Supplementary Material: Temporal Action Localization for Inertial-based Human Activity Recognition

MARIUS BOCK, University of Siegen, Germany
MICHAEL MOELLER, University of Siegen, Germany
KRISTOF VAN LAERHOVEN, University of Siegen, Germany

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing design and evaluation methods**; • **Computing methodologies → Neural networks**.

Additional Key Words and Phrases: Deep Learning, Inertial-based Human Activity Recognition, Body-worn Sensors, Temporal Action Localization

A METHODOLOGICAL TRANSPARENCY APPENDIX

The source code that was used to conduct all experiments, postprocessing and evaluation steps mentioned in this paper is available via Github (https://github.com/mariusbock/tal_for_har). 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 to add additional datasets 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. All experiments were conducted on a cluster-based system, with each run being assigned one Tesla V100 GPU along with 10 GB of RAM and two AMD EPYC 7452 CPUs. Further, the repository provides links to download each of the six mentioned dataset as used during experiments. Each dataset download is structured into the (1) '.json'-formatted annotations, (2) raw inertial and (3) precomputed vectorized feature embeddings as mentioned in the main paper.

B SUPPLEMENTARY RESULTS OF OFFLINE ACTIVITY RECOGNITION EXPERIMENTS

B.1 Two-stage Training

Table 1 provides extended offline activity recognition results of the two-stage training using both LSTM- and attention-based features. Figure 1 and 2 present additional confusion matrices of the ActionFormer being applied both with single-stage and two-stage training.

B.2 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.

With the TAL architectures used in this paper being set to predict 2000 segments and are not trained to predict the NULL-class, score thresholding the predicted segments by their associated confidence is essential in order to preserve breaks. As evident by results presented in Table 2 and Figure 3, without score thresholding (almost) all timestamps get classified into one of the activity classes. Applying a score threshold thus results in significant increases in NULL-class accuracy across all datasets. Though mAP scores decrease, overall F1-scores increase by a significant margin across all datasets. The optimal score threshold for each dataset was chosen by trying to balance the increase in F1-score compared to the decrease in mAP score.

Authors' addresses: Marius Bock, marius.bock@uni-siegen.de, University of Siegen, Ubiquitous Computing, Computer Vision, Siegen, Germany; Michael Moeller, michael.moeller@uni-siegen.de, University of Siegen, Computer Vision, Siegen, Germany; Kristof Van Laerhoven, kvl@eti.uni-siegen.de, University of Siegen, Ubiquitous Computing, Siegen, Germany.

Table 1. Average single-stage vs. two-stage LOSO per-sample classification metrics (Precision (P), Recall (R), F1-Score (F1)), misalignment ratios and average mAP applied at different tIoU thresholds (0.3:0.1:07) obtained on six inertial HAR benchmark datasets [4, 5, 7–9, 11] of the evaluated TAL architectures [10, 12, 15]. Best results per dataset are in **bold**.

	Model	Two-Stage	P (↑)	R (↑)	F1 (↑)	UR (↓)	OR (↓)	DR (↓)	IR (↓)	FR (↓)	MR (↓)	mAP (↑)
Opportunity	ActionFormer		54.63	**58.78**	51.93	**0.19**	13.34	0.41	33.82	0.01	**0.42**	**51.24**
	TemporalMaxer		44.44	55.63	44.74	0.21	14.13	0.40	43.78	0.01	0.55	46.31
	TriDet		48.72	57.69	48.79	0.23	13.20	**0.39**	40.07	0.01	0.58	49.70
	ActionFormer	L	60.58	38.29	40.85	0.26	10.38	0.78	26.43	**0.00**	0.46	29.81
	TemporalMaxer	L	59.52	40.10	41.49	0.28	12.48	0.73	27.26	0.01	0.65	29.44
	TriDet	L	60.80	39.25	41.47	0.27	**10.26**	0.76	27.78	**0.00**	0.51	30.77
	ActionFormer	A	63.97	53.24	53.41	0.24	12.55	0.52	**24.10**	**0.00**	0.77	46.71
	TemporalMaxer	A	60.63	54.43	52.98	0.27	15.70	0.46	24.39	0.01	0.91	43.50
	TriDet	A	**64.29**	54.31	**54.74**	0.25	11.96	0.49	24.23	0.01	0.87	47.17
SBHAR	ActionFormer		87.02	84.37	84.43	0.36	5.41	0.16	5.09	**0.00**	0.20	**95.46**
	TemporalMaxer		86.52	83.37	83.66	0.41	6.02	0.19	4.41	**0.00**	0.24	94.39
	TriDet		88.95	86.15	86.45	0.38	5.41	**0.12**	**3.95**	0.01	0.29	94.75
	ActionFormer	L	86.52	76.49	78.96	0.38	3.16	0.35	6.72	**0.00**	0.04	84.82
	TemporalMaxer	L	85.66	75.53	78.04	0.42	3.22	0.37	6.88	**0.00**	0.14	84.38
	TriDet	L	85.89	76.87	79.15	0.36	**2.94**	0.34	7.68	0.01	0.17	83.65
	ActionFormer	A	90.00	85.69	86.60	0.33	4.60	0.19	4.64	**0.00**	**0.02**	93.31
	TemporalMaxer	A	89.21	83.87	85.02	0.37	4.95	0.20	4.93	**0.00**	0.13	93.23
	TriDet	A	**90.04**	**86.83**	**87.23**	**0.30**	4.88	0.20	4.31	0.01	0.10	93.50
Wetlab	ActionFormer		40.71	49.25	40.71	0.56	9.41	0.79	52.44	0.07	0.79	33.53
	TemporalMaxer		**50.43**	36.65	37.09	0.59	9.37	0.83	53.59	0.10	0.61	35.72
	TriDet		44.13	49.15	42.85	0.58	8.85	0.75	48.63	0.10	0.84	34.05
	ActionFormer	L	47.87	43.36	43.06	0.69	**4.66**	0.93	36.92	**0.05**	0.59	26.61
	TemporalMaxer	L	46.68	44.54	42.87	0.70	4.96	0.91	40.09	0.08	**0.58**	25.66
	TriDet	L	47.37	43.10	42.66	0.65	4.08	0.95	**35.95**	0.07	0.79	27.06
	ActionFormer	A	47.01	56.12	47.30	0.55	9.69	**0.64**	40.46	0.06	1.09	43.07
	TemporalMaxer	A	44.14	**56.96**	45.36	**0.51**	9.73	0.67	44.92	0.08	1.03	40.55
	TriDet	A	48.00	54.51	**47.57**	0.53	7.94	0.71	39.78	0.07	1.00	**41.20**
WEAR	ActionFormer		71.88	76.70	72.43	0.22	6.48	0.65	7.12	**0.01**	2.06	73.80
	TemporalMaxer		69.54	72.80	69.52	0.23	6.87	0.83	5.50	**0.01**	1.50	69.18
	TriDet		73.57	77.54	73.18	0.28	4.79	0.64	6.21	0.03	1.64	77.12
	ActionFormer	L	80.32	76.63	75.99	0.27	**2.46**	0.49	8.15	0.12	0.66	63.38
	TemporalMaxer	L	79.61	77.31	75.80	0.28	3.50	0.49	7.94	0.10	0.90	59.23
	TriDet	L	80.18	76.63	75.65	0.29	3.22	0.47	8.24	0.13	**0.55**	61.41
	ActionFormer	A	85.01	**86.34**	84.56	**0.20**	4.51	0.30	**2.94**	**0.01**	1.42	85.17
	TemporalMaxer	A	82.04	84.31	82.13	**0.20**	4.81	0.39	3.25	**0.01**	1.40	82.97
	TriDet	A	**86.29**	85.77	**84.78**	0.24	3.35	**0.29**	3.39	0.02	1.33	**85.94**
Hang-Time	ActionFormer		49.19	57.57	51.23	0.62	11.63	0.51	47.65	0.48	0.64	29.26
	TemporalMaxer		45.01	54.56	47.17	0.71	10.97	0.45	52.71	1.13	0.65	27.86
	TriDet		49.59	55.14	50.67	0.73	9.85	0.52	48.55	0.69	0.62	29.24
	ActionFormer	L	**56.10**	51.63	51.76	0.57	8.90	0.57	42.21	0.31	**0.18**	23.80
	TemporalMaxer	L	54.92	51.99	51.51	0.57	8.87	0.53	43.87	0.41	0.23	23.72
	TriDet	L	56.72	50.55	51.43	0.59	**8.39**	0.60	**41.82**	0.35	0.20	22.70
	ActionFormer	A	53.64	62.47	55.81	**0.55**	11.13	0.39	43.25	0.28	0.50	37.52
	TemporalMaxer	A	51.39	61.24	53.79	0.56	11.18	**0.36**	45.52	0.72	0.47	**36.77**
	TriDet	A	54.83	**62.27**	**56.66**	0.59	9.50	0.39	43.40	**0.26**	0.50	36.84
RWHAR	ActionFormer		63.76	67.64	61.24	2.48	11.12	2.06	11.23	**0.09**	**0.00**	65.40
	TemporalMaxer		63.20	67.60	60.59	2.59	11.95	1.81	13.96	0.36	0.05	50.60
	TriDet		69.27	73.04	67.86	1.48	6.88	2.03	10.24	0.23	**0.00**	69.98
	ActionFormer	L	71.95	74.76	69.67	1.97	10.63	1.56	6.87	0.16	**0.00**	71.37
	TemporalMaxer	L	75.35	77.72	72.61	1.55	7.76	1.33	8.00	0.37	**0.00**	39.79
	TriDet	L	77.30	80.00	75.56	1.23	6.07	1.37	7.47	0.29	**0.00**	73.53
	ActionFormer	A	80.32	80.94	78.09	1.81	9.16	0.87	5.24	0.16	**0.00**	75.07
	TemporalMaxer	A	76.87	78.10	73.66	2.11	9.41	1.07	7.54	0.20	0.06	60.77
	TriDet	A	**85.82**	**86.86**	**84.08**	**0.96**	**4.64**	**0.78**	**5.09**	0.21	**0.00**	**77.86**

