SUPPLEMENTARY MATERIAL: WEAR: AN OUTDOOR SPORTS DATASET FOR WEARABLE AND EGOCENTRIC ACTIVITY RECOGNITION

Anonymous authors
Paper under double-blind review

A DATASET OVERVIEW AND CONTENTS

The outdoor sports dataset WEAR features data of 18 participants performing each a total of 18 different workout activities with untrimmed inertial (acceleration) and camera (egocentric video) data recorded at 10 different outside locations. It provides a challenging prediction scenario marked by purposely introduced activity variations and an overall small information overlap across modalities. Figure 1 provides a dataset nutrition label inspired by Holland et al. (2018) in a table-like manner.

WEAR Dataset Key Facts

Motivation	An outdoor sports dataset dataset (egocentric-video & inertial data) with small information overlap across modalities
Example	Inertial-based, vision-based & multimodal
Use Cases	Human Activity Recognition
Authors	[Anonymized]

Meta information

Dataset

Locations	10
Activities	18 workout activities + NULL-class
Action segments	615
Total duration	908 min
Per-participant	50.5 ± 10.5 min

Subjects

Count	18 (10 male, 8 female)
Age	28 ± 5
Height	175.4 ± 10.8 cm
Weight	69.26 ± 12.43 kg

Modalities

3D-Acceleration

Sensor	Bangle.js Version 1
Settings	50 Hz (± 8g)
Format	.csv
Placement	Ankles & Wrists
Size	589.2 MB

Egocentric Vision

Sensor	GoPro Hero 8
Settings	1080p 60 FPS; SuperView FOV
Format	.mp4
Placement	Head (tilted 25° downwards)
Size	137.58 GB

Figure 1: Dataset nutrition label of the WEAR dataset. The dataset nutrition label was originally proposed by Holland et al. (2018). Our adaptation is inspired by DelPreto et al. (2022).

A.1 Intended Uses and Ethical Considerations

Before participating in the study, participants were notified that by nature the data they provide can only be pseudonymised. This means that, though requiring a substantial amount of effort, the identity of a person can be reconstructed. Although participants agreed to include their egocentric videos in a public dataset, it is essential to refrain from actively identifying the individuals featured in the WEAR dataset. If other researchers decide to contribute to the WEAR dataset by recording additional participants, societal and ethical implications should be considered. As with the participants part of the original release of the WEAR dataset, all participants must be briefed before their first recording, making them aware of all necessary information and implications that come with providing to the WEAR dataset. Recording locations should only be chosen if video recordings are allowed at said location and participants are given enough space to perform each activity safely. If the recording location involves pedestrians walking within close proximity, pedestrians should be notified that they are being recorded and, if applicable, captured faces should be blurred during postprocessing.

The WEAR dataset and associated code are made public for research purposes. With the accurate detection of physical activities that we perform in our daily lives having been identified as valuable information, the WEAR dataset focuses on one of the most popular application scenarios of wearable smartwatches and action cameras, i.e. self-tracking of workout activities. With the ease of reproducability we hope to make WEAR a collaborative, expanding dataset which researchers from different locations and backgrounds can contribute to. For example, as the current selection of participants is biased towards healthy, young people, we hope to overcome said limitation by including people from more diverse backgrounds and age groups in future iterations of the dataset.

Lastly, the authors took great care of avoiding any infringement of rights during the data collection process. Yet, in case of conflicts, they are of course committed to taking appropriate actions, such as promptly removing data associated with such concerns.

A.2 Data Availability and Licensing

WEAR and all associated files are offered under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. The dataset is hosted via the cloud-storage platform [Anonymized], which is a service hosted by [anonymized] (`https://www.anonymous.edu/anon`). It is a non-commercial cloud storage service for research, studying and teaching and is provided to participating institutions exclusively. With locations exclusively in [anonymized], [anonymized] is subject to the strict [anonymized] directives on data protection and data security. The complete dataset can be downloaded via [anonyimized] (`https://www.anonymous.edu/anon`). The dataset download is structured into the (1) '.json'-formatted annotations, (2) raw, synchronized inertial and vision data and (3) precomputed feature embeddings as mentioned in the main paper. Third party data-hosting services will be explored once the dataset paper is published and in a non-changing state. We will involve the ethics council of [anonymized] during our decision process to ensure a each selected hosting platform is inline with our data privacy standards.

The source code that was used to conduct all experiments is available via [Anonymized] (`https://www.anonymous.edu/anon`). A snapshot of the code is provided as part of the supplementary material download. The repository is written in such a way that other architectures (both inertial- and vision-based) can be added in the future. The repository provides Readme files which give details on the overall structure of the repository, how collect additional data and how to set up an Anaconda environment with the needed packages to run experiments. Experiments are defined via '.json'-format configuration files which allow for easy sharing of used hyperparameter settings.

B Experimental Protocol

Table 1 gives an overview of the 18 activity classes featured in the WEAR dataset and provides number of coherent sequences as well as total duration per workout activity class. In order to properly explain participants the activities they needed to perform and give insights on the overall study design a recording plan (see Section D) was provided to participants prior to their first session. The recording plan details all necessary materials and is written in such a way that the can easily be reproduced by persons other than the authors. The plan further outlines the study protocol as well

Table 1: Overview of the activity classes featured in the WEAR dataset. Each activity is categorized into as either a running (R), flexibility (F) or strength (S) exercise. The total duration of each class is provided in minutes and averaged across all activities. The total duration of the null class, i.e. samples not belonging to any of the classes of interest, is provided. A detailed description of each activity can be found in the recording plan attached at the end of the supplementary material.

Label ID	Activity Class	Category	Action Segments	Total duration (in min)
1	jogging	R	28	32:30
2	jogging (rotating arms)	R	22	29:34
3	jogging (skipping)	R	34	29:11
4	jogging (sidesteps)	R	30	33:33
5	jogging (butt-kicks)	R	37	28:43
6	stretching (triceps)	F	25	29:13
7	stretching (lunging)	F	23	31:09
8	stretching (shoulders)	F	22	30:04
9	stretching (hamstrings)	F	23	30:50
10	stretching (lumbar rotation)	F	27	31:36
12	push-ups	S	57	27:33
13	push-ups (complex)	S	41	29:14
14	sit-ups	S	43	30:50
15	sit-ups (complex)	S	32	31:02
16	burpees	S	49	31:25
17	lunges	S	31	31:54
18	lunges (complex)	S	35	33:19
19	bench-dips	S	56	28:38
0	null	-	592	358:29

informs about any risks of harm, data collection, usage, anonymisation and publication, as well as how to revoke data usage rights at any point in the future. Besides a written description of each activity, the original document provides short video-clips of each activity, showing the correct execution of exercises. To avoid any misunderstandings, the participants further received a one-on-one session with the researchers being able to ask their questions about the plan and activities in it. Other than the used sensors for video and acceleration recording, the exercises only require a yoga mat and a chair (or similar items). Sessions can be recorded at any location outside as long as the privacy and safety of the participants as well as pedestrians is ensured.

B.1 Participant and Session Information

The location and the time of day at which the sessions were performed were not fixed and thus vary across subjects. As participants were allowed to split activities across more (or less) than two sessions, session counts vary across subjects. Table 2 provides information on all 10 recording locations that are part of the WEAR dataset. The table details general information such as surface conditions of the location as well as which direction the static camera seen in videos is facing. Table 3 provides supplementary information on all separate sessions contained in the dataset. For each session, we detail its overall length in minutes, the number of distinct activities performed by the participant, the location it was recorded at, the month and time of day it was recorded, as well as the overall weather conditions during the duration of the session.

After having completed all sessions, participants were asked to take part in a questionnaire which was used to gather vital information (gender, age, height and weight) as well as workout-specific questions, aiming towards assessing the overall fitness level and experience with the activities detailed in the study protocol. The workout-specific questions were:

1. How many workouts (longer than 15 min) do you usually do per week?
2. Which kind of workout do you usually do (cycling, team sport, gym, cardio, yoga etc.)?
3. How many activities that are part of the workout plan did you know in advance?
4. How many activities that are part of the workout plan do you perform regularly yourself as part of your own workouts?

Table 4 shows the answers to the questionnaire items for each participant. Note that, to protect the privacy of our study participants, we only asked for age, height and weight in ranges instead of exact values, and always provided the option to not answer the questions if preferred.

Table 2: Description of the 10 locations featured in the WEAR dataset. For each location we provide information on surface conditions, overall surroundings and direction the static camera is facing.

Location ID	Description
1	Meadow in proximity to a larger building. Area is surrounded by trees with from November on-wards, fallen leaves laying on the ground. Static camera faces North-West.
2	Parking lot in proximity to building. Concrete surface. Static camera faces West.
3	Small square with concrete surface. Surrounded by bushes and buildings. Static camera faces West.
4	Meadow enclosed by bungalow-style living quarters. Static camera faces North-East
5	Covered walkway next to a building. Concrete surface. Walkway enclosed by building and bushes. Static camera faces North.
6	Football field with ash surface build behind a supermarket next a road and crop-fields. Long side of the football field is surrounded by bushes. Static camera faces mostly North-West.
7	Backyard in an urban-village with both concrete and grass surface. Terrace has a garden table and chairs standing around. Static camera faces mostly West.
8	Parking lot next to allotments in a city-area. Static camera faces mostly North-East.
9	Meadow next to a building. Static camera faces South.
10	City park in a metropolitan area behind a city mall. Park is surrounded by buildings, a playing ground, football and basketball fields. Static camera faces mostly North.

B.2 Hardware Overview

In order to capture the accelerometer data, four open-source Bangle.js Version 1 smartwatches running a custom, open-source firmware (Van Laerhoven et al., 2022) were used. The Bangle.js Version 1 comes with a Nordic 64MHz nRF52832 ARM Cortex-M4 processor with Bluetooth LE, 64kB RAM, 512kB on-chip flash, 4MB external flash, a heart rate monitor, a 3D accelerometer and a 3D magnetometer. The raw 3D acceleration was captured at 50 Hz with a sensitivity of $\pm$ 8g. As outlined in the recording plan (see Section D), watches were placed by the researchers before each session in a predetermined orientation on the participants' limbs and ankles. Egocentric video data was captured using a GoPro Hero 8 action camera. The camera was mounted using a headstrap with the camera tilted downwards in a roughly 45 degree angle. The GoPro was set to record at 1080p with 60 frames using a *SuperView* FOV with *Hypersmooth 2.0* electronic image stabilization and *Auto Low-Light* correction turned on. As the recorded egocentric video of participants makes accurate ground truth annotations more difficult (due to e.g. participants not looking at the actions they perform), a second camera was placed on a tripod in the proximity to the participants. Using again a large FOV setting, the second camera was placed in a way such that as much area as possible was captured. To allow for even more freedom of movement, participants were allowed to move out of the FOV of the second camera, but were asked to start and end their activities within the camera's FOV. This allowed participants, especially during running exercises, to run straight distances and overall commence activities in a more natural way. To preserve the privacy of our participants, the second camera's video stream and all audio streams captured during the experiments are not part of the WEAR dataset.

B.3 Postprocessing and Annotation Process

The open-source firmware (Van Laerhoven et al., 2022) running on each Bangle.js smartwatch stores the lossless, delta-compressed inertial data in separate files on the internal memory of each watch. During post-processing, said compressed files were extracted, uncompressed and concatenated to a single CSV file per session. Being a common issue with accelerometers sampling at a high sampling rate, the Bangle.js smartwatch is not able to maintain an exact sampling rate of 50 Hz, with the true sampling rate being closer to 48 Hz with fluctuations ranging between $\pm$ 1 Hz. The firmware

Table 3: Per-session meta-information. We provide the individual session count, duration of each session, number of activities performed during the session, location ID (LID) the session was performed at, approximate time of the year and day and weather conditions during recording time. More detailed information on each location can be found in Table 2 using the location ID.

