] Imen Charfi, Johel Miteran, Julien Dubois, Mohamed Atri, and Rached Tourki. Optimized spatio-temporal descriptors for real-time fall detection: comparison of support vector machine and adaboost-based classification. 22(4):041106. Publisher: SPIE.

[7] Zhongwei Cheng, Lei Qin, Yituo Ye, Qingming Huang, and Qi Tian. Human daily action analysis with multi-view and color-depth data. In *Computer Vision–ECCV 2012. Workshops and Demonstrations: Florence, Italy, October 7-13, 2012, Proceedings, Part II 12*, pages 52–61. Springer, 2012.

[8] Jia-Luen Chua, Yoong Choon Chang, and Wee Keong Lim. A simple vision-based fall detection technique for indoor video surveillance. *Signal, Image and Video Processing*, 9:623–633, 2015.

[9] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019.

[10] Stefan Denkovski, Shehroz S. Khan, Brandon Malamis, Sae Young Moon, Bing Ye, and Alex Mihailidis. Multi visual modality fall detection dataset. 10:106422–106435. Conference Name: IEEE Access.

[11] Dave Epstein, Boyuan Chen, and Carl. Vondrick. Oops! predicting unintentional action in video. *arXiv preprint arXiv:1911.11206*, 2019.

[12] Yazan Abu Farha and Jurgen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3575–3584, 2019.

[13] Jane Fleming and Carol Brayne. Inability to get up after falling, subsequent time on floor, and summoning help: prospective cohort study in people over 90. *Bmj*, 337, 2008.

[14] José Camilo Eraso Guerrero, Elena Muñoz España, Mariela Muñoz Añasco, and Jesús Emilio Pinto Lopera. Dataset for human fall recognition in an uncontrolled environment. 45:108610.

[15] Bogdan Kwolek and Michal Kepski. Human fall detection on embedded platform using depth maps and wireless accelerometer. 117(3):489–501.

[16] Saturnino Maldonado-Bascon, Cristian Iglesias-Iglesias, Pilar Martín-Martín, and Sergio Lafuente-Arroyo. Fallen people detection capabilities using assistive robot. *Electronics*, 8(9):915, 2019.

[17] Lourdes Martínez-Villaseñor, Hiram Ponce, Jorge Brieva, Ernesto Moya-Albor, José Núñez-Martínez, and Carlos Peñafort-Asturiano. UP-fall detection dataset: A multimodal approach. 19(9):1988. Number: 9 Publisher: Multidisciplinary Digital Publishing Institute.

[18] Saquib Sarfraz, Marios Koulakis, Constantin Seibold, and Rainer Stiefelhagen. Hierarchical nearest neighbor graph embedding for efficient dimensionality reduction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 336–345, 2022.

[19] Mary E Tinetti, Wen-Liang Liu, and Elizabeth B Claus. Predictors and prognosis of inability to get up after falls among elderly persons. *Jama*, 269(1):65–70, 1993.

[20] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

[21] Thanh-Hai Tran, Thi-Lan Le, Dinh-Tan Pham, Van-Nam Hoang, Van-Minh Khong, Quoc-Toan Tran, Thai-Son Nguyen, and Cuong Pham. A multi-modal multi-view dataset for human fall analysis and preliminary investigation on modality. In *2018 24th International Conference on*

Pattern Recognition (ICPR), pages 1947–1952. ISSN: 1051-4651.

[22] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[23] World Health Organization. Falls. `https://web.archive.org/web/20250423060144/https://www.who.int/news-room/fact-sheets/detail/falls/`, 2021. Fact sheet, last updated 26 April 2021. Estimates 684 000 fatal falls and 37.3 million falls requiring medical attention each year.

[24] Zhong Zhang, Christopher Conly, and Vassilis Athitsos. Evaluating depth-based computer vision methods for fall detection under occlusions. In George Bebis, Richard Boyle, Bahram Parvin, Darko Koracin, Ryan McMahan, Jason Jerald, Hui Zhang, Steven M. Drucker, Chandra Kambhamettu, Maha El Choubassi, Zhigang Deng, and Mark Carlson, editors, *Advances in Visual Computing*, volume 8888, pages 196–207. Springer International Publishing. Series Title: Lecture Notes in Computer Science.