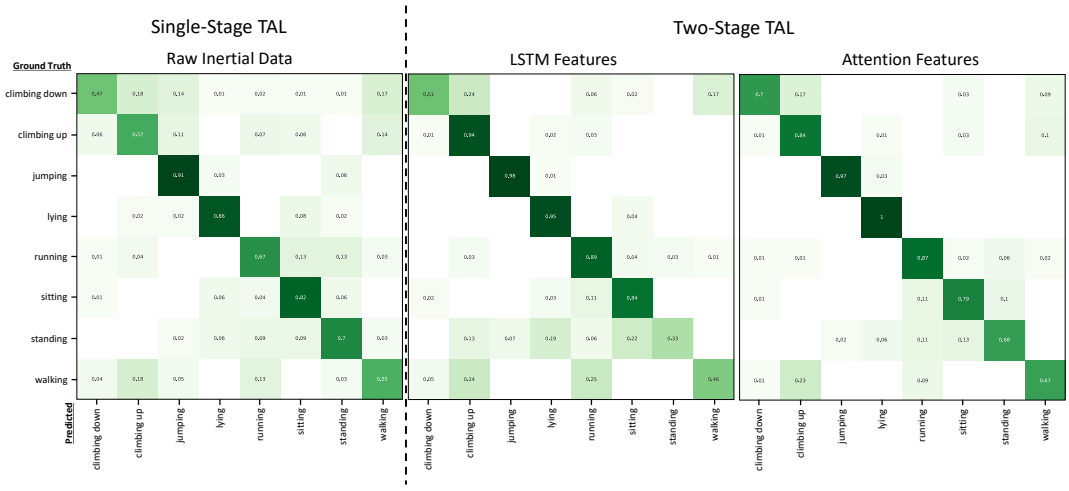

Fig. 1. Single-Stage vs. Two-Stage confusion matrices of the ActionFormer [15] applied on the RWHAR dataset [11]. Note that confusions which are 0 are omitted.

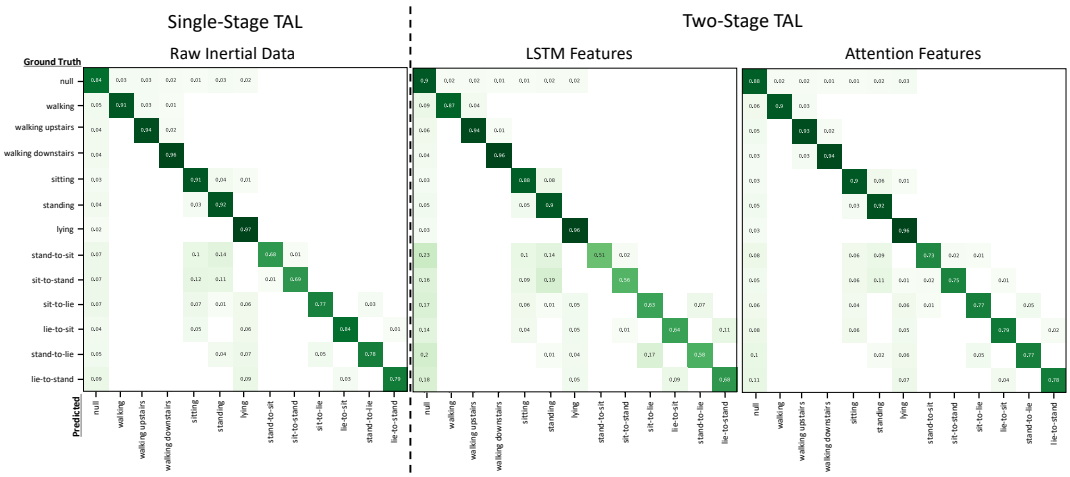

Fig. 2. Single-Stage vs. Two-Stage confusion matrices of the ActionFormer [15] applied on the SBHAR dataset [7]. Note that confusions which are 0 are omitted.

The inertial architectures in this paper are optimized to predict per-window class labels. This causes unprocessed prediction streams of all three models to show frequently occurring, short activity switches as the context of prior windows is disregarded during the prediction of the current timestamp. This causes segments of activities to get fragmented frequently, ultimately leading the mAP score of the inertial models to be significantly lower than that of TAL architectures. Therefore, all prediction streams of the inertial architectures were smoothed using a majority filter. The exact size of each majority filter was chosen dataset-dependent, trying out different

filter sizes and choosing the one which balanced changes in F1-score and mAP to the best extend. Table 3 and Figure 4 illustrate the effect of the majority filter on the inertial architecture's prediction streams.

Table 2. Unprocessed versus postprocessed average LOSO results obtained on the six inertial HAR benchmark datasets [4, 5, 7–9, 11] using the three temporal action localization models [10, 12, 15]. The table provides along the applied score threshold s the per-sample classification metrics, i.e. Precision (P), Recall (R), F1-Score (F1), misalignment ratios [14] and average mAP applied at different tIoU thresholds (0.3:0.1:07). All experiments were conducted the same style as reported in the main paper. Best results per dataset are in **bold**.

	s	Model	P (↑)	R (↑)	F1 (↑)	UR (↓)	OR (↓)	DR (↓)	IR (↓)	FR (↓)	MR (↓)	mAP (↑)
Opportunity	0.0	ActionFormer	48.08	61.44	49.37	0.20	13.83	0.35	40.38	**0.01**	0.43	**53.74**
		TemporalMaxer	41.34	56.50	43.11	0.21	13.85	0.37	47.41	**0.01**	0.56	47.32
		TriDet	41.41	59.30	44.61	0.25	**13.11**	**0.31**	48.08	**0.01**	0.54	51.97
	0.2	ActionFormer	**54.63**	**58.78**	**51.93**	**0.19**	13.34	0.41	**33.82**	**0.01**	**0.42**	51.24
		TemporalMaxer	44.44	55.63	44.74	0.21	14.13	0.40	43.78	**0.01**	0.55	46.31
		TriDet	48.72	57.69	48.79	0.23	13.20	0.39	40.07	**0.01**	0.58	49.70
SBHAR	0.0	ActionFormer	60.16	79.85	66.84	**0.26**	10.05	0.09	24.13	0.02	0.63	**97.08**
		TemporalMaxer	61.96	81.18	68.66	0.29	10.21	0.09	23.80	0.02	0.61	96.28
		TriDet	61.96	81.18	68.66	0.28	7.93	**0.06**	24.15	0.04	0.61	97.16
	0.3	ActionFormer	87.02	84.37	84.43	0.36	**5.41**	0.16	5.09	**0.00**	**0.20**	95.46
		TemporalMaxer	86.52	83.37	83.66	0.41	6.02	0.19	4.41	**0.00**	0.24	94.39
		TriDet	**88.95**	**86.15**	**86.45**	0.38	**5.41**	0.12	**3.95**	0.01	0.29	94.75
Wetlab	0.0	ActionFormer	27.54	51.37	23.74	0.59	7.99	0.53	71.98	0.16	0.54	39.56
		TemporalMaxer	27.98	50.02	23.65	0.62	7.47	0.59	72.28	0.21	**0.44**	38.68
		TriDet	27.92	51.82	24.12	0.60	**6.85**	**0.43**	73.41	0.24	0.54	**41.72**
	0.15	ActionFormer	40.71	**49.25**	40.71	**0.56**	9.41	0.79	52.44	**0.07**	0.79	33.53
		TemporalMaxer	**50.43**	36.65	37.09	0.59	9.37	0.83	53.59	0.10	0.61	35.72
		TriDet	44.13	49.15	**42.85**	0.58	8.85	0.75	**48.63**	0.10	0.84	34.05
WEAR	0.0	ActionFormer	56.87	78.35	60.81	**0.18**	9.13	0.46	26.22	0.02	2.76	81.63
		TemporalMaxer	56.72	79.76	61.70	**0.18**	8.51	0.42	28.13	0.03	2.49	78.69
		TriDet	53.97	78.22	60.63	0.24	6.52	**0.36**	31.61	0.09	2.22	**83.08**
	0.2	ActionFormer	71.88	76.70	72.43	0.22	6.48	0.65	7.12	**0.01**	2.06	73.80
		TemporalMaxer	69.54	72.80	69.52	0.23	6.87	0.83	**5.50**	**0.01**	**1.50**	69.18
		TriDet	**73.57**	**77.54**	**73.18**	0.28	**4.79**	0.64	6.21	0.03	1.64	77.12
Hang-Time	0.0	ActionFormer	41.26	58.23	44.11	**0.54**	10.75	0.33	58.70	**0.42**	0.63	31.21
		TemporalMaxer	39.21	56.20	42.07	0.55	9.73	0.36	61.94	0.79	0.60	29.23
		TriDet	39.98	57.17	43.05	0.55	**8.40**	**0.31**	62.55	0.54	**0.58**	**31.75**
	0.15	ActionFormer	49.19	**57.57**	**51.23**	0.62	11.63	0.51	47.65	0.48	0.64	29.26
		TemporalMaxer	45.01	54.56	47.17	0.71	10.97	0.45	52.71	1.13	0.65	27.86
		TriDet	**49.59**	55.14	50.67	0.73	9.85	0.52	**48.55**	0.69	0.62	29.24

B.3 Additional Confusion Matrices

The following provides additional confusion matrices which were not part of the results chapter. Specifically, we provide confusion matrices of the TAL and inertial architectures being applied on the Opportunity [8] (see Figures 5 and 5), the Wetlab [9] (see Figures 7 and 8), the WEAR [4] (see Figures 9 and 10) and the Hang-Time [5] (see Figures 11 and 12) dataset. All experiments were conducted as specified in the main paper.