Subject	Session	Duration	# Activities	Month	Time-of-day	LID	Weather conditions
sbj_0	1	16:33:30	7	mid-Oct.	morning	1	sunny, ≈10°C
sbj_0	2	11:55:00	6	mid-Oct.	afternoon	1	partly-cloudy, ≈10°C
sbj_0	3	18:06:00	7	late-Oct.	afternoon	1	partly-cloudy, ≈20°C
sbj_1	1	20:20:00	9	late-Oct.	afternoon	1	sunny, ≈15°C
sbj_1	2	25:58:00	9	early-Nov.	afternoon	1	sunny, ≈10°C
sbj_2	1	32:24:00	9	early-Nov.	morning	1	sunny, ≈10°C
sbj_2	2	25:08:00	9	mid-Jan.	afternoon	2	cloudy, after rain, ≈0°C
sbj_2	3	01:52:00	1	mid-Feb.	afternoon	3	sunny, ≈5°C
sbj_3	1	33:34:00	10	mid-Nov.	afternoon	4	sunny, ≈5°C
sbj_3	2	25:52:00	6	mid-Nov.	afternoon	4	partly-cloudy, ≈10°C
sbj_3	3	06:24:00	2	mid-Nov.	afternoon	4	sunny, ≈10°C
sbj_3	4	03:41:00	2	late-Jan.	afternoon	5	cloudy, snowy, ≈-5°C
sbj_4	1	24:07:30	9	mid-Nov.	midday	1	foggy, cloudy, windy, ≈5°C
sbj_4	2	29:04:00	9	late-Nov.	afternoon	1	partly-cloudy, ≈5°C
sbj_5	1	19:48:30	9	mid-Nov.	afternoon	1	sunny, ≈10°C
sbj_5	2	16:02:00	9	end-Nov.	afternoon	1	cloudy, ≈5°C
sbj_6	1	23:52:00	10	end-Nov.	afternoon	1	foggy, ≈5°C
sbj_6	2	17:51:30	8	end-Jan.	morning	5	cloudy, snowy, ≈-5°C
sbj_7	1	22:48:00	9	late-Dec.	morning	6	partly-sunny, ≈10°C
sbj_7	2	24:45:00	9	late-Dec.	midday	6	partly-sunny, ≈10°C
sbj_8	1	20:00:00	9	late-Dec.	midday	6	partly-cloudy, ≈10°C
sbj_8	2	21:35:00	9	late-Jan.	afternoon	7	cloudy, ≈0°C
sbj_9	1	18:50:00	9	early-Jan.	afternoon	8	cloudy, ≈10°C
sbj_9	2	17:16:00	9	early-Jan.	afternoon	8	cloudy, ≈10°C
sbj_10	1	21:42:00	9	mid-Jan.	afternoon	5	rainy, windy, ≈5°C
sbj_10	2	21:04:00	9	early-Feb.	afternoon	5	rainy, windy, ≈5°C
sbj_10	3	23:39:00	9	mid-Feb.	afternoon	9, 3	sunny, cloudy, windy, ≈5°C
sbj_11	1	17:41:00	9	mid-Jan.	morning	5	cloudy, rainy, ≈5°C
sbj_11	2	19:21:00	9	mid-Jan.	midday	5	cloudy, rainy, ≈5°C
sbj_12	1	27:08:00	9	mid-Jan.	afternoon	5	cloudy, windy, ≈0°C
sbj_12	2	27:22:00	9	late-Feb.	afternoon	5	partly-sunny, windy, ≈0°C
sbj_13	1	30:08:00	9	mid-Jan.	afternoon	5, 3	sunny, ≈0°C
sbj_13	2	36:10:00	9	mid-Jan.	afternoon	5, 3	sunny, ≈0°C
sbj_14	1	22:18:00	9	mid-Jan.	afternoon	5, 3	sunny, ≈-5°C
sbj_14	2	31:03:00	9	mid-Jan.	afternoon	5, 3	cloudy, ≈-5°C
sbj_15	1	23:17:00	9	late-Jan.	afternoon	5, 3	cloudy, ≈0°C
sbj_15	2	20:06:00	9	late-Jan.	afternoon	5, 3	cloudy, ≈0°C
sbj_16	1	26:34:00	9	early-Feb.	midday	10	partly-sunny, ≈10°C
sbj_16	2	31:56:00	9	early-Feb.	midday	10	partly-sunny, ≈10°C
sbj_17	1	23:16:00	9	early-Feb.	afternoon	1	sunny, ≈0°C
sbj_17	2	28:15:00	9	early-Feb.	afternoon	3	sunny, ≈0°C

Table 4: Per subject answers to the questionnaire handed to participants after having completed all sessions. The questionnaire collected vital information (gender (G), left- or righthanded (L/R), age, height and weight) as well as workout-specific questions, i.e. frequency and type of private workouts and number of activities, part of the WEAR dataset, which were known in advance and regularly conducted in private workouts.

Subject	G	L/R	Age	Height	Weight	Private Workouts		Activities	
						Frequency	Type	Known	Regularly
sbj_0	M	R	≥40	180-189	70-79	5	Cycling	5	0
sbj_1	M	R	25-29	170-179	60-69	3	Hiking	11	0
sbj_2	M	R	25-29	180-189	80-89	5	Gym, Cardio	18	9
sbj_3	M	R	35-39	170-179	70-79	4-5	Gym, Basketball, Cardio	18	9
sbj_4	M	R	25-29	180-189	60-69	0	Table-tennis	18	0
sbj_5	F	R	30-34	160 -169	N/A	2-3	Freeletics	16	9
sbj_6	F	R	25-29	150-159	50-59	1	Gym	9	0
sbj_7	M	R	30-34	180-189	80-89	5	Gym, Cardio	18	5
sbj_8	F	R	25-29	170-179	60-69	2-3	Volleyball, Yoga	15	7
sbj_9	F	R	25-29	150-159	50-59	7	Gym, Bicycling, Cardio, Ballet	18	7
sbj_10	F	R	20-24	160-169	50-59	5	Gym, Dancing, Yoga	15	7
sbj_11	F	R	25-29	160-169	50-59	3	Volleyball, Cardio, Yoga	18	11
sbj_12	F	R	20-24	170-179	60-69	4	Gym	17	8
sbj_13	M	R	20-24	≥190	90-99	2	Gym, Cardio	16	8
sbj_14	M	R	30-34	170-179	80-89	0	N/A	11	2
sbj_15	F	L	25-29	180-189	60-69	8	Rowing, Gym, Cycling, Cardio	18	9
sbj_16	M	R	20-24	180-189	60-69	2-3	Gym	15	3
sbj_17	M	R	25-29	180-189	70-79	4	Badminton, Bouldering, Hiking	15	5

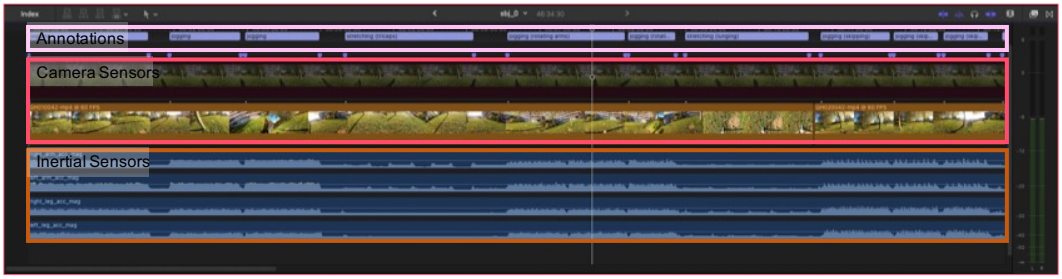


Figure 2: Snapshot along with descriptions of the annotation process using Final Cut Pro. Importing the converted video and inertial data (as '.wav'-files) allowed for an easy validation of the synchronization process. Labels were added via subtitles, exported as '.srt'-files and converted such that they can be appended to the respective '.csv'-files.

(Van Laerhoven et al., 2022) provides for each file a timestamp that was set by the on-board real-time clock, which allows correcting individual times of all delta-compressed samples. Therefore, in order to obtain the true sampling rate and correct the timestamps of the concatenated CSV-file, synchronisation jumps were performed by each participant at the start and end of each session. The synchronization jumps involved participants move in front of the tripod-mounted camera, stand still for approximately 10 seconds, jump three times while raising the arms while jumping and stand still for another 10 seconds. This allowed to map peaks in the inertial sensor streams to be mapped to points in the video stream and thus obtain a start and end point within both modality data streams. Lastly, assuming recorded inertial data records are equidistant, all records within the span of the start and end-point were evenly distributed across the experiment's duration and, as a final step, resampled to have a sampling rate of 50 Hz via linear interpolation. Similar to the inertial data, the video data recorded by the head-mounted GoPro was not recording at a true frame rate of 60 FPS, but slightly deviated from that (i.e. 59.94 FPS). We therefore also resampled the egocentric videos to be of a frame rate of 60 FPS.

In order to validate our synchronization process we made use of the similarities between sensor and audio data and converted each axis of the 3D accelerometer as well as their combined magnitude to four separate WAV-files. This approach is inspired by the works of Scholl et al. (2019) and Morshed et al. (2022). We calculated the magnitude as the summed norm of each individual inertial sensor channels, i.e. $\sqrt{x^2 + y^2 + z^2}$ with x, y and z being the x-, y- and z-axis of the 3D accelerometer data. Having converted the CSV data to WAV files allowed us to import both video data and inertial data into a standard video editing tool, in our case we used Final Cut Pro (see Figure 2). The user interface of Final Cut offers to see previews of sound files being in our case equivalent to a graph-like visualization of the acceleration data. This feature enabled us to have a visualised data stream of all modalities simultaneously while annotating. On average, the combined magnitude proved to be most useful when verifying the correctness of our synchronization across time. Labels of the activities were added by a single expert annotator as subtitles in SRT-format. A final script then converted the exported SRT-file to CSV-format, filling gaps within the subtitles with a *NULL* label and appended this to the respective final inertial sensor data CSV-file.

C Supplementary Experiments and Figures

C.1 Attend-and-Discriminate Improvements

Instead of employing a plain Attend-and-Discriminate model as proposed by Abedin et al. (2021), we incorporate architecture improvements suggested by Bock et al. (2021). Said architecture improvements are (1) using one instead of two recurrent layers, (2) increasing the amount of hidden units in the recurrent layer from 128 to 1024 and (3) scaling the convolutional kernel by the same factor the window size increases or decreases. Table 5 shows performance difference gained from employing the improved Attend-and-Discriminate architecture by comparing it to the original architecture. Note that results were obtained using longer training times along with a learning rate

Table 5: Results demonstrating the effectiveness of made modifications to the Attend-and-Discriminate model (Abedin et al., 2021). We compare the plain original model with an optimised version (1-layered LSTM with 1024 hidden units and an adjusted convolutional kernel sizes). We report results on the three employed window sizes (0.5, 1.0 and 2.0 seconds) each with a 50% overlap. Note that results are reported with no postprocessing applied.

	Model	P	R	F1	mAP					
					0.3	0.4	0.5	0.6	0.7	Avg
0.5s	Original A-and-D	71.87	72.78	71.63	1.86	1.54	1.35	1.07	0.87	1.34
	Optimised A-and-D	76.29	69.08	71.60	1.69	1.14	0.83	0.63	0.48	0.96
1.0s	Original A-and-D	72.37	72.38	71.60	3.07	2.46	1.96	1.49	1.31	2.06
	Optimised A-and-D	78.90	73.25	75.22	4.35	3.38	2.76	2.22	1.76	2.90
2.0s	Original A-and-D	74.48	73.99	73.26	8.75	7.1	5.94	4.85	3.95	6.12
	Optimised A-and-D	81.13	76.47	77.90	11.13	9.35	7.42	6.04	5.17	7.82

Table 6: Results demonstrating the effectiveness of longer training times on the inertial-based models. Compared are the shallow DeepConvLSTM (Bock et al., 2021) and improved Attend-and-Discriminate (Abedin et al., 2021) model using either a short training time (30 epochs and no step-wise learning rate schedule (LRS)) or long training time (300 epochs and LRS). We report results on the three employed window sizes (0.5, 1.0 and 2.0 seconds) each with a 50% overlap. Note that results are reported with no postprocessing applied.

	Model	Epochs	LRS	P	R	F1	mAP					
							0.3	0.4	0.5	0.6	0.7	Avg
0.5s	Shallow D.	30		70.51	72.92	70.71	2.13	1.82	1.55	1.33	1.22	1.61
	Shallow D.	300	✓	77.29	69.13	71.91	2.50	1.97	1.65	1.37	1.16	1.73
	A-and-D	30		72.15	71.87	71.24	1.97	1.61	1.33	1.04	0.84	1.36
	A-and-D	300	✓	76.29	69.08	71.60	1.69	1.14	0.83	0.63	0.48	0.96
1.0s	Shallow D.	30		73.35	76.25	73.78	4.83	4.01	3.38	2.81	2.32	3.47
	Shallow D.	300	✓	81.09	72.07	75.29	5.71	4.50	3.66	2.77	2.50	3.83
	A-and-D	30		74.00	74.96	73.70	4.48	3.47	3.00	2.35	2.01	3.06
	A-and-D	300	✓	78.90	73.25	75.22	4.35	3.38	2.76	2.22	1.76	2.90
2.0s	Shallow D.	30		74.97	78.21	75.63	11.68	10.44	8.71	7.8	6.42	9.01
	Shallow D.	300	✓	82.95	74.63	77.60	13.24	11.1	8.79	7.77	6.77	9.53
	A-and-D	30		77.04	79.01	77.29	10.55	8.74	7.21	6.17	5.08	7.55
	A-and-D	300	✓	81.13	76.47	77.90	11.13	9.35	7.42	6.04	5.17	7.82

schedule (see Section C.2 for more details) and are reported without having applied any postprocessing.