Table 3. Unprocessed versus postprocessed average LOSO results obtained on the six inertial HAR benchmark datasets [4, 5, 7–9, 11] using the four inertial architectures [1, 3, 6, 16]. The table provides along the applied majority filter m the per-sample classification metrics, i.e. Precision (P), Recall (R), F1-Score (F1), misalignment ratios [14] and average mAP applied at different tIoU thresholds (0.3:0.1:07). Best results per dataset are in **bold**.

	m	Model	P (↑)	R (↑)	F1 (↑)	UR (↓)	OR (↓)	DR (↓)	IR (↓)	FR (↓)	MR (↓)	mAP (↑)
Opportunity	0 sec.	DeepConvLSTM	45.29	34.91	34.75	0.50	9.61	0.49	47.27	0.09	0.12	6.16
		Shallow D.	39.62	27.93	27.33	0.41	8.88	0.71	52.50	0.08	**0.09**	4.75
		A-and-D	30.61	45.61	32.87	0.46	**7.89**	**0.27**	65.02	0.10	0.17	5.63
		TinyHAR	42.12	51.65	43.26	0.46	11.37	0.23	49.12	0.09	0.29	9.35
	2.5 sec	DeepConvLSTM	50.22	33.88	34.41	0.30	17.29	0.78	**31.26**	**0.01**	0.23	13.97
		Shallow D.	42.08	27.18	26.46	**0.24**	14.28	0.97	35.06	**0.01**	0.15	10.61
		A-and-D	35.25	48.55	36.35	0.32	15.28	0.45	52.55	0.02	0.35	13.75
		TinyHAR	**48.09**	**54.27**	**47.09**	0.34	19.99	0.39	34.01	0.02	0.46	**19.78**
SBHAR	0 sec.	DeepConvLSTM	54.04	59.98	54.67	0.61	3.31	0.29	44.48	0.51	**0.00**	18.58
		Shallow D.	72.29	**75.80**	70.89	0.65	8.58	0.34	19.20	0.13	0.02	52.03
		A-and-D	55.08	65.19	56.12	0.51	3.83	**0.25**	43.47	0.40	0.01	21.99
		TinyHAR	50.91	59.43	50.48	0.59	**3.20**	**0.25**	48.60	0.47	**0.00**	16.32
	5 sec.	DeepConvLSTM	67.54	63.72	62.31	0.41	7.19	0.59	21.44	0.07	0.10	49.60
		Shallow D.	**72.98**	75.41	**71.13**	0.60	10.19	0.46	**14.23**	**0.02**	0.09	**65.15**
		A-and-D	68.64	71.07	66.49	**0.31**	9.47	0.45	21.71	0.05	0.11	55.79
		TinyHAR	58.91	63.70	56.29	**0.31**	8.16	0.54	31.67	0.05	0.13	45.38
Wetlab	0 sec.	DeepConvLSTM	25.62	38.69	26.31	**0.48**	**0.84**	0.10	81.78	1.47	**0.00**	0.45
		Shallow D.	38.4	37.29	35.01	0.79	6.77	1.04	57.65	0.28	0.14	6.08
		A-and-D	27.38	44.65	26.60	0.51	1.16	**0.06**	79.95	1.37	**0.00**	0.65
		TinyHAR	26.76	42.84	24.44	0.68	1.28	0.15	80.51	1.26	**0.00**	1.22
	20 sec.	DeepConvLSTM	38.65	47.34	**37.87**	0.51	7.13	0.69	48.53	0.21	0.64	11.88
		Shallow D.	**39.01**	35.42	34.42	0.51	9.52	1.57	**34.51**	**0.06**	0.43	**15.40**
		A-and-D	37.75	**55.71**	37.49	0.56	9.69	0.63	57.60	0.16	0.57	12.27
		TinyHAR	34.31	50.84	31.48	0.61	8.41	1.00	61.26	0.11	0.59	10.05
WEAR	0 sec.	DeepConvLSTM	74.86	70.40	70.04	0.21	**1.06**	0.18	22.16	0.73	**0.02**	6.57
		Shallow D.	82.94	77.99	77.09	0.31	2.55	0.28	9.60	0.26	0.07	36.42
		A-and-D	73.37	76.71	72.87	0.16	1.51	0.08	25.03	0.61	0.05	9.08
		TinyHAR	75.33	81.72	76.78	**0.13**	1.82	**0.06**	22.95	0.46	0.05	12.41
	15 sec.	DeepConvLSTM	80.68	76.25	75.78	0.28	2.35	0.52	6.68	0.11	0.32	61.03
		Shallow D.	80.78	78.91	77.71	0.27	3.23	0.50	**5.21**	**0.04**	0.43	67.89
		A-and-D	82.34	83.29	80.61	0.20	4.03	0.33	7.18	0.09	0.52	64.78
		TinyHAR	**83.63**	**87.79**	**83.96**	0.15	5.09	0.20	7.87	0.09	0.54	**68.97**
Hang-Time	0 sec.	DeepConvLSTM	34.35	39.84	35.82	0.61	4.75	0.21	53.94	0.91	0.03	2.91
		Shallow D.	38.12	44.24	39.20	0.57	9.29	0.57	44.73	0.39	0.63	5.95
		A-and-D	32.48	**45.94**	35.14	0.63	**4.39**	**0.17**	56.78	0.98	0.04	3.41
		TinyHAR	29.99	42.25	31.80	0.65	3.69	0.23	60.35	1.20	**0.01**	**2.60**
	5 sec.	DeepConvLSTM	**44.13**	33.95	35.25	**0.28**	10.60	0.88	**20.16**	0.27	1.46	5.44
		Shallow D.	37.97	38.19	36.85	0.35	14.00	1.07	36.38	**0.21**	2.33	5.00
		A-and-D	40.54	43.32	**40.39**	0.35	14.14	0.72	34.81	0.30	1.44	6.73
		TinyHAR	37.09	41.13	36.89	0.41	11.59	0.75	42.90	0.43	1.26	4.73
RWHAR	0 sec.	DeepConvLSTM	74.22	75.27	71.35	0.57	**0.24**	0.19	24.99	3.01	**0.00**	0.00
		Shallow D.	**89.89**	87.40	85.53	0.43	0.44	**0.04**	9.39	1.43	**0.00**	0.00
		A-and-D	75.17	76.30	72.25	0.59	0.32	0.30	23.67	2.70	**0.00**	0.00
		TinyHAR	80.90	80.94	77.86	0.46	0.26	0.04	18.84	2.46	**0.00**	0.00
	40 sec.	DeepConvLSTM	79.05	81.93	77.56	0.65	2.05	1.09	13.63	1.07	**0.00**	**0.11**
		Shallow D.	88.59	**89.01**	**86.85**	**0.31**	1.14	0.38	**6.93**	0.98	**0.00**	0.00
		A-and-D	79.46	82.68	78.09	0.66	2.31	1.17	11.28	**0.84**	**0.00**	0.04
		TinyHAR	83.59	86.25	82.62	0.53	1.14	0.76	11.65	0.89	**0.00**	0.02

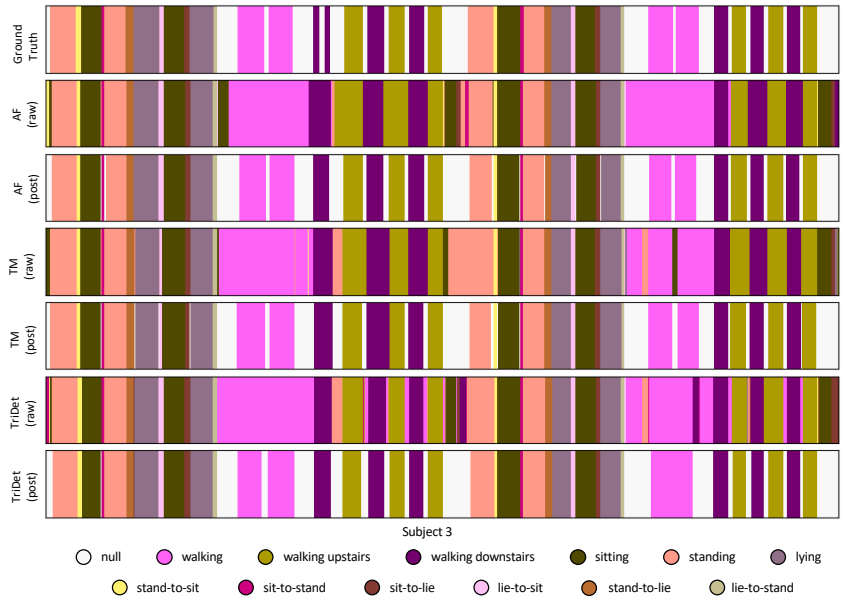

Fig. 3. Colored visualization of unprocessed versus postprocessed prediction streams obtained on the inertial SBHAR dataset [7] using the three TAL architectures [10, 12, 15]. Postprocessing was performed using majority filter of 15 seconds. All experiments were conducted the same style as reported in the main paper. One can see that score thresholding predictions causes not every timestamp to be predicted and NULL-class accuracy to rise. Visualized results are that of Subject 3.