C.2 Longer vs. Shorter Training Runs

As mentioned in the main paper, all inertial-based architectures are trained for 300 epochs as compared to 30 epochs reported in Bock et al. (2021). These longer training times are inspired by the training reported in Abedin et al. (2021). To compensate for longer training times we employ a step-wise learning rate schedule as seen in Abedin et al. (2021) with a step size of 10 epochs and a decay rate of 0.9. Table 6 shows the improvement gained from using such longer training times by comparing it to a shorter training time of 30 epochs.

C.3 Ablation Study on Postprocessing

The following details ablation experiments conducted to demonstrate the effectiveness and validity of the applied postprocessing described in the experiments section of the main paper.

Figure 3 illustrates the effect the majority vote filter has on the prediction stream of the inertial-based models. One can see that without applying a majority vote filter, inertial-based architectures produce a large amount of non-coherent segments. This is due to the fact that during training, inertial models

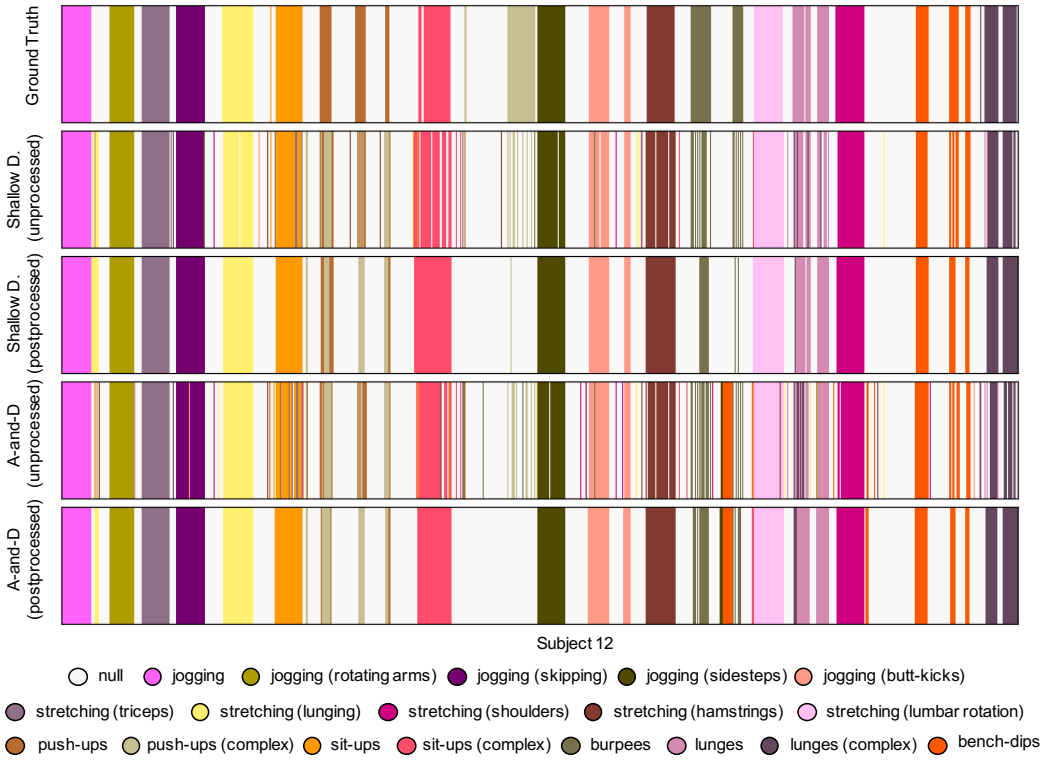


Figure 3: Color-coded comparison of the ground truth data (top row) with the raw and postprocessed (15 sec majority vote filter) activity streams of the shallow DeepConvLSTM (Bock et al., 2021) and Attend-and-Discriminate (Abedin et al., 2021) (A-and-D) model. The illustrated activity stream is of a sample subject having trained using inertial data which is windowed using a 1 second sliding window with 50% overlap.

such as Bock et al. (2021) and Abedin et al. (2021) are not explicitly trained to predict coherent segments, but rather predict a continuous stream of windowed data. The models therefore tend to show a lot of intermediate switches in-between activity labels which causes mAP scores of inertial-based architectures to be substantially lower than scores of vision-based models. We therefore make use of a majority vote filter to erase short activity-label switches. Table 7 and 8 shows experimental results of applying different-sized majority vote filters (5, 10, 15, 20 and 25 seconds) compared to applying no filter. Interestingly, results (see Table 7 and 8) not only demonstrate the effectiveness of the majority vote filter through a substantial increase in mAP scores, yet also show that said increase does not come at the cost of a decreased F1-score, but rather an increase. Table 7 and 8 further show a majority vote filter of 15 seconds being most effective resulting in the highest F1-score.

Temporal action localization models such as the ActionFormer (Zhang et al., 2022) and TriDet architecture (Shi et al., 2023) are not trained on an explicitly modelled NULL-class. This means, that unlike Bock et al. (2021) and Abedin et al. (2021), models are only able to predict segments with activity labels other than the NULL-class. With both models being set to predict up to 2000 action segments per video, the unprocessed prediction results resulted in activity streams such as illustrated in Figure 4. One can see that almost all samples have been assigned an activity label, leaving only a few data to be predicted as NULL, ultimately resulting in a substantially lower NULL-class accuracy than compared to inertial-based models mentioned in this paper. We therefore increased the score-threshold of both the ActionFormer and TriDet model, eliminating low-scoring segments and replacing them with NULL (see Figure 4). This improved classification performance of the ActionFormer (see Table 9) and TriDet model (see Table 10) significantly across all experiments (i.e. using inertial, vision and a combined setup as input data), while only marginally decreasing

Table 7: Ablation experiments on the effect of different-sized majority vote (MV) filters (5, 10, 15, 20 and 25 seconds) on the raw prediction results (0 seconds) of the shallow DeepConvLSTM model (Bock et al., 2021). We report results on the three employed window sizes (0.5, 1.0 and 2.0 seconds) each with a 50% overlap. Best results per clip-length are in **bold**.

	MV filter	P	R	F1	mAP					
					0.3	0.4	0.5	0.6	0.7	Avg
0.5 window	0 sec	77.29	69.13	71.91	2.50	1.97	1.65	1.37	1.16	1.73
	5 sec	85.87	75.63	79.04	37.83	36.02	34.30	32.48	29.89	34.10
	10 sec	86.63	**75.81**	**79.38**	50.43	47.92	46.17	43.86	41.94	46.06
	15 sec	**86.77**	75.42	79.18	54.36	51.67	49.42	47.40	44.70	49.51
	20 sec	86.72	74.71	78.66	56.90	53.97	51.65	49.06	**46.25**	51.57
	25 sec	86.56	73.89	78.02	**57.22**	**54.16**	**52.20**	**49.56**	45.89	**51.81**
1.0 window	0 sec	81.09	72.07	75.29	5.71	4.50	3.66	2.77	2.50	3.83
	5 sec	87.27	77.21	80.73	42.92	40.25	38.10	35.38	32.51	37.83
	10 sec	87.87	**77.35**	**81.05**	52.02	49.72	47.21	44.42	41.94	47.06
	15 sec	**88.02**	77.03	80.86	57.09	55.32	53.61	50.59	47.85	52.89
	20 sec	87.98	76.44	80.44	59.24	57.28	55.49	52.17	50.07	54.85
	25 sec	87.74	75.81	79.93	**61.50**	**59.63**	**57.41**	**53.88**	**51.13**	**56.71**
2.0 window	0 sec	82.95	74.63	77.60	13.24	11.10	8.79	7.77	6.77	9.53
	5 sec	86.92	77.88	81.08	42.44	40.53	37.92	35.18	32.81	37.78
	10 sec	87.80	**78.37**	**81.71**	55.51	52.62	49.75	47.09	44.87	49.97
	15 sec	**87.92**	78.16	81.60	59.89	57.00	54.69	51.77	48.99	54.47
	20 sec	87.90	77.74	81.32	61.04	58.99	57.05	53.31	50.49	56.18
	25 sec	87.70	77.22	80.89	**62.35**	**60.18**	**57.95**	**54.64**	**50.97**	**57.22**

Table 8: Ablation experiments on the effect of different-sized majority vote (MV) filters (5, 10, 15, 20 and 25 seconds) on the raw prediction results (0 seconds) of the improved Attend-and-Discriminate model (Abedin et al., 2021). We report results on the three employed window sizes (0.5, 1.0 and 2.0 seconds) each with a 50% overlap. Best results per clip length are in **bold**.

	MV filter	P	R	F1	mAP					
					0.3	0.4	0.5	0.6	0.7	Avg
0.5 window	0 sec	76.29	69.08	71.60	1.69	1.14	0.83	0.63	0.48	0.96
	5 sec	86.18	76.16	79.40	36.38	34.07	31.09	28.14	26.25	31.19
	10 sec	87.25	**76.31**	**79.78**	49.15	46.28	43.86	41.41	39.61	44.06
	15 sec	87.54	75.98	79.59	53.57	51.08	48.51	45.82	42.87	48.37
	20 sec	**87.61**	75.38	79.20	56.13	53.42	50.90	47.50	44.76	50.54
	25 sec	87.40	74.53	78.52	**59.00**	**55.82**	**53.45**	**49.49**	**45.90**	**52.73**
1.0 window	0 sec	78.90	73.25	75.22	4.35	3.38	2.76	2.22	1.76	2.90
	5 sec	86.58	78.95	81.56	40.43	37.88	35.03	32.22	29.61	35.03
	10 sec	87.61	**79.24**	**82.09**	51.81	49.59	47.73	44.88	41.55	47.11
	15 sec	87.87	79.02	82.01	56.38	54.47	52.28	50.07	46.92	52.03
	20 sec	**87.94**	78.59	81.74	57.80	57.80	55.88	52.95	49.58	55.19
	25 sec	87.82	77.92	81.23	**61.65**	**59.71**	**58.10**	**54.85**	**51.44**	**57.15**
2.0 window	0 sec	81.13	76.47	77.90	11.13	9.35	7.42	6.04	5.17	7.82
	5 sec	86.57	80.10	82.22	38.81	36.58	33.69	31.05	28.98	33.82
	10 sec	87.91	**80.71**	83.06	52.89	50.98	48.32	45.34	42.49	48.00
	15 sec	88.24	80.55	**83.08**	58.32	56.68	54.44	51.58	48.34	53.87
	20 sec	**88.37**	80.22	82.89	61.18	59.97	57.99	54.69	51.07	56.98
	25 sec	88.24	79.76	82.51	**62.83**	**61.06**	**58.96**	**56.20**	**52.87**	**58.38**

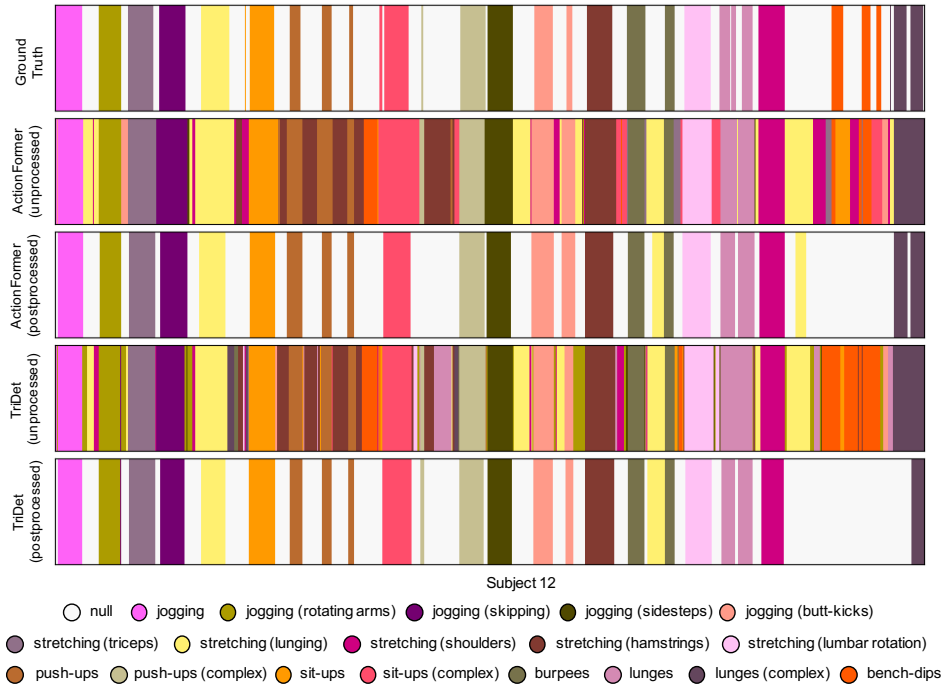


Figure 4: Color-coded comparison of the ground truth data (top row) with the raw and score-thresholded (0.2) activity streams of the TriDet (Shi et al., 2023) model. The illustrated activity stream is of sample subject having trained the model using both inertial and vision data which is windowed using a 1 second sliding window with 50% overlap.

mAP scores. Table 10 further shows 0.2 being the most effective identified threshold of our ablation study, resulting in the highest F1-score of the temporal action localization models.

Table 9: ActionFormer score thresholding results (Zhang et al., 2022) ablation experiments on the effect of different score thresholds (0.05, 0.1, 0.15, 0.2 and 0.25) on the raw prediction results (0.0 threshold) of experiments involving the ActionFormer model. We report results on the ActionFormer being applied to only inertial, camera and a combined (inertial + camera) features using three clip length window sizes (0.5, 1.0 and 2.0 seconds) each with a 50% overlap. Best results per modality are in **bold**.

	Threshold	CL	P	R	F1	mAP					
						0.3	0.4	0.5	0.6	0.7	Avg
Inertial	0.0	0.5s	55.65	77.50	61.41	73.55	70.70	62.51	48.14	32.02	57.38
	0.05	0.5s	65.11	78.53	67.78	72.61	69.73	61.60	47.24	31.25	56.49
	0.1	0.5s	71.15	77.77	72.15	70.28	67.59	59.74	45.25	29.67	54.51
	0.15	0.5s	76.27	75.18	73.96	67.45	64.89	57.02	42.55	28.10	52.00
	0.2	0.5s	78.73	70.50	72.51	63.71	61.28	53.90	39.81	26.40	49.02
	0.25	0.5s	81.76	64.25	69.46	59.09	56.93	49.66	36.17	24.36	45.24
	0.0	1.0s	58.46	78.94	61.91	**80.02**	**78.14**	**74.28**	**69.19**	**61.32**	**72.59**
	0.05	1.0s	67.40	**80.21**	70.60	79.24	77.40	73.55	68.45	60.59	71.85
	0.1	1.0s	74.00	79.14	74.72	77.63	75.80	72.05	67.14	59.34	70.39
	0.15	1.0s	78.82	77.21	76.41	75.15	73.46	70.03	65.55	57.90	68.42
	0.2	1.0s	81.69	75.37	**76.86**	72.90	71.30	68.28	64.14	56.65	66.65
	0.25	1.0s	**84.12**	73.38	**76.86**	70.25	69.01	66.15	62.49	55.35	64.65
	0.0	2.0s	54.47	74.61	57.84	74.85	71.16	67.88	63.67	56.53	66.82
	0.05	2.0s	61.67	75.41	64.98	73.92	70.13	66.81	62.62	55.69	65.84
	0.1	2.0s	68.67	73.72	68.95	71.70	68.20	65.00	61.01	54.27	64.04
	0.15	2.0s	74.59	71.78	71.00	69.22	66.05	63.02	59.22	52.51	62.00
	0.2	2.0s	78.18	69.15	71.15	66.43	63.30	60.47	56.66	50.26	59.43
	0.25	2.0s	80.99	66.93	70.90	64.11	61.06	58.31	54.73	48.47	57.34
Camera	0.0	0.5s	49.81	70.46	54.24	67.44	65.10	59.96	47.89	31.61	54.40
	0.05	0.5s	60.74	71.84	61.94	65.69	63.30	58.37	46.42	30.48	52.85
	0.1	0.5s	64.66	69.02	63.90	61.28	59.04	54.71	43.57	28.40	49.40
	0.15	0.5s	66.39	63.73	62.19	56.29	54.18	50.10	39.76	25.88	45.24
	0.2	0.5s	68.06	57.68	58.47	51.27	49.45	45.74	36.10	23.38	41.19
	0.25	0.5s	66.90	51.37	53.93	45.81	44.29	40.92	32.29	20.96	36.85
	0.0	1.0s	55.09	71.87	55.96	**74.07**	**72.05**	**69.54**	**65.81**	**59.28**	**68.15**
	0.05	1.0s	65.74	**73.82**	65.65	72.63	70.60	68.14	64.44	58.04	66.77
	0.1	1.0s	69.02	72.32	66.98	69.87	67.99	65.71	62.29	56.17	64.41
	0.15	1.0s	71.61	70.33	67.18	66.59	64.76	62.83	59.75	54.23	61.63
	0.2	1.0s	72.63	68.87	**67.26**	63.99	62.32	60.62	57.88	52.79	59.52
	0.25	1.0s	**73.27**	66.99	66.84	61.76	60.27	58.78	56.31	51.42	57.71
	0.0	2.0s	53.31	68.90	53.53	71.61	68.95	65.86	63.05	56.53	65.20
	0.05	2.0s	59.24	69.98	59.74	70.45	67.70	64.52	61.81	55.38	63.97
	0.1	2.0s	64.35	69.29	62.97	67.74	65.15	62.23	59.79	53.64	61.71
	0.15	2.0s	66.97	67.45	63.83	64.14	62.31	59.9	57.64	51.75	59.15
	0.2	2.0s	69.67	65.79	64.15	61.32	59.92	57.96	55.91	50.39	57.10
	0.25	2.0s	69.90	63.15	63.00	58.07	56.88	55.16	53.31	48.22	54.33
Inertial + Camera	0.0	0.5s	58.49	81.20	64.60	76.95	75.25	69.60	54.99	38.62	63.08
	0.05	0.5s	70.50	82.93	73.57	75.67	73.92	68.28	53.51	37.45	61.76
	0.1	0.5s	75.70	80.32	76.13	72.35	70.82	65.52	50.69	35.30	58.94
	0.15	0.5s	79.23	75.95	75.91	68.58	67.28	62.02	47.49	33.52	55.78
	0.2	0.5s	82.40	70.96	73.76	64.95	63.89	58.49	44.67	31.77	52.75
	0.25	0.5s	83.87	64.68	70.06	60.10	59.29	53.92	40.80	29.37	48.70
	0.0	1.0s	60.91	82.08	64.96	**84.41**	**82.67**	**79.73**	**76.01**	**68.01**	**78.16**
	0.05	1.0s	72.45	**83.75**	75.61	83.50	81.77	78.76	75.07	67.02	77.22
	0.1	1.0s	77.00	82.96	78.46	81.63	79.83	76.97	73.38	65.52	75.46
	0.15	1.0s	79.84	81.61	79.43	79.70	77.92	75.01	71.70	64.22	73.71
	0.2	1.0s	82.38	80.30	80.15	77.63	75.97	73.28	70.31	63.04	72.05
	0.25	1.0s	**84.48**	78.66	**80.20**	75.58	74.02	71.52	68.65	61.80	70.31
	0.0	2.0s	56.73	77.66	60.37	78.90	75.83	72.84	69.29	63.15	72.00
	0.05	2.0s	64.75	78.75	68.25	77.56	74.55	71.65	68.11	62.09	70.80
	0.1	2.0s	71.04	77.83	72.35	75.64	72.87	70.07	66.46	60.54	69.12
	0.15	2.0s	75.27	75.80	73.75	73.23	70.66	68.04	64.52	58.82	67.06
	0.2	2.0s	79.19	73.88	74.52	71.10	68.79	66.38	63.00	57.54	65.36
	0.25	2.0s	81.26	72.13	74.26	69.17	66.79	64.40	61.18	56.14	63.53

Table 10: TriDet score thresholding results (Shi et al., 2023) ablation experiments on the effect of different score thresholds (0.05, 0.1, 0.15, 0.2 and 0.25) on the raw prediction results (0.0 threshold) of experiments involving the TriDet model. We report results on the TriDet being applied to only inertial, camera and a combined (inertial + camera) features using three clip length sizes (0.5, 1.0 and 2.0 seconds) each with a 50% overlap. Best results per modality are in **bold**.

	Threshold	CL	P	R	F1	mAP					
						0.3	0.4	0.5	0.6	0.7	Avg
Inertial	0.0	0.5s	54.94	77.88	61.53	76.30	73.57	67.90	59.18	49.16	65.22
	0.05	0.5s	68.51	**79.26**	70.92	75.39	72.72	67.04	58.34	48.35	64.37
	0.1	0.5s	77.48	78.04	76.26	73.34	70.84	65.04	56.14	46.43	62.36
	0.15	0.5s	82.56	75.02	**77.19**	70.28	67.93	62.10	53.15	43.92	59.48
	0.2	0.5s	86.06	70.10	75.25	66.01	63.71	57.70	49.30	41.09	55.56
	0.25	0.5s	**87.78**	63.97	71.09	60.73	58.58	52.94	45.06	37.85	51.03
	0.0	1.0s	55.34	78.01	60.87	b	**78.45**	**76.11**	**72.94**	**67.48**	**75.03**
	0.05	1.0s	66.81	79.22	70.05	79.42	77.66	75.28	72.22	66.73	74.26
	0.1	1.0s	75.83	77.89	75.28	77.92	76.26	73.95	70.86	65.47	72.89
	0.15	1.0s	80.70	75.94	76.91	75.97	74.38	72.16	69.09	64.04	71.13
	0.2	1.0s	83.85	73.76	77.12	73.27	71.66	69.83	66.79	62.25	68.76
	0.25	1.0s	85.73	71.77	76.59	70.96	69.39	67.51	64.72	60.43	66.60
	0.0	2.0s	50.57	75.56	58.19	74.94	72.67	70.35	67.05	61.67	69.33
	0.05	2.0s	63.35	75.93	66.64	73.77	71.50	69.14	66.04	60.82	68.26
	0.1	2.0s	71.97	74.06	71.04	71.23	68.99	66.91	63.87	59.06	66.01
	0.15	2.0s	77.71	71.69	72.47	68.04	65.97	64.04	61.08	56.62	63.15
	0.2	2.0s	81.72	69.37	72.53	65.57	63.65	61.86	59.07	54.82	60.99
	0.25	2.0s	84.13	67.14	71.99	63.01	61.13	59.28	56.78	52.98	58.64
Camera	0.0	0.5s	49.81	70.46	54.24	67.44	65.10	59.96	47.89	31.61	54.40
	0.05	0.5s	60.74	71.84	61.94	65.69	63.30	58.37	46.42	30.48	52.85
	0.1	0.5s	64.66	69.02	63.90	61.28	59.04	54.71	43.57	28.40	49.40
	0.15	0.5s	66.39	63.73	62.19	56.29	54.18	50.10	39.76	25.88	45.24
	0.2	0.5s	68.06	57.68	58.47	51.27	49.45	45.74	36.10	23.38	41.19
	0.25	0.5s	66.90	51.37	53.93	45.81	44.29	40.92	32.29	20.96	36.85
	0.0	1.0s	55.09	71.87	55.96	**74.07**	**72.05**	**69.54**	**65.81**	**59.28**	**68.15**
	0.05	1.0s	65.74	**73.82**	65.65	72.63	70.60	68.14	64.44	58.04	66.77
	0.1	1.0s	69.02	72.32	66.98	69.87	67.99	65.71	62.29	56.17	64.41
	0.15	1.0s	71.61	70.33	67.18	66.59	64.76	62.83	59.75	54.23	61.63
	0.2	1.0s	72.63	68.87	**67.26**	63.99	62.32	60.62	57.88	52.79	59.52
	0.25	1.0s	**73.27**	66.99	66.84	61.76	60.27	58.78	56.31	51.42	57.71
	0.0	2.0s	53.31	68.90	53.53	71.61	68.95	65.86	63.05	56.53	65.20
	0.05	2.0s	59.24	69.98	59.74	70.45	67.70	64.52	61.81	55.38	63.97
	0.1	2.0s	64.35	69.29	62.97	67.74	65.15	62.23	59.79	53.64	61.71
	0.15	2.0s	66.97	67.45	63.83	64.14	62.31	59.9	57.64	51.75	59.15
	0.2	2.0s	69.67	65.79	64.15	61.32	59.92	57.96	55.91	50.39	57.10
	0.25	2.0s	69.90	63.15	63.00	58.07	56.88	55.16	53.31	48.22	54.33
Inertial + Camera	0.0	0.5s	58.91	80.98	64.74	80.30	78.52	74.52	67.53	56.76	71.53
	0.05	0.5s	73.96	82.69	75.74	78.95	77.12	73.10	66.16	55.44	70.15
	0.1	0.5s	81.07	79.82	78.94	75.31	73.67	69.70	62.86	52.32	66.77
	0.15	0.5s	84.88	75.43	78.36	71.95	70.36	66.35	59.14	49.20	63.40
	0.2	0.5s	87.85	70.34	75.90	67.65	66.05	62.22	55.55	46.12	59.52
	0.25	0.5s	**88.98**	63.95	71.24	62.04	60.46	56.70	50.70	42.00	54.38
	0.0	1.0s	58.69	81.51	64.16	**84.95**	**83.77**	**82.05**	**79.49**	**74.19**	**80.89**
	0.05	1.0s	71.84	**83.39**	75.19	84.03	82.83	81.13	78.55	73.17	79.94
	0.1	1.0s	78.61	82.79	79.37	82.69	81.48	79.76	77.15	71.88	78.59
	0.15	1.0s	82.64	81.42	80.93	80.99	79.76	78.09	75.60	70.57	77.00
	0.2	1.0s	84.99	79.55	**81.08**	78.64	77.45	75.74	73.40	68.79	74.81
	0.25	1.0s	86.81	77.11	80.38	75.60	74.39	72.76	70.40	66.20	71.87
	0.0	2.0s	51.17	78.44	60.46	79.51	77.74	75.56	72.54	68.28	74.73
	0.05	2.0s	66.62	79.44	70.01	78.08	76.28	74.17	71.23	67.13	73.38
	0.1	2.0s	74.65	78.03	74.43	75.36	73.74	71.85	69.13	65.12	71.04
	0.15	2.0s	79.86	76.49	76.35	73.39	71.72	69.95	67.56	63.76	69.28
	0.2	2.0s	83.10	74.55	76.72	71.20	69.69	67.88	65.49	61.77	67.20
	0.25	2.0s	84.29	72.39	76.03	68.64	67.07	65.30	63.02	59.49	64.70