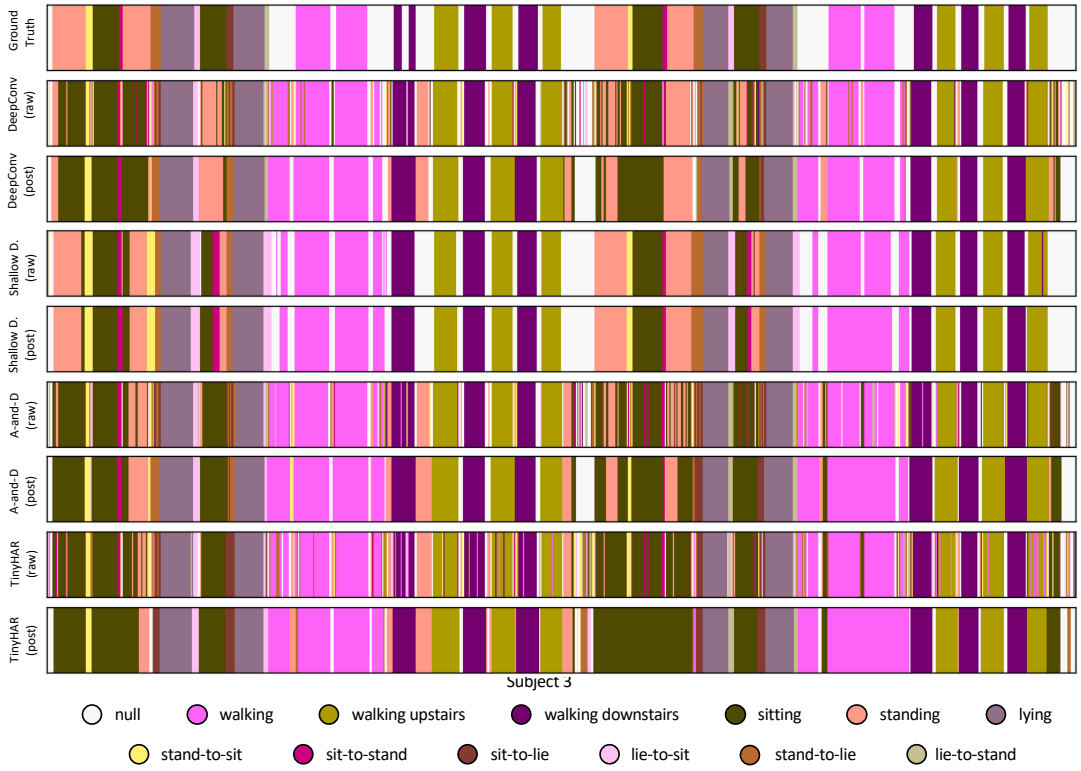

Fig. 4. Colored visualization of unprocessed versus postprocessed prediction streams obtained on the inertial SBHAR dataset [7] using the four inertial architectures [1, 3, 6, 16]. Postprocessing was performed using majority filter of 15 seconds. All experiments were conducted the same style as reported in the main paper. One can see that majority filtering the predictions causes short activity switches to disappear with more coherent segments being created. Visualized results are that of Subject 3.

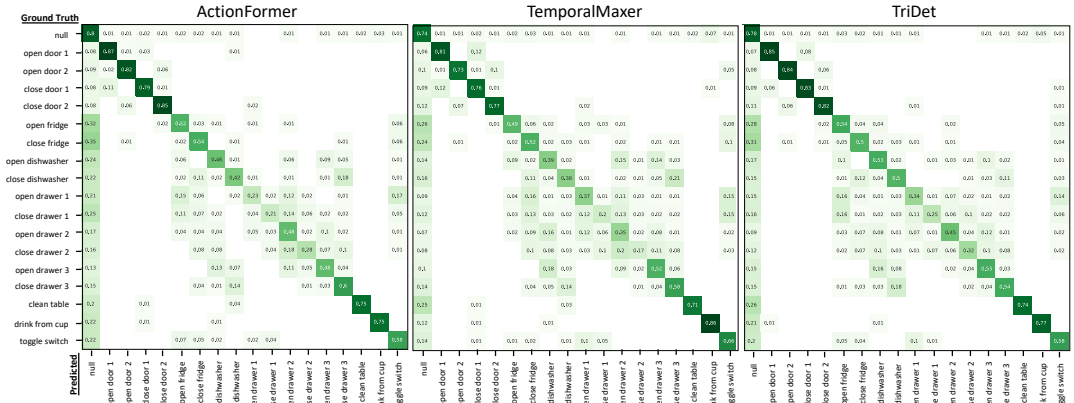

Fig. 5. Confusion matrices of the ActionFormer [15], TemporalMaxer [12] and TriDet model [10] applied on the Opportunity dataset [8] with a one second sliding window and 50% overlap. Results are postprocessed using a score threshold of 0.2. Note that confusions which are 0 are omitted.

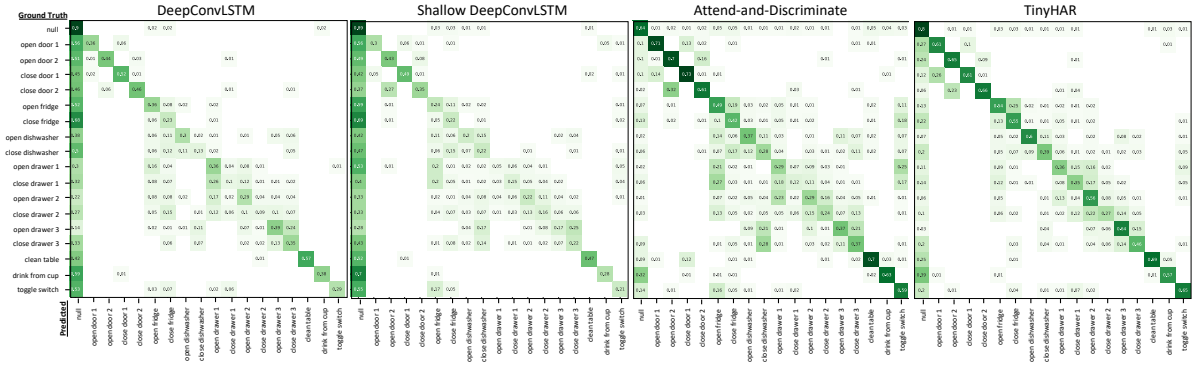

Fig. 6. Confusion matrices of the DeepConvLSTM [6], shallow DeepConvLSTM [3], Attend-and-Discriminate [1] and TinyHAR model [16] applied on the Opportunity dataset [8] with a one second sliding window and 50% overlap. Results are postprocessed using a ,majority filter of 2.5 seconds. Note that confusions which are 0 are omitted.

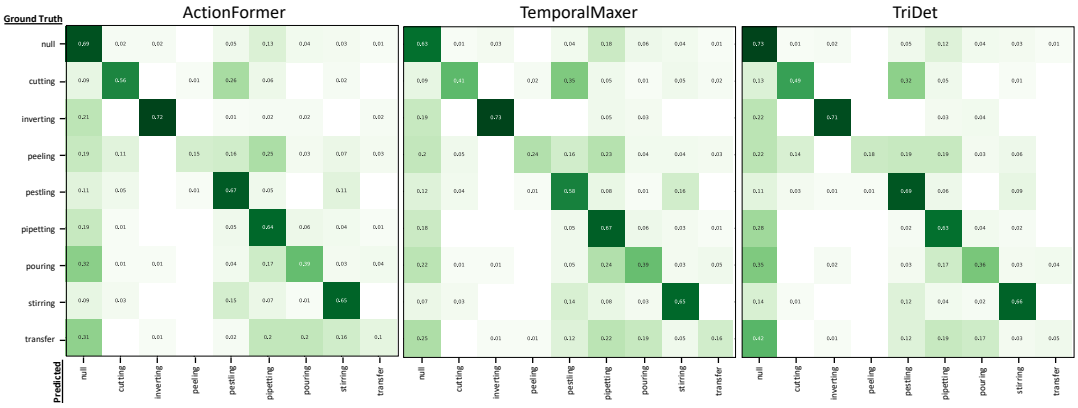

Fig. 7. Confusion matrices of the ActionFormer [15], TemporalMaxer [12] and TriDet model [10] applied on the Wetlab dataset [9] with a one second sliding window and 50% overlap. Results are postprocessed using a score threshold of 0.15. Note that confusions which are 0 are omitted.

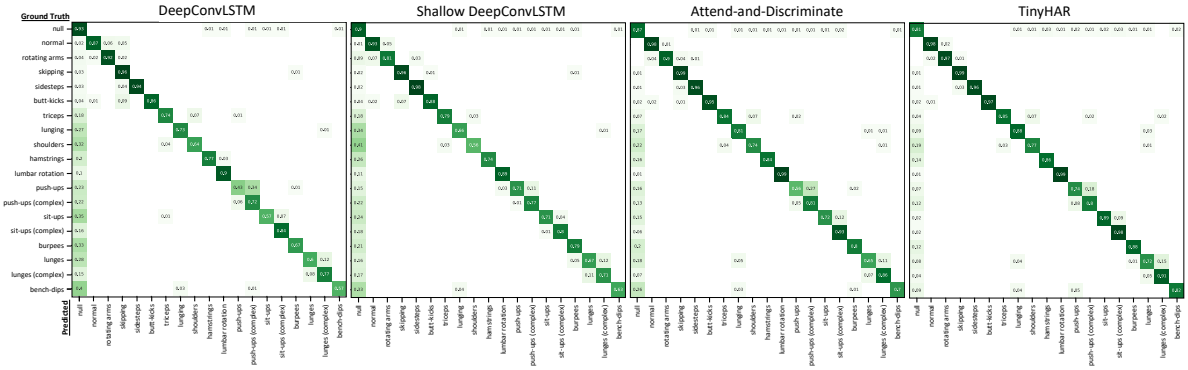

Fig. 8. Confusion matrices of the DeepConvLSTM [6], shallow DeepConvLSTM [3], Attend-and-Discriminate [1] and TinyHAR model [16] applied on the Wetlab dataset [9] with a one second sliding window and 50% overlap. Results are postprocessed using a majority filter 20 seconds. Note that confusions which are 0 are omitted.

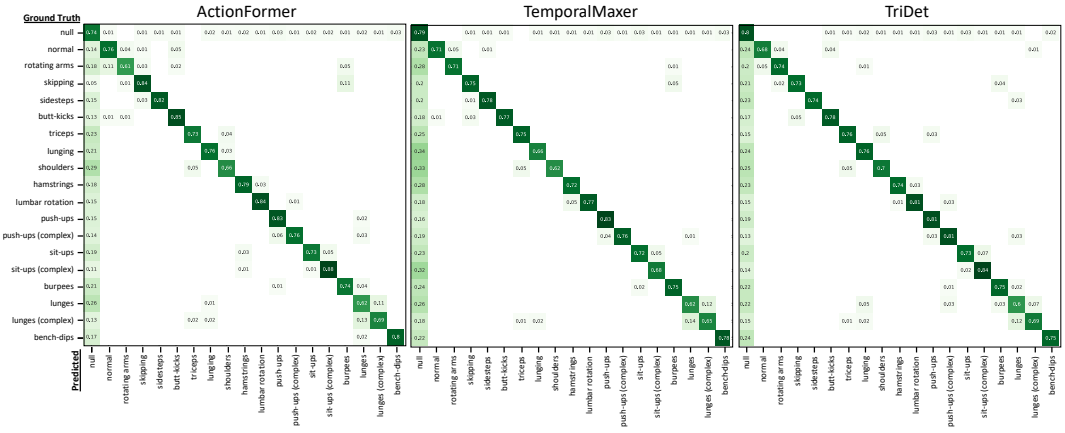

Fig. 9. Confusion matrices of the ActionFormer [15], TemporalMaxer [12] and TriDet model [10] applied on the WEAR dataset [4] with a one second sliding window and 50% overlap. Compared to inertial architectures such as the DeepConvLSTM [6] transition activities get more reliably predicted. Results are postprocessed using a score threshold of 0.2. Note that confusions which are 0 are omitted.

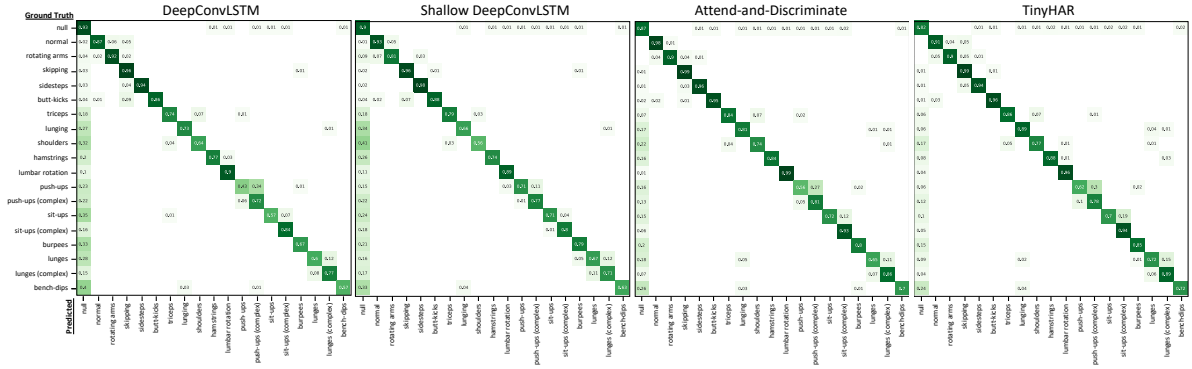

Fig. 10. Confusion matrices of the DeepConvLSTM [6], shallow DeepConvLSTM [3], Attend-and-Discriminate [1] and TinyHAR model [16] applied on the WEAR dataset [4] with a one second sliding window and 50% overlap. Results are postprocessed using a majority filter 15 seconds. Note that confusions which are 0 are omitted.

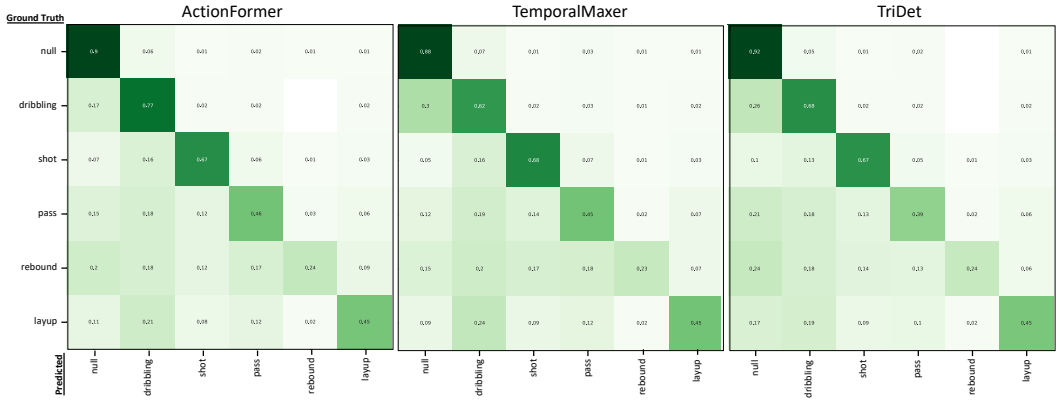

Fig. 11. Confusion matrices of the ActionFormer [15], TemporalMaxer [12] and TriDet model [10] applied on the Hang-Time dataset [5] with a one second sliding window and 50% overlap. Results are postprocessed using a score threshold of 0.15. Note that confusions which are 0 are omitted.

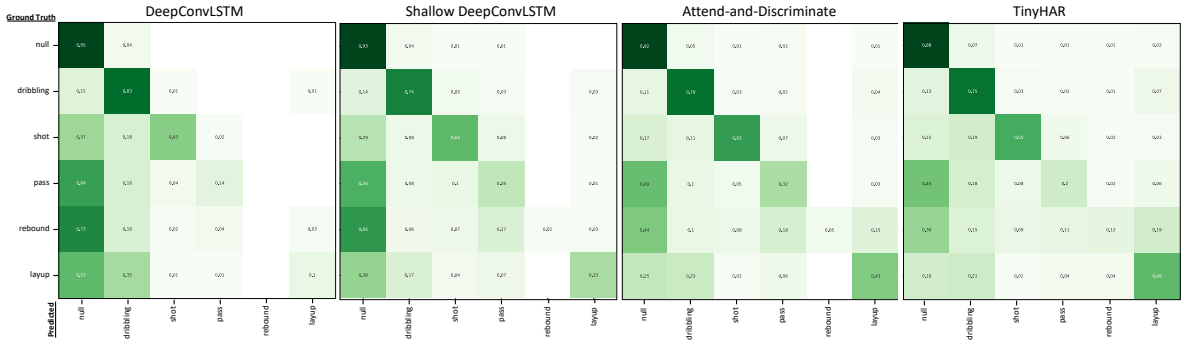

Fig. 12. Confusion matrices of the DeepConvLSTM [6], shallow DeepConvLSTM [3], Attend-and-Discriminate [1] and TinyHAR model [16] applied on the Hang-Time dataset [5] with a one second sliding window and 50% overlap. Results are postprocessed using a majority filter 15 seconds. Note that confusions which are 0 are omitted.