C.4 Single-Stage Temporal Action Localization for Inertial Data

In this paper we demonstrated the applicability of vision-based single-stage temporal action localization models on a previously unexplored modality, i.e. inertial data. As the investigated architectures, namely the TriDet (Shi et al., 2023) and ActionFormer (Zhang et al., 2022), both require clip-based, one-dimensional feature embeddings as input, data of both camera and inertial sensors had to be preprocessed. Figure 5 summarizes the applied preprocessing on both modalities. First step for both modalities included windowing the data streams using a predefined clip length and overlap. In total three different clip lengths were tested (0.5, 1 and 2 seconds). Having windowed the inertial data left us with a 3-dimensional feature array, i.e. $[no.\ windows, window\ length, no.\ sensor\ axis]$. In order to obtain a vectorized feature embedding per sliding window, individual sensor axis were concatenated. Depending on the window length this left us with a one-dimensional feature vector of size 300 (0.5 second), 600 (1 second) and 1200 (2 seconds) per video clip, i.e. sliding window. Contrarily, as also applied in Shi et al. (2023), we extracted two-stream I3D feature embeddings (Carreira & Zisserman, 2017) pretrained on Kinetics-400 (Kay et al., 2017) from the raw video stream, resulting in a vision-based embedding of size 2048 per video clip. Having vectorized both modalities we were able to train both temporal action localization architectures on either (1) inertial, (2) camera or (3) a concatenation of the two (inertial + camera). Even though our concatenation approach results in varying input dimensions, said change does not come at increased computational costs. More specifically, while amount of learnable parameters marginally increases (not more than 10%) with an increased input dimension, unlike other approaches, no additional embedding needs to be extracted from the inertial data and raw data streams can directly be used.

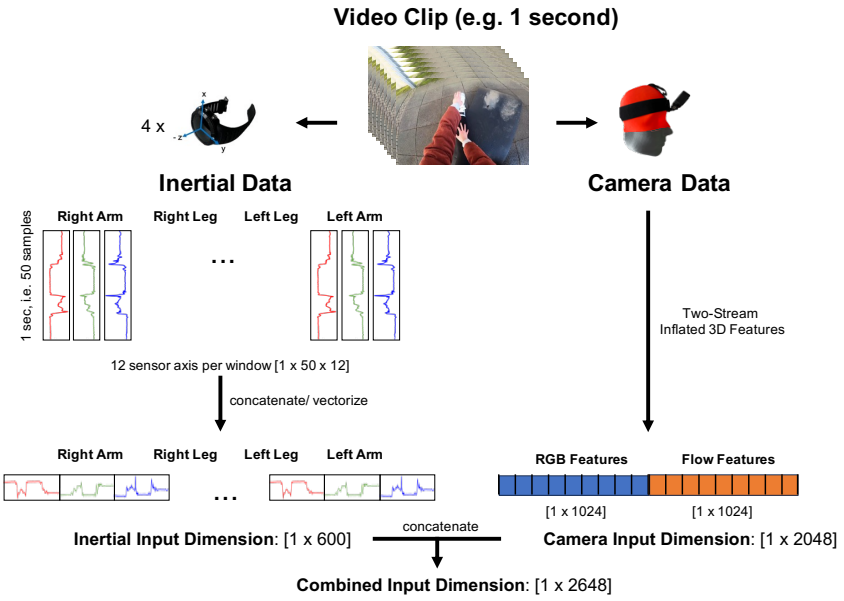


Figure 5: Visualization of the applied preprocessing on inertial and camera data in order to make to create a feature embedding which can be used to train the TriDet Shi et al. (2023) and ActionFormer Zhang et al. (2022) network.

C.5 Ablation Study on Influence of Frequency of Inputs

With the frequencies both the camera (60 FPS) and inertial sensors (50 HZ) being set fairly high, the WEAR dataset allows to explore lower frequency experiments and their effect fewer datapoints per second might have on the predictive quality of the trained models. Table 11 summarizes experiments conducted using only 50% and 20% of the available frequency for both types of sensors. Note that a clip length of 0.5 seconds was not explored during experiments as it was not possible anymore to extract two-stream I3D feature embeddings (Carreira & Zisserman, 2017) as the amount of frames was lower than the required minimum input frames. Looking at results presented in Table 11 one can

see that all models trained using only inertial data suffered from lower frequency inputs with both classification and mAP scores decreasing. Contrarily, models trained using camera-based improved when using features extracted from videos with a lower FPS, which might be caused by Kinetics-400 (Kay et al., 2017), which was used for pretraining the I3D extraction method, on consisting of videos with a lower FPS than the WEAR dataset.

C.6 Ablation Study on Inertial Sensor Selection

As reported experimental results are based on acceleration recordings of all limbs, the following experiments investigate how the predictive performance of each algorithm is affected by using only (1) acceleration recorded from the right wrist and (2) acceleration recorded from both the right wrist and right ankle. Results in Table 12 show that using only acceleration data obtained from the right wrist significantly decreases predictive performance across all algorithms across all metrics. Moreover, Table 13 clearly underlines the value of additionally measuring acceleration at the ankles' of participants, as results again significantly increase, being mostly on par compared to using all four inertial sensor locations. Interestingly, unlike the inertial-based architectures, results of vision-based models improve when excluding data captured by the left-wrist and left-ankle inertial sensors, which could be caused by the dataset being biased towards right-handed participants (see Table 4) and dominant hand movement might being overall more consistent. Figure 6 shows the per-class results of the TriDet model (Shi et al., 2023) being trained using (1) data obtained from the right wrist inertial sensor, (2) right wrist and right ankle inertial sensor and (3) all four inertial sensors (right and left wrists and ankles).

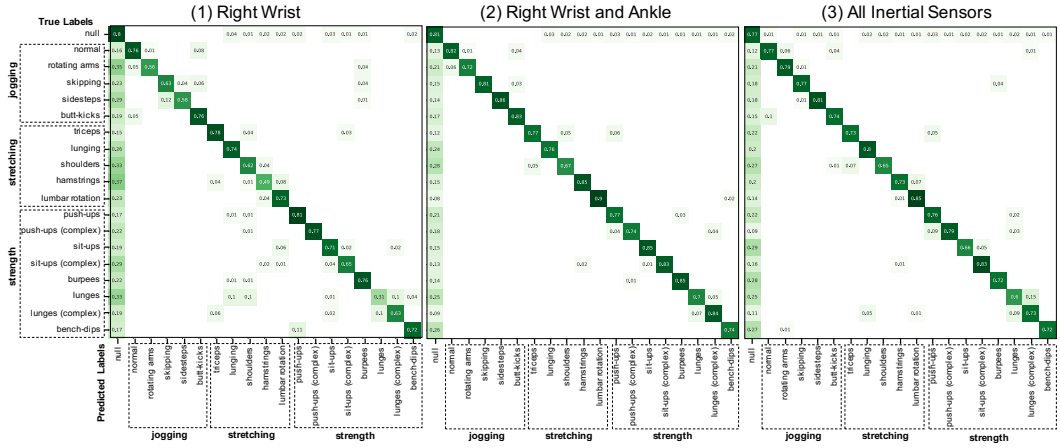


Figure 6: Confusion matrices of the TriDet model (Shi et al., 2023) being applied using only inertial obtained from the (1) right wrist, (2) right wrist and ankle and (3) right and left wrists and ankles with a one second sliding window and 50% overlap.

C.7 Ablation Study on Second Execution of Workout Sessions

In order to explore the robustness of obtained results, we recorded all activities of two participants (sbj_0 and sbj_14) a second time in August. Both participants recording conditions significantly differed from their first recording, with temperatures being around 25 degrees Celsius with overall more sunny weather conditions. Further, as not all participants knew all activities beforehand (see Table 4), recording the same participants a second time would allow to analyse how a certain degree of familiarity with the recording setup can be seen in altered movements (e.g., via a smoother execution of activities) as well as subject-specific finetuning affects the overall recognition performance. Table 14 compares validation results obtained on the original, first recording of sbj_0 and sbj_14 with their second execution of the workout plan. Unlike our prior experiments, each algorithm is trained using the data of all but the validation subjects' recordings, ensuring the validation subjects (sbj_0 and sbj_14) remain unseen during the training of each algorithm. All results are postprocessed as reported in the main paper. While, one can see improved results regarding sbj_0, this trend does not

Table 11: Results of evaluating different frequencies (Freq.) as input for different clip lengths (CL) on our WEAR dataset. Both inertial- and camera-based features were downsampled to be only 50% (i.e. 30 FPS and 25 Hz) and 20% (i.e. 12 FPS and 10 Hz) of the original frequency input (i.e. 60 FPS and 50 Hz). One can see that the predictive performance of inertial models decreases with a lower frequency input. Interestingly camera and combined models increase in performance when lower frequency inputs with I3D being calculated on 12 FPS videos resulting in the highest classification and mAP scores during camera-based experiments. Experiments are evaluated in terms of precision (P), recall (R), F1-score and mean average precision (mAP) for different temporal intersection over union (tIoU) thresholds. Best results per modality are in **bold**.