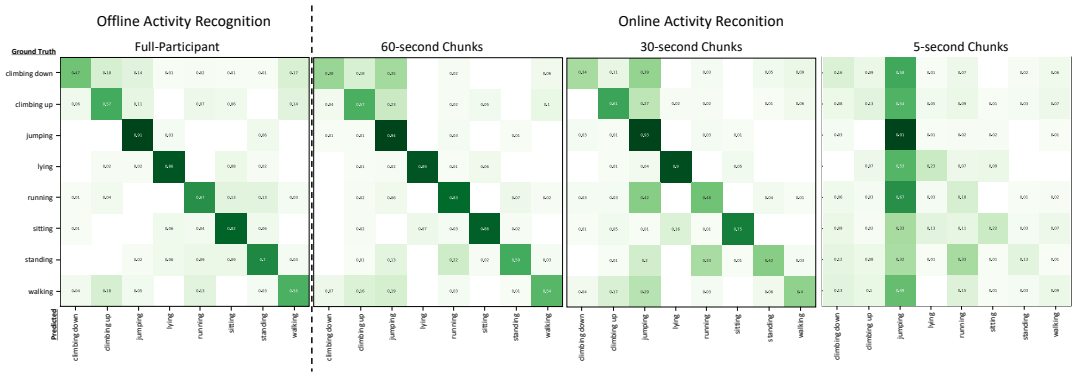

Fig. 13. Offline vs. online prediction results presented as confusion matrices of the ActionFormer [15] applied on the RWHAR dataset [11]. Illustrated are results applying a chunking of 60, 30 and 5 seconds. Note that confusions which are 0 are omitted.

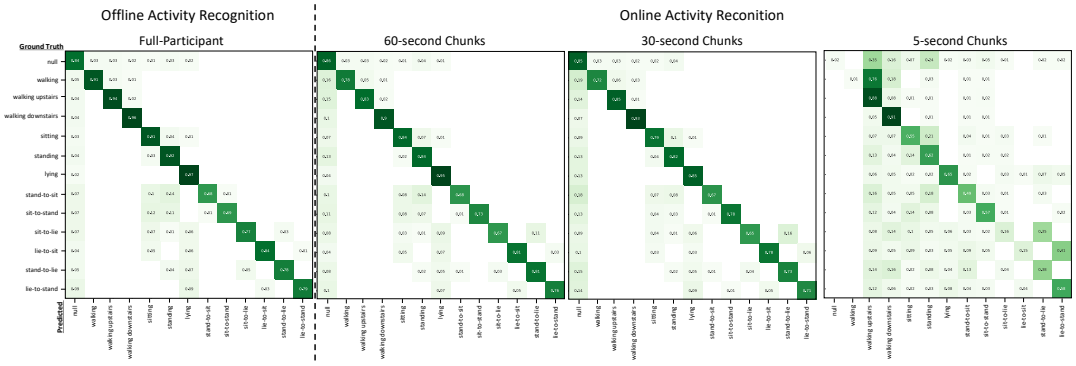

Fig. 14. Offline vs. online prediction results presented as confusion matrices of the ActionFormer [15] applied on the SBHAR dataset [7]. Illustrated are results applying a chunking of 60, 30 and 5 seconds. Note that confusions which are 0 are omitted.

C SUPPLEMENTARY RESULTS OF ONLINE ACTIVITY RECOGNITION

Figure 13 and 14 present additional confusion matrices of the ActionFormer being applied in an online prediction scenario.

D DETAD ANALYSIS

In 2018 Alwassel et al. [2] released a tool called Diagnosing Error in Temporal Action Detectors (DETAD). The tool, targeted towards analyzing temporal action localization algorithms consists of three analyses: a false positive analysis, a localization metric sensitivity analysis, and a false negative analysis. Similar to Ward et al. [13]'s efforts in the inertial community, DETAD is supposed to provide an easy-to-use tool which analyzes the performance of temporal action detectors in videos beyond a single scalar metric, i.e. mAP. Given a dataset, DETAD divides segments into bins according to three criterion: coverage, length, and the number of instances. According to coverage, being the relative size of a segment compared to the length of the complete sequence, each segment can be divided into five bins: XS: (0, 0.02], S: (0.02, 0.04], M: (0.04, 0.06], L: (0.06, 0.08], and XL: (0.08, 1.0]. Length, being the absolute length (in seconds) of segments, can be divided into five bins: XS: (0 seconds, 3 seconds], S:

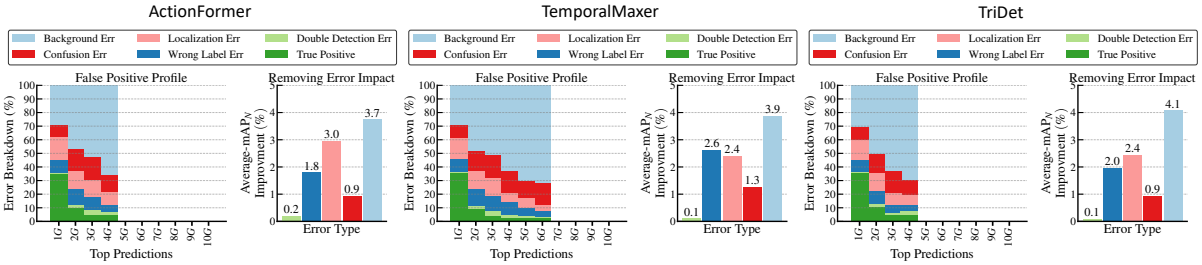

Fig. 15. DETAD False Positive (FP) analysis of the TAL architectures [10, 12, 15] on the Wetlab dataset [9]. Experiments are those described in the results chapter of the main paper with the random seed being fixed at 1. Each subplot exists of a FP error breakdown (left) and an impact analysis of error (right). The FP error breakdown is divided considering the top-10 predictions. One can see that the background error and localization error are most crucial. Furthermore, the TemproalMaxer's predictions show a larger overlap amongst each other with there being a higher degree of top-N predictions.

(3 seconds, 6 seconds], M: (6 seconds, 12 seconds], L: (12 seconds, 18 seconds], and XL: more than 18 seconds. Number of instances refers to the number of segments of the same class within a video. This can be divided into five bins: XS: 1, Small S: [2, 40], M: [40, 80], and L: > 80. Our repository integrates the scripts provided by Alwassel et al. [2] such that the DETAD analysis can easily be run on both inertial and temporal action localization models. The following provides sample results obtained on the Wetlab dataset [9]. All other plots created by the three DETAD analyses performed on the other datasets can be found in the repository.

In general, the DETAD analyses reveal that TAL models are less prone to background errors than inertial models. Furthermore, results obtained on the shallow DeepConvLSTM shows similarties to that of the TAL models, further suggesting that the inertial model is capable of learning temporal dependencies across windows. In case of the Wetlab dataset the three types of analysis (see Figure 15, 16 and 17) show that the TAL models especially struggle with predicting small segments correctly. In general, differentiation with the null class as well localizing the activities is one of the largest error sources for both models.

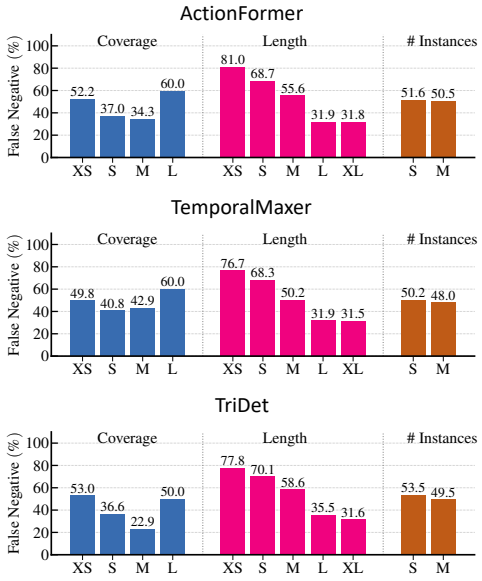

Fig. 16. DETAD False Negative (FN) analysis of the TAL architectures [10, 12, 15] on the Wetlab dataset [9]. Experiments are those described in the results chapter of the main paper with the random seed being fixed at 1. The figure shows FN rates for different types of video segments. One can see that all architectures struggle with predicting small segments correctly.

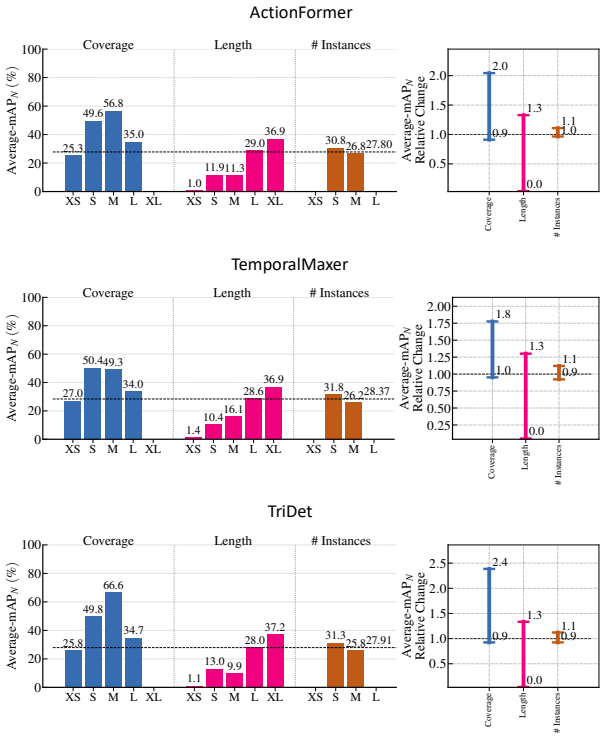

Fig. 17. DETAD Sensitivity analysis of the TAL architectures [10, 12, 15] on the Wetlab dataset [9]. Experiments are those described in the results chapter of the main paper with the random seed being fixed at 1. Each subplot exists of a plot showing the mAP across different video segments (left) and as well as visualized change in mAP for the three segment categorizations (right). One can see that especially longer segments are predicted more reliably with especially extra small segments being not recognized at all.