	Threshold	CL	P	R	F1	mAP						
						0.3	0.4	0.5	0.6	0.7	Avg	
Inertial	Shallow D.	Orig	1s	88.02	77.03	80.86	57.09	55.32	53.61	50.59	47.85	52.89
	Shallow D.	50%	1s	87.02	76.51	80.10	55.33	52.70	51.02	48.30	45.67	50.61
	Shallow D.	20%	1s	86.38	76.10	79.59	53.94	51.95	50.05	47.60	45.19	49.75
	A-and-D	Orig	1s	87.87	79.02	**82.01**	56.38	54.47	52.28	50.07	46.92	52.03
	A-and-D	50%	1s	87.57	78.25	81.23	55.88	53.51	51.33	48.16	44.78	50.73
	A-and-D	20%	1s	86.10	79.88	81.67	56.76	54.87	53.03	49.85	47.39	52.38
	ActionFormer	Orig	1s	81.69	75.37	76.86	72.90	71.30	68.28	64.14	56.65	66.65
	ActionFormer	50%	1s	80.93	73.66	75.62	71.40	69.69	66.77	63.09	56.01	65.39
	ActionFormer	20%	1s	80.73	72.43	74.51	70.07	68.34	65.44	60.83	54.45	63.82
	TriDet	Orig	1s	83.85	73.76	77.12	**73.27**	**71.66**	**69.83**	**66.79**	**62.25**	**68.76**
	TriDet	50%	1s	84.52	72.82	76.67	72.01	70.62	68.86	65.13	60.32	67.39
	TriDet	20%	1s	84.30	71.72	75.68	70.60	69.38	67.34	63.75	57.61	65.74
	Shallow D.	Orig	2s	87.92	78.16	81.60	59.89	57.00	54.69	51.77	48.99	54.47
	Shallow D.	50%	2s	85.08	76.13	79.09	53.85	51.57	49.30	46.50	43.79	49.00
	Shallow D.	20%	2s	84.59	75.39	78.33	53.21	50.98	48.66	45.69	43.08	48.32
	A-and-D	Orig	2s	**88.24**	**80.55**	83.08	58.32	56.68	54.44	51.58	48.34	53.87
	A-and-D	50%	2s	87.24	78.05	80.88	53.69	51.36	48.58	45.99	42.66	48.46
	A-and-D	20%	2s	86.98	77.94	80.83	55.63	53.32	49.76	46.45	43.74	49.78
	ActionFormer	Orig	2s	78.18	69.15	71.15	66.43	63.30	60.47	56.66	50.26	59.43
	ActionFormer	50%	2s	77.85	67.46	70.24	64.88	62.47	59.26	55.65	49.35	58.32
	ActionFormer	20%	2s	76.84	65.71	68.69	62.51	59.90	56.87	52.64	45.50	55.48
	TriDet	Orig	2s	81.72	69.37	72.53	65.57	63.65	61.86	59.07	54.82	60.99
	TriDet	50%	2s	80.61	65.51	69.77	62.32	60.60	58.20	55.90	51.50	57.71
	TriDet	20%	2s	78.59	63.55	67.79	59.87	58.24	56.09	53.33	47.92	55.09
Camera	ActionFormer	Orig	1s	72.63	**68.87**	67.26	63.99	62.32	60.62	57.88	52.79	59.52
	ActionFormer	50%	1s	74.62	68.58	67.81	64.61	63.12	61.28	58.56	53.77	60.27
	ActionFormer	20%	1s	74.36	67.82	67.92	65.92	64.36	62.98	59.99	55.31	61.71
	TriDet	Orig	1s	75.32	68.07	67.95	64.36	63.30	61.38	59.13	54.64	60.56
	TriDet	50%	1s	77.21	68.41	68.82	66.01	65.06	63.46	61.53	57.56	62.72
	TriDet	20%	1s	**75.95**	68.41	**69.10**	**66.61**	**65.71**	**63.72**	**61.85**	**57.66**	**63.11**
	ActionFormer	Orig	2s	69.67	65.79	64.15	61.32	59.92	57.96	55.91	50.39	57.10
	ActionFormer	50%	2s	72.64	67.53	66.43	64.22	62.25	60.65	57.71	52.93	59.55
	ActionFormer	20%	2s	71.94	65.39	65.62	61.94	59.93	58.10	54.62	49.88	56.89
	TriDet	Orig	2s	73.85	64.09	64.25	60.95	60.03	57.75	55.55	52.19	57.30
	TriDet	50%	2s	75.08	66.27	67.10	64.18	62.98	61.37	59.95	56.11	60.92
	TriDet	20%	2s	74.04	63.20	64.98	59.48	58.27	56.82	55.70	52.05	56.47
Inertial + Camera	ActionFormer	Orig	1s	82.38	**80.30**	80.15	77.63	75.97	73.28	70.31	63.04	72.05
	ActionFormer	50%	1s	82.04	80.24	79.84	76.98	75.34	73.35	69.60	63.07	71.67
	ActionFormer	20%	1s	81.89	79.33	79.24	76.24	74.92	72.95	70.43	63.08	71.52
	TriDet	Orig	1s	84.99	79.55	81.08	**78.64**	**77.45**	**75.74**	**73.40**	68.79	**74.81**
	TriDet	50%	1s	**86.25**	79.48	**81.46**	77.68	77.07	75.26	73.05	**68.94**	74.40
	TriDet	20%	1s	84.79	79.13	80.45	77.55	76.83	75.05	71.94	68.19	73.91
	ActionFormer	Orig	2s	79.19	73.88	74.52	71.10	68.79	66.38	63.00	57.54	65.36
	ActionFormer	50%	2s	79.24	75.14	75.55	70.84	68.13	65.71	62.77	57.33	64.96
	ActionFormer	20%	2s	78.47	73.78	74.15	68.60	66.69	63.24	60.14	55.78	62.89
	TriDet	Orig	2s	83.10	74.55	76.72	71.20	69.69	67.88	65.49	61.77	67.20
	TriDet	50%	2s	81.33	73.58	75.69	70.52	69.06	67.31	64.91	61.32	66.62
	TriDet	20%	2s	82.51	72.91	75.48	69.09	67.08	64.87	62.59	59.41	64.61

Table 12: Results of using only inertial features captured by the sensor placed on the right wrist for different clip lengths (CL) on our WEAR dataset evaluated in terms of precision (P), recall (R), F1-score and mean average precision (mAP) for different temporal intersection over union (tIoU) thresholds. One can see a clear overall decrease across all evaluation metrics. Best results per modality are in **bold**.

	Model	CL	P	R	F1	mAP					
						0.3	0.4	0.5	0.6	0.7	Avg
Inertial	Shallow D.	0.5s	64.54	68.08	64.23	23.26	21.57	19.27	17.31	16.00	19.48
	A-and-D	0.5s	75.34	64.09	66.93	27.08	25.33	22.55	20.53	18.94	22.89
	ActionFormer	0.5s	72.96	63.59	65.30	54.45	52.42	45.70	34.82	22.11	41.90
	TriDet	0.5s	**79.48**	62.89	66.98	54.32	52.10	47.57	40.39	30.38	44.95
	Shallow D.	1s	66.98	68.81	66.19	25.53	23.62	22.11	19.56	18.28	21.82
	A-and-D	1s	75.56	64.31	67.21	29.18	26.39	23.52	21.60	19.57	24.05
	ActionFormer	1s	73.07	65.51	66.91	61.00	58.05	52.69	47.32	39.82	51.78
	TriDet	1s	78.04	**67.88**	**70.44**	**63.08**	**62.09**	**60.07**	**57.07**	**50.36**	**58.54**
	Shallow D.	2s	66.79	67.68	65.34	28.34	26.46	24.05	21.58	19.40	23.97
	A-and-D	2s	76.71	65.87	68.63	31.93	28.51	25.86	23.46	21.31	26.21
	ActionFormer	2s	69.44	61.33	63.06	55.42	53.22	51.32	47.34	39.90	49.44
	TriDet	2s	70.73	58.22	61.08	52.06	50.54	48.51	46.05	40.93	47.62
I + C	ActionFormer	0.5s	76.91	65.69	67.69	58.60	57.36	51.30	38.48	26.15	46.38
	TriDet	0.5s	**81.87**	65.19	69.57	60.83	59.12	55.57	48.84	40.71	53.01
	ActionFormer	1s	79.62	**77.00**	**76.56**	72.06	70.65	68.94	66.26	60.49	67.68
	TriDet	1s	79.96	76.26	76.45	**74.39**	**73.55**	**71.84**	**69.52**	**65.88**	**71.03**
	ActionFormer	2s	74.43	73.48	72.01	68.87	66.86	64.51	60.95	55.89	63.42
	TriDet	2s	77.07	71.70	72.43	67.87	66.93	64.93	62.12	58.30	64.03

Table 13: Results using only inertial features captured by the sensor placed on the right wrist and right ankle for different clip lengths (CL) on our WEAR dataset evaluated in terms of precision (P), recall (R), F1-score and mean average precision (mAP) for different temporal intersection over union (tIoU) thresholds. Comparing results to 12 one can see the increase in performance one can achieve when tracking acceleration measured at the ankle in addition to a wrist-worn inertial sensor. Best results per modality are in **bold**.

	Model	CL	P	R	F1	mAP					
						0.3	0.4	0.5	0.6	0.7	Avg
Inertial	Shallow D.	0.5s	78.73	74.71	75.24	42.19	40.40	37.77	34.94	32.07	37.47
	A-and-D	0.5s	81.88	69.35	73.02	41.76	39.42	35.94	32.48	30.29	35.98
	ActionFormer	0.5s	78.62	74.39	74.47	68.28	64.52	54.70	39.06	25.95	50.50
	TriDet	0.5s	**84.83**	73.00	76.38	68.73	65.80	60.60	50.99	40.43	57.31
	Shallow D.	1s	80.63	74.87	76.32	44.49	42.87	40.70	37.12	34.95	40.03
	A-and-D	1s	82.83	72.72	75.77	43.75	41.17	38.26	34.65	32.20	38.00
	ActionFormer	1s	81.33	**78.60**	78.64	76.64	74.60	70.97	65.88	58.13	69.24
	TriDet	1s	84.03	78.16	**79.75**	**77.84**	**75.93**	**73.69**	**70.80**	**64.39**	**72.53**
	Shallow D.	2s	80.41	75.62	76.76	46.07	44.36	41.52	38.50	35.63	41.21
	A-and-D	2s	84.56	76.65	79.07	50.25	47.27	43.27	40.19	36.96	43.59
	ActionFormer	2s	77.72	73.34	73.63	70.06	67.29	64.45	60.00	52.62	62.88
	TriDet	2s	79.75	72.70	74.49	68.19	66.31	64.34	61.33	57.12	63.46
I + C	ActionFormer	0.5s	81.20	73.51	75.19	66.99	65.65	60.32	44.59	30.61	53.63
	TriDet	0.5s	86.97	71.16	75.78	67.41	65.68	61.54	53.14	43.45	58.24
	ActionFormer	1s	83.01	**82.35**	81.47	79.17	77.84	75.34	71.12	65.54	73.80
	TriDet	1s	**85.39**	81.59	**82.47**	**80.22**	**79.12**	**77.01**	**73.81**	**71.07**	**76.25**
	ActionFormer	2s	78.22	77.84	76.53	73.87	71.83	69.07	64.91	59.47	67.83
	TriDet	2s	80.93	77.67	77.90	73.47	72.01	70.12	68.21	64.43	69.65

Table 14: Comparison of obtained results of repeated sessions for participants sbj_0 and sbj_14 for different clip lengths (CL) on our WEAR dataset evaluated in terms of F1-score and mean average precision (mAP). The two participants were invited to perform the recording plan a second time. While one can see that improved results regarding sbj_0, suggesting potential learning effects of the correct execution of activities, this trend does not apply to sbj_14. Note that weather conditions (temperature and sunlight) significantly differ amongst the recordings – winter (first recording) compared to summer (2nd recording). These figures are, as in the earlier results, averaged across 3 runs using 3 different random seeds. For the first recording, both subjects' best results per modality are in underlined. For the second recording, both subjects' best results per modality are in **bold**. Unlike our prior experiments, each algorithm is trained using the data of all but the validation subjects' recordings, ensuring the validation subjects (sbj_0 and sbj_14) remain unseen during the training of each algorithm. All results are postprocessed as reported in the main paper.

	Model	CL	sbj_0				sbj_14			
			1st Recording		2nd Recording		1st Recording		2nd Recording	
			F1	mAP	F1	mAP	F1	mAP	F1	mAP
Inertial	Shallow D.	0.5s	69.75	42.18	85.15	75.20	77.52	65.18	77.60	62.40
	A-and-D	0.5s	73.51	38.02	84.10	70.60	79.09	59.26	75.88	61.70
	ActionFormer	0.5s	76.05	69.17	81.98	72.68	79.02	78.01	69.09	62.24
	TriDet	0.5s	74.84	67.57	80.07	74.97	81.59	85.67	72.29	67.90
	Shallow D.	1s	73.72	49.62	**85.46**	**76.47**	82.77	69.30	77.15	62.19
	A-and-D	1s	78.08	48.78	84.52	71.40	79.75	62.44	76.24	63.27
	ActionFormer	1s	77.98	75.22	75.37	83.92	84.01	91.39	75.67	80.49
	TriDet	1s	76.54	75.68	72.88	81.31	84.92	93.81	74.75	80.08
	Shallow D.	2s	69.75	42.18	85.01	77.84	84.60	73.36	79.72	67.69
	A-and-D	2s	78.33	50.22	84.78	73.66	82.12	65.43	**85.40**	71.87
	ActionFormer	2s	61.20	59.68	68.56	70.85	78.72	87.25	72.91	**81.73**
	TriDet	2s	68.30	62.51	69.88	74.12	82.09	91.70	74.04	79.84
Camera	ActionFormer	0.5s	48.97	42.98	68.39	62.88	58.60	53.70	70.43	73.19
	TriDet	0.5s	51.25	50.85	70.96	69.08	60.98	55.52	68.52	66.87
	ActionFormer	1s	64.29	62.62	**77.37**	87.25	74.11	78.20	63.00	82.72
	TriDet	1s	60.74	62.18	76.84	87.84	66.60	73.17	62.31	**84.26**
	ActionFormer	2s	59.88	57.27	76.76	**90.20**	73.65	82.32	**71.78**	78.15
	TriDet	2s	55.54	58.63	76.39	84.31	66.91	78.75	61.31	75.83
I + C	ActionFormer	0.5s	79.81	69.93	80.94	74.64	81.65	84.12	71.43	75.35
	TriDet	0.5s	79.98	71.20	74.35	69.08	84.83	85.85	77.16	80.15
	ActionFormer	1s	83.55	80.74	**87.30**	**94.12**	85.71	94.71	75.98	82.18
	TriDet	1s	81.75	83.90	86.51	92.22	88.27	97.60	75.16	80.27
	ActionFormer	2s	67.90	69.20	80.19	91.18	82.72	94.50	77.35	88.00
	TriDet	2s	70.16	72.56	78.35	87.84	83.30	94.62	76.97	85.00

apply to sbj_14. More specifcially, improvements and decline rates between the two recordings lie within the expected standard deviation across participants (between 15% to 20%). Though being a small sample size of only two participants, the results suggest that in order to guarantee a reliable detection of activities, each participant would need to be recorded multiple times under different conditions. Nevertheless, in order to come up with reliable conclusions, future extensions of the WEAR dataset would need focus on re-recording more participants multiple times under varying conditions.

C.8 Additional Visualizations of Final Runs

In addition to the visualisations supplied in the main paper, the following provides supplementary visualizations for further analysis. All models mentioned in this section were trained using a clip length of 1.0 second with a 50% overlap. Predictions made by the temporal action localization models (Zhang et al., 2022; Shi et al., 2023) were filtered using a score threshold of 0.2 and predictions made by inertial-based architectures (Bock et al., 2021; Abedin et al., 2021) were filtered using a majority vote filter of 15 seconds. Figure 7 provides confusion matrices of the ActionFormer (Zhang et al., 2022) being applied using inertial, camera and combined (inertial + camera) features. Figure 8 provides Confusion matrices of the shallow DeepConvLSTM (Bock et al., 2021) and improved Attend-and-Discriminate (Abedin et al., 2021) applied on inertial data. Figure 9 shows a color-coded visualisation of predictions streams of all models mentioned in the results table of the main paper. Figure 10 delivers a side-by-side comparison of the confusion matrices of all models involved in the *Oracle*-late-fusion-approach analysis mentioned in the main paper. Figure 10 shows that a joint

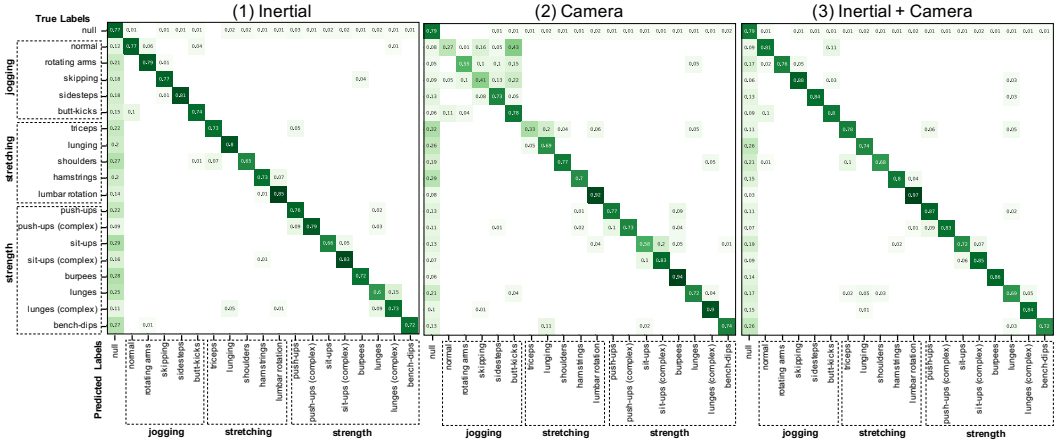


Figure 7: Confusion matrices of the ActionFormer (Zhang et al., 2022) being applied using only inertial, vision (camera) and both combined (inertial + camera).

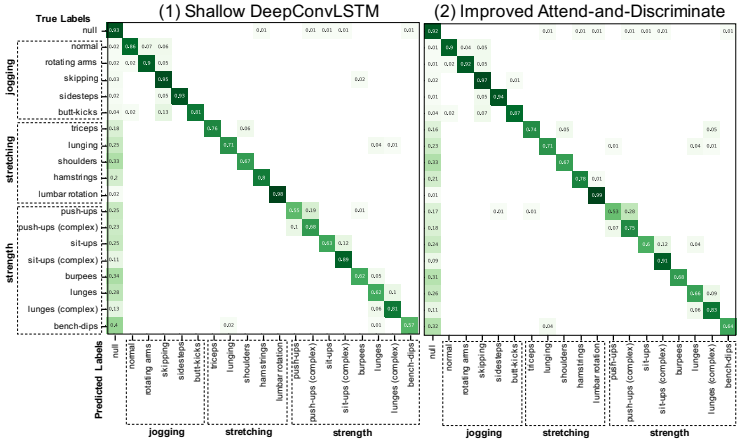


Figure 8: Confusion matrices of the shallow DeepConvLSTM (Bock et al., 2021) and improved Attend-and-Discriminate (Abedin et al., 2021).

learning of both modalities particularly improves differentiation between the NULL-class and the activity classes resulting in better action boundaries, i.e. higher mAP scores, and classification scores.

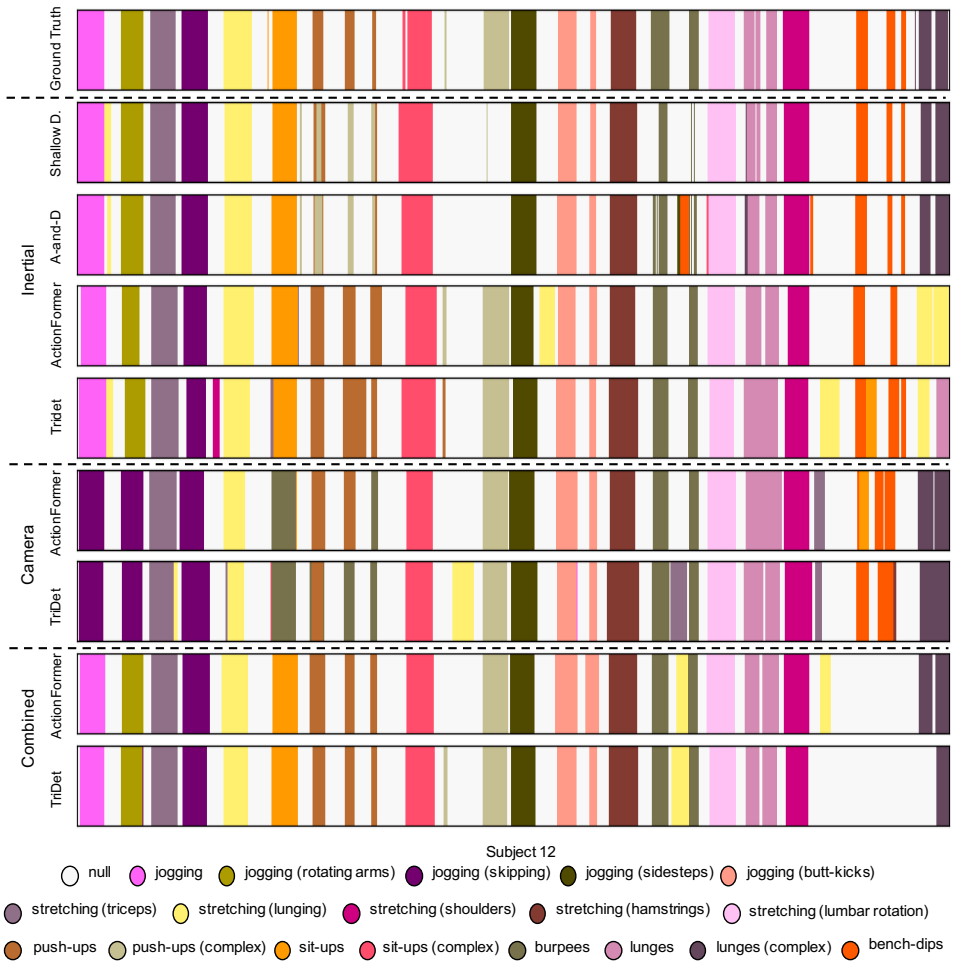


Figure 9: Color-coded comparison of the ground truth data (top row) with the shallow DeepConvLSTM (Bock et al., 2021), improved Attend-and-Discriminate (Abedin et al., 2021), ActionFormer (Zhang et al., 2022) and TriDet (Shi et al., 2023) model on varying input modalities.