E ABLATION STUDY ON HYPERPARAMETERS OF TAL MODELS

The following will provide a performed ablation study on the most significant hyperparameters, as specified by the original authors of the ActionFormer, TemporalMaxer and TriDet model. All experiments were conducted the same way as specified in the main paper, being the average across three runs using a set of three random seeds. To limit the amount of experiments we only assess results obtained on the WEAR dataset. All mentioned results are postprocessed applying a score threshold of 0.2.

E.1 Maximum Input Sequence Length during Training

The TAL models investigated in this paper all aim to predict all activity segments within an untrimmed sequence in one iteration, rather than batching the sequence like the inertial models do. In order to limit complexity, all three TAL models randomly truncate input sequence to not exceed a maximum sequence length during training. The maximum sequence length therefore becomes an important hyperparameter, especially when dealing with long inital input sequences per subject. Table 4 shows results trying out different maximum sequence lenghts for each of the three architectures being applied on the WEAR dataset. For reference, given a window size of 1 second with an overlap of 50%, per-subject sequences within the WEAR dataset contain on average more than 6000 windows per participant. Results show that in case of all architectures a small maximum input sequence length results in the highest classification and average mAP scores. Particularly, the TemporalMaxer benefits the most from a smaller input sequence legnth.

Table 4. Ablation results obtained by applying different maximum sequence lengths (T_{max}) for the ActionFormer (AF) [15], TemporalMaxer (TM) [12] and TriDet (TD) model [10] on the WEAR dataset [4]. Results show that a small maximum input sequence length results in the highest classification and average mAP scores. Note that the ActionFormer cannot be trained on a maximum sequence length of 1152, as the sequence length had do be divisible by the kernel sizes of the model. Best results per architecture are in **bold**.

	T_{max}	P (↑)	R (↑)	F1 (↑)	UR (↓)	OR (↓)	DR (↓)	IR (↓)	FR (↓)	MR (↓)	Avg. mAP (↑)
AF	512	70.67	**76.88**	71.36	0.22	7.09	**0.64**	8.27	0.01	1.70	71.71
	1280*	**72.61**	76.36	**72.69**	0.20	6.28	0.70	5.73	0.01	1.47	**72.71**
	2304	72.53	74.87	72.16	**0.19**	6.11	0.77	4.83	0.01	1.44	71.42
	4608	72.00	72.69	70.94	0.21	**5.76**	0.83	**3.43**	**0.00**	**1.26**	69.31
TM	512	69.34	73.67	69.57	0.22	7.43	**0.77**	7.04	0.02	1.22	69.45
	1152	**71.02**	**74.10**	**70.88**	**0.19**	6.48	0.81	5.28	**0.01**	1.12	**71.23**
	2304	69.29	70.79	68.51	0.20	6.38	0.93	**3.70**	**0.01**	1.01	67.59
	4608	66.05	65.78	64.40	0.22	**5.83**	1.09	4.00	**0.01**	**0.97**	61.48
TD	512	71.77	**76.03**	71.25	0.26	5.58	**0.61**	9.16	0.03	1.34	**73.97**
	1152	73.87	75.73	**72.83**	**0.20**	4.59	0.70	6.75	0.03	1.41	73.39
	2304	**74.60**	73.81	72.29	0.24	4.18	0.76	4.76	0.03	1.35	72.48
	4608	72.11	70.01	69.46	0.24	**4.13**	0.90	**4.04**	**0.02**	**1.03**	68.56

E.2 Feature Pyramid

One of the central parts of the ActionFormer, TemporalMaxer and TriDet model is the scaled feature pyramid. By applying a downsampling operator all three architectures aim to learn features at different temporal scales which can then be used to classify each timestamp based on a different temporal granularity. Table 5 shows results when applying each model with different amounts of layers employed in the feature pyramid. Same as the ablation study performed by Shi et al. [10], we use both fixed and scaled number of bins when varying the number of layers in the TriDet model. Results show that the amount of layers significantly affect prediction performance across all architectures. In case of the TriDet and ActionFormer six layers, as also used in the main paper, are

resulting in highest classification and average mAP scores. In case of the TriDet model, applying a scaled number of bins does result in better performance as more bins are needed to capture activities with a longer duration.

Table 5. Ablation results obtained by trying out different number of layers within the feature pyramid when training the TAL architectures [10, 12, 15] on the WEAR dataset [4]. Results show that the amount of layers significantly affect prediction performance across all architectures, with more layers needed to achieve good prediction results. In case of the TriDet model, applying a scaled number of bins does result in better performance as more bins are needed to capture activities with a longer duration. Best results per architecture are in **bold**.

	Levels	Bins	P (↑)	R (↑)	F1 (↑)	UR (↓)	OR (↓)	DR (↓)	IR (↓)	FR (↓)	MR (↓)	Avg. mAP (↑)
ActionFormer	1	-	23.90	17.61	17.91	0.34	**4.96**	2.66	3.68	0.03	**0.08**	6.78
	2	-	34.13	26.95	27.66	0.42	7.64	2.23	4.07	0.04	0.20	13.36
	3	-	44.90	41.95	41.32	0.38	10.86	1.76	**2.83**	0.02	0.61	28.91
	4	-	55.64	55.47	53.41	0.35	10.72	1.28	4.29	0.02	0.85	40.74
	5	-	66.55	69.89	66.34	0.28	9.31	0.86	5.01	**0.01**	0.88	59.20
	6	-	**71.88**	**76.70**	**72.43**	**0.22**	6.48	0.65	7.12	**0.01**	2.06	**73.80**
	7	-	70.03	76.55	70.84	0.27	8.57	**0.60**	7.20	**0.01**	3.84	73.70
TemporalMaxer	1	-	27.81	20.41	20.84	0.43	**6.27**	2.44	6.55	0.04	**0.10**	7.93
	2	-	36.42	27.02	28.54	0.47	7.98	2.17	3.28	0.05	0.32	13.15
	3	-	42.01	38.25	37.73	0.39	10.30	1.87	4.03	0.03	0.74	24.20
	4	-	47.83	45.88	44.48	0.36	11.21	1.64	**2.55**	0.02	0.77	31.76
	5	-	58.48	57.84	56.24	0.32	9.29	1.27	2.92	0.03	0.63	46.70
	6	-	**69.54**	72.80	69.52	**0.23**	6.87	0.83	5.50	**0.01**	1.50	69.18
	7	-	68.52	**75.53**	**69.77**	0.24	8.14	**0.68**	5.61	**0.01**	4.37	**71.78**
TriDet	1	16	21.42	12.2	11.98	0.38	**1.92**	2.76	5.05	0.07	**0.01**	1.53
	2	16	34.53	22.24	24.07	0.49	4.57	2.34	3.32	0.05	0.08	9.08
	3	16	44.70	36.78	38.04	0.38	5.22	1.95	3.32	0.02	0.28	26.05
	4	16	59.86	52.80	54.01	0.31	4.51	1.42	3.14	0.06	0.19	42.45
	5	16	69.92	67.56	67.07	0.29	4.72	0.93	4.50	0.03	0.65	60.99
	6	16	**73.57**	**77.54**	**73.18**	0.28	4.79	0.64	6.21	0.03	1.64	**77.12**
	7	16	71.59	76.25	71.90	**0.26**	5.90	0.64	6.60	**0.01**	3.97	73.52
	1	512	22.16	20.18	19.13	**0.26**	5.80	2.65	5.15	0.02	0.53	11.47
	2	256	33.46	29.39	29.11	0.31	7.09	2.27	**2.76**	0.02	0.36	19.55
	3	128	44.90	38.69	39.31	0.33	5.85	1.92	3.61	0.03	0.24	29.35
	4	64	58.95	52.31	53.44	0.34	4.97	1.42	3.80	0.03	0.16	43.26
	5	32	69.71	67.29	66.89	0.29	4.97	0.93	4.63	0.03	0.67	60.42
	6	16	**73.57**	**77.54**	**73.18**	0.28	4.79	0.64	6.21	0.03	1.64	**77.12**
	7	8	72.49	76.92	72.50	0.27	5.65	**0.60**	5.91	**0.01**	4.31	73.80

E.3 Number of Bins (TriDet)

As described in the previous section, the number of bins employed in the Trident head of the TriDet model influences the maximum length at which activities can be recognized. Table 6 demonstrates that in case of the WEAR dataset an increase in bins improves both classification and average mAP scores, with the optimal value being 16.

E.4 Local Window Size for Self-Attention (ActionFormer)

The next set of experiments investigates the effect the size of the local self-attention window has on the results of the ActionFormer. As evident by Table 7 results are only marginally affected by a changing window size. Though a larger window size of 25 does results in both higher classification scores and average mAP, improvements are less than a percent for both categories compared to the original window size of 9.