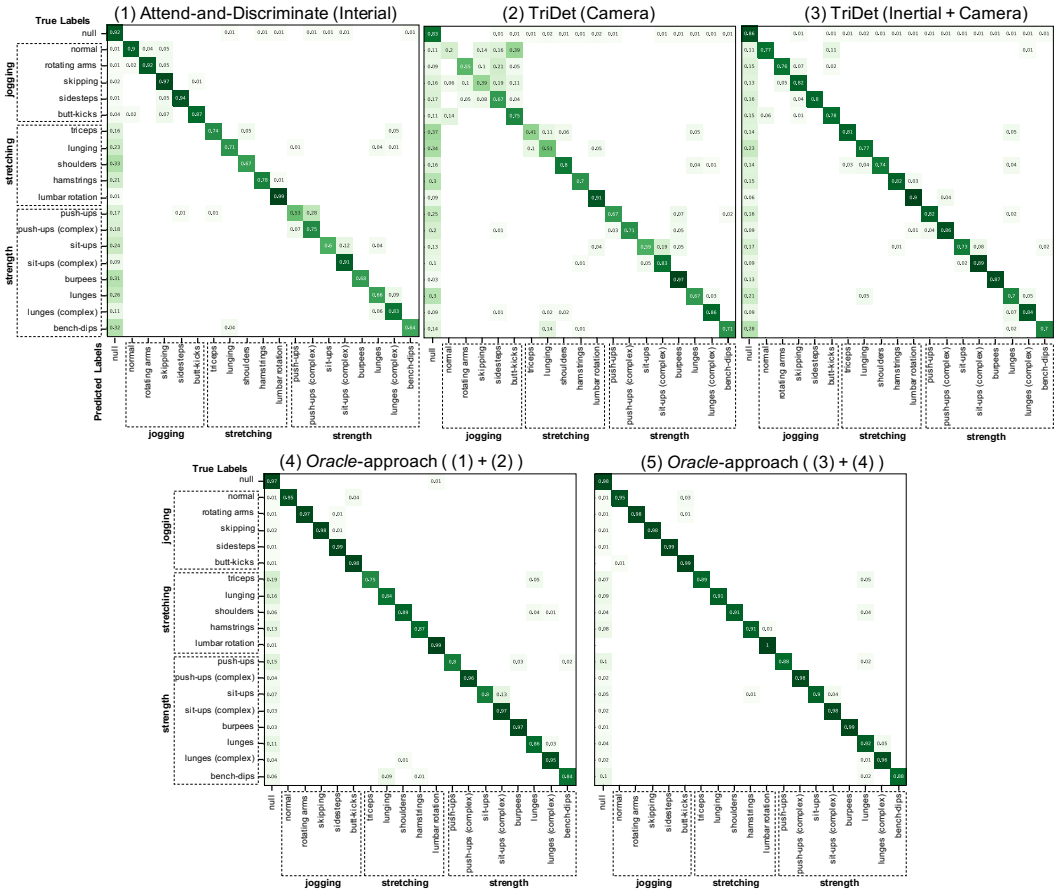
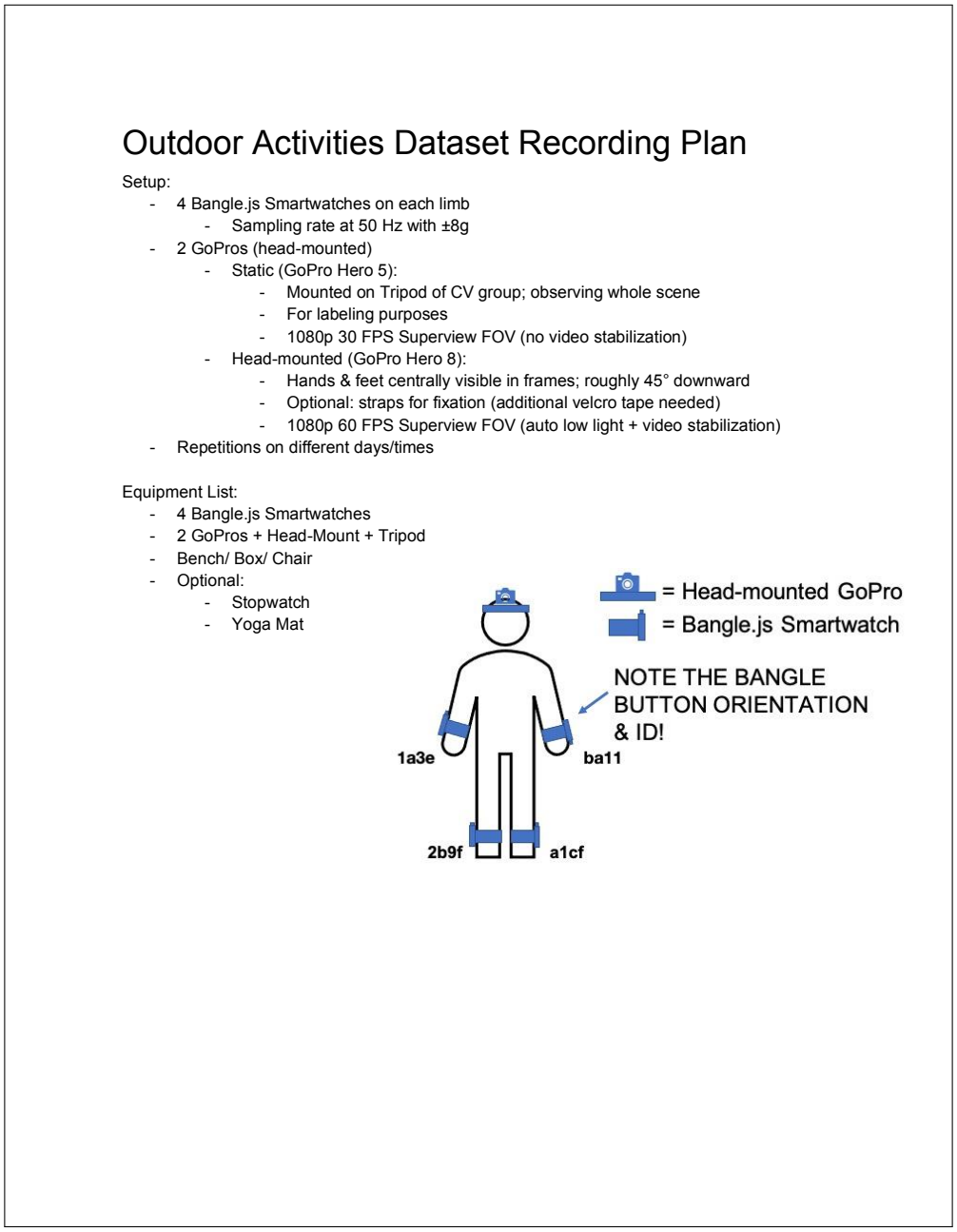


Figure 10: Confusion matrices of the best (1) inertial model (Attend-and-Dicriminate (Abedin et al., 2021)), (2) vision model (TriDet (Shi et al., 2023)) and (3) combined model (vision + inertial) compared with (4) an *Oracle*-combination of the inertial and camera as well (5) *Oracle*-combination of the previous oracle with the combined approach.

D FULL-TEXT RECORDING PLAN



Outdoor Activities Dataset Recording Plan

Setup:

- 4 Bangle.js Smartwatches on each limb
  - Sampling rate at 50 Hz with ±8g
- 2 GoPros (head-mounted)
  - Static (GoPro Hero 5):
    - Mounted on Tripod of CV group; observing whole scene
    - For labeling purposes
    - 1080p 30 FPS Superview FOV (no video stabilization)
  - Head-mounted (GoPro Hero 8):
    - Hands & feet centrally visible in frames; roughly 45° downward
    - Optional: straps for fixation (additional velcro tape needed)
    - 1080p 60 FPS Superview FOV (auto low light + video stabilization)
- Repetitions on different days/times

Equipment List:

- 4 Bangle.js Smartwatches
- 2 GoPros + Head-Mount + Tripod
- Bench/ Box/ Chair
- Optional:
  - Stopwatch
  - Yoga Mat

Figure 11: First page of the recording plan of the WEAR dataset.

1st Session (ca. 30 min)

➔ **Each activity 3 sets à repetitions at will (ca. 30 sec); short break after each set (ca. 30 sec)**

Activity	Sub-Activity	Description	Normal Variant	Easy Variant
Running	Sidestep	Start with both feet together; jump on one foot to the side and repeat same motion for other foot to create a jumping motion; repeat		
	Butt-Kicks	Fold hands on butt; jog while trying to lift alternating each heel as close to butt as possible ("kicking" it)		
Stretching	Shoulder	Start by standing straight; stretch left arm to the right while keeping it parallel to the ground; use lower right arm to press against left arm upper left arm (close to the elbow) trying to move it closer to the body; hold stretch; repeat by switching arms		
	Hamstrings	Start by sitting down; have left leg stretched out straight in a 45° angle to the left side and keep right leg as sitting cross-legged with right foot touching the left knee's side; try to reach for left foot; hold stretch; repeat by switching legs		
	Lumbar Rotation	Start by laying on back; reach out with both arms to the side; raise legs; move legs to the left side as close as possible to the ground while keeping them straight; do not move torso; hold stretch; repeat by moving legs to opposite side		

Figure 12: Second page of the recording plan of the WEAR dataset. Note that pictures are blurred for anonymization and are short video-clips in the original document.

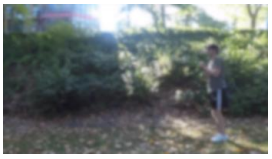
Burpees	Normal	Start by standing straight; put your hands on the ground and jump back with your feet into a push-up position; do a push-up; jump forward with your feet into the starting position and jump up with raised arms; repeat		No Push Up, but lay flat on ground; get up by standing up instead of jumping
Walking Lunges	Normal	Stand straight; Keep your hands crossed in front of your torso; move your right foot forward, bending your left knee to the ground and right knee into a 90° angle; step your left foot forward going back into the position your started in; repeat to create walking motion		
	Complex	Stand straight; keep your arms raised in front of you pointing forward parallel to the floor; do a lunge, but when at the bottom turn your torso to the side of the foot being forward; repeat to create walking motion		
Bench Dips	Normal	Stand with your back facing a chair; go into dip position with your hands on the chair and your legs straight in front of you; move your body downwards by bending your arms into a 90° angle; keep legs and back as straight as possible; move up again; repeat		

Figure 13: Third page of the recording plan of the WEAR dataset. Note that pictures are blurred for anonymization and are short video-clips in the original document.

2nd Session (ca. 30 min)

➔ **Each activity 3 sets à repetitions at will (ca. 30 sec); short break after each set (ca. 30 sec)**

Activity	Sub-Activity	Description	Normal Variant	Easy Variant
Running	Jogging	Normal jogging		
	Jogging with rotating arms	Jogging while rotating arms backwards; first only left arm; then only right arm; then both arms simultaneously		
	Skipping	Alternate between jumping from left leg while lifting the right knee high; land on the left leg; repeat by switching legs to create running motion		
	Triceps	Start by standing straight; raise left arm behind head and try to touch the right shoulder; use right arm to grab left elbow trying to move left hand further down the shoulder; hold stretch; repeat by switching arms		
	Lunging	Start by standing in a split stance with right front forward and left foot straight back; bend right knee about 90°; place hands on your forward knee; hold stretch; repeat by switching legs		

Figure 14: Fourth page of the recording plan of the WEAR dataset. Note that pictures are blurred for anonymization and are short video-clips in the original document.

Push-ups	Normal	Normal Push-Up		On knees
	Complex	Move into a push-up position; do a push-up by lowering your body to the ground; after moving it back up, reach out with the right arm to the sky, opening your torso so that it faces to the right; move back into push-up; repeat for left arm; repeat sequence		On knees
Sit-ups	Normal	Lay on your back; have your hands touch the sides of your head; move your legs into a 90° angle with your feet on the ground; move your torso towards your knees while keeping the legs in place; repeat		Straight legs
	Complex	Lay on your back with your hands touching the sides of your head; move your legs into a 90° angle (feet on the ground) while also (!) moving your torso towards your knees; when reaching highest point, touch first your right heel with your right hand; and left heel with left hand; repeat		First move up with upper body; then with lower body so that legs are straight during situp

Figure 15: Fifth page of the recording plan of the WEAR dataset. Note that pictures are blurred for anonymization and are short video-clips in the original document.

REFERENCES

Alireza Abedin, Mahsa Ehsanpour, Qinfeng Shi, Hamid Rezatofighi, and Damith C. Ranasinghe. Attend and discriminate: Beyond the state-of-the-art for human activity recognition using wearable sensors. *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(1):1–22, 2021. URL https://doi.org/10.1145/3448083.

Marius Bock, Alexander Hoelzemann, Michael Moeller, and Kristof Van Laerhoven. Improving Deep Learning for HAR with Shallow LSTMs. In *ACM International Symposium on Wearable Computers*, 2021. URL https://doi.org/10.1145/3460421.3480419.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *IEEE Conference on Computer Vision and Pattern Recognition*, July 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Carreira_Quo_Vadis_Action_CVPR_2017_paper.html.

Joseph DelPreto, Chao Liu, Yiyue Luo, Michael Foshey, Yunzhu Li, Antonio Torralba, Wojciech Matusik, and Daniela Rus. ActionSense: A multimodal dataset and recording framework for human activities using wearable sensors in a kitchen environment. In *Neural Information Processing Systems Track on Datasets and Benchmarks*, 2022. URL https://action-sense.csail.mit.edu.

Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. The dataset nutrition label: A framework to drive higher data quality standards. *CoRR*, abs/1805.03677, 2018. URL https://arxiv.org/abs/1805.03677.

Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017. URL http://arxiv.org/abs/1705.06950.

Mehrab Bin Morshed, Harish Haresamudram, Dheeraj Bandaru, Gregory Abowd, and Thomas Plötz. A personalized approach for developing a snacking detection system using earbuds in a semi-naturalistic setting. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing and International Symposium on Wearable Computers*, 2022. URL https://doi.org/10.1145/3544794.3558469.

Philipp M. Scholl, Benjamin Völker, Bernd Becker, and Kristof Van Laerhoven. A multi-media exchange format for time-series dataset curation. In Nobuo Kawaguchi, Nobuhiko Nishio, Daniel Roggen, Sozo Inoue, Susanna Pirttikangas, and Kristof Van Laerhoven (eds.), *Human Activity Sensing*, pp. 111–119. Springer, 2019. URL https://doi.org/10.1007/978-3-030-13001-5_8.

Dingfeng Shi, Yujie Zhong, Qiong Cao, Lin Ma, Jia Li, and Dacheng Tao. TriDet: Temporal action detection with relative boundary modeling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Kristof Van Laerhoven, Alexander Hoelzemann, Iris Pahmeier, Andrea Teti, and Lars Gabrys. Validation of an open-source ambulatory assessment system in support of replicable activity studies. *German Journal of Exercise and Sport Research*, 52(2):262–272, 2022. URL https://doi.org/10.1007/s12662-022-00813-2.

Chen-Lin Zhang, Jianxin Wu, and Yin Li. ActionFormer: Localizing moments of actions with transformers. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *European Conference on Computer Vision*, 2022. URL https://doi.org/10.1007/978-3-031-19772-7_29.