Table 6. Ablation results obtained by trying out different number of bins when training the TriDet [10] on the WEAR dataset [4]. One can see that a larger bin count results in better prediction performance, with the optimal value being 16. Best results are in **bold**.

Bins	P (↑)	R (↑)	F1 (↑)	UR (↓)	OR (↓)	DR (↓)	IR (↓)	FR (↓)	MR (↓)	Avg. mAP (↑)
4	73.88	76.48	73.41	**0.24**	4.88	0.66	6.53	**0.01**	1.78	73.88
8	74.38	75.30	72.99	0.29	4.74	0.66	6.36	0.02	1.51	74.54
10	74.38	75.49	73.15	0.27	5.06	0.66	**6.08**	0.03	**1.36**	74.12
12	**74.84**	75.67	**73.54**	0.28	4.60	0.65	6.30	0.02	1.51	74.67
14	74.83	75.56	73.35	0.29	4.69	**0.64**	6.51	0.03	1.44	74.53
16	73.57	**77.54**	73.18	0.28	4.79	**0.64**	6.21	0.03	1.64	**77.12**
20	73.47	74.71	72.10	0.27	**4.59**	0.69	6.33	0.02	1.50	73.87

Table 7. Ablation results obtained by trying out different local self-attention window sizes when training the ActionFormer [15] on the WEAR dataset [4]. Though a larger window size of 25 results in both higher classification and average mAP scores, improvements are less than a percent for both compared to the original window size of 9. Best results are in **bold**.

Window Size	P (↑)	R (↑)	F1 (↑)	UR (↓)	OR (↓)	DR (↓)	IR (↓)	FR (↓)	MR (↓)	Avg. mAP (↑)
9	71.88	76.70	72.43	**0.22**	**6.48**	**0.65**	7.12	0.01	2.06	73.80
19	71.38	76.20	71.84	**0.22**	6.64	0.67	7.45	0.01	1.90	73.84
25	**72.13**	**76.81**	**72.50**	**0.22**	6.51	0.66	**6.78**	0.01	2.03	**74.07**
37	71.50	76.32	71.86	0.24	6.87	0.66	6.96	**0.00**	**1.89**	73.63
Full	71.72	76.02	71.81	**0.22**	6.62	0.68	7.42	0.01	1.91	73.26

E.5 SGP Window Sizes (TriDet)

Similarly to the local self-attention window size, the TriDet model's SGP layers are defined by two window size parameters, namely w and k. Table 8 shows results when trying out different values for w and k, fixing one of them to 1 when trying out larger values for the other hyperparameter. One can see that the optimal value for w is between 9 and 13, while the optimal value for k is around 5. Nevertheless, as for the local self-attention, differences amongst results are only marginal with the initial setup ($w = 1$ and $k = 4$) being only marginally worse than the optimized parameter values.

E.6 Loss Weight

As TAL architectures predict segments consisting of a class label and activity onsets and offsets, the models are optimized using both a classfication and regression loss. Table 9 explores different weighting schemes between the two losses. The weight λ_{reg} mentioned in the table refers to the relative weight of the regression loss in calculating the final loss of each of the three models. Results show that a larger regression loss weight results in more precise boundaries and higher mAP scores. Nevertheless, one can achieve the best trade-off between classification and mAP scores when using an unweighted, combined loss.

Table 8. Ablation results of trying out different hyperparameter values for both w and k in the SGP layers when training the TriDet [10] model on the WEAR dataset [4]. While optimal values for w and k are between 9 and 13 (w) and around 5 (k), while the optimal value for k is around 5, differences amongst results are only marginal. Best results per hyperparameter are in **bold**.

k	w	P (↑)	R (↑)	F1 (↑)	UR (↓)	OR (↓)	DR (↓)	IR (↓)	FR (↓)	MR (↓)	Avg. mAP (↑)
k = 1	1	73.91	74.92	72.49	0.30	4.74	0.65	7.32	0.03	1.53	74.54
	3	73.96	75.01	72.58	0.29	4.74	0.67	6.75	**0.01**	1.53	74.01
	5	73.83	75.01	72.67	0.28	**4.50**	0.67	7.30	0.02	1.48	74.04
	7	74.39	75.28	73.03	0.29	4.84	0.65	6.64	0.04	1.59	74.69
	9	74.99	75.68	73.51	0.29	4.87	0.65	6.63	0.02	1.47	**74.90**
	11	73.86	75.49	72.97	**0.26**	4.88	0.68	6.19	0.02	1.54	74.00
	13	**75.73**	**76.35**	**74.40**	**0.26**	4.60	0.65	5.69	0.02	1.50	74.43
	15	75.12	75.99	73.89	0.28	4.61	**0.64**	**5.50**	**0.01**	1.69	74.70
	17	73.96	75.08	72.88	0.30	4.97	0.66	6.68	**0.01**	1.55	73.89
	19	73.93	74.74	72.36	0.29	5.02	0.67	6.32	0.03	**1.40**	74.69
	21	74.33	75.23	72.67	0.30	4.88	0.64	6.21	0.03	1.76	74.68
1	w = 1	73.91	74.92	72.49	0.30	4.74	0.65	7.32	0.03	1.53	74.54
2		75.31	75.64	73.61	0.29	**4.52**	0.63	6.66	0.02	1.62	74.67
3		74.55	75.58	73.42	0.27	4.92	0.66	**5.64**	0.02	1.43	74.03
4		73.84	74.80	72.36	0.29	4.78	0.67	6.63	0.03	1.45	74.29
5		**75.54**	**76.14**	**74.03**	0.27	4.87	0.62	6.53	0.04	**1.34**	74.03
6		74.27	75.38	73.02	0.28	4.96	0.66	6.46	0.02	1.48	74.54
7		74.29	75.46	73.01	0.28	4.73	0.64	6.68	0.03	1.39	74.31
8		74.63	75.87	73.50	**0.26**	4.68	0.67	5.89	0.02	1.69	74.72
9		74.75	75.74	73.50	0.28	4.71	0.64	6.90	0.03	1.44	74.68
10		74.75	75.28	73.00	0.28	4.80	0.66	6.04	0.02	1.36	74.32
11		75.30	75.83	73.34	0.32	4.81	**0.60**	6.49	**0.01**	1.68	**75.05**

Table 9. Ablation results on trying out different relative regression loss weights λ_{reg} when training the three TAL architectures [10, 12, 15] model on the WEAR dataset [4]. Once can see that a higher regression loss weight results in more precise boundaries, though the best tradeoff amongst classification and mAP scores can be achieved when using an unweighted combination of regression and classification loss. Best results per architecture are in **bold**.

	λ_{reg}	P (↑)	R (↑)	F1 (↑)	UR (↓)	OR (↓)	DR (↓)	IR (↓)	FR (↓)	MR (↓)	Avg. mAP (↑)
ActionFormer	0.5	71.40	76.61	72.21	**0.21**	6.86	0.66	7.21	**0.01**	1.95	73.64
	0.2	71.9	76.36	72.03	0.23	6.97	0.65	**6.45**	**0.01**	**1.84**	73.06
	1	71.88	76.70	72.43	0.22	**6.48**	0.65	7.12	**0.01**	2.06	73.80
	2	**72.73**	**76.96**	**72.72**	0.24	6.68	**0.64**	6.87	**0.01**	2.07	73.58
	5	71.47	76.86	71.97	0.22	6.49	0.66	7.96	**0.01**	1.96	**74.19**
T.Maxer	0.5	68.81	71.59	68.44	0.23	6.99	0.86	**5.22**	**0.01**	**1.46**	68.03
	0.2	68.67	71.57	68.23	0.25	7.60	0.84	5.96	**0.01**	**1.46**	68.11
	1	69.54	72.80	69.52	0.23	6.87	0.83	5.50	**0.01**	1.50	69.18
	2	69.44	73.55	**69.80**	**0.22**	**6.50**	0.80	6.35	**0.01**	1.63	70.82
	5	**69.61**	**73.98**	69.49	0.25	6.99	**0.73**	7.94	0.02	1.72	**71.00**
TriDet	0.5	71.46	74.87	70.92	0.29	7.14	0.65	6.44	0.02	1.73	72.74
	0.2	68.62	74.47	68.92	**0.27**	8.44	0.68	6.93	**0.01**	2.40	70.30
	1	73.57	**77.54**	**73.18**	0.28	4.79	**0.64**	**6.21**	0.03	1.64	77.12
	2	74.16	75.29	72.61	**0.27**	**4.65**	0.66	7.09	0.02	1.56	74.44
	5	**74.32**	75.82	72.93	**0.27**	4.98	**0.64**	7.15	0.03	**1.46**	**74.82**

REFERENCES

[1] Alireza Abedin, Mahsa Ehsanpour, Qinfeng Shi, Hamid Rezatofighi, and Damith C. Ranasinghe. 2021. 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 (2021), 1–22. https://doi.org/10.1145/3448083

[2] Humam Alwassel, Fabian Caba Heilbron, Victor Escorcia, and Bernard Ghanem. 2018. Diagnosing Error in Temporal Action Detectors. In *European Conference on Computer Vision*. https://doi.org/10.1007/978-3-030-01219-9_16

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

[4] Marius Bock, Hilde Kuehne, Kristof Van Laerhoven, and Michael Moeller. 2023. WEAR: An Outdoor Sports Dataset for Wearable and Egocentric Activity Recognition. *CoRR* abs/2304.05088 (2023). https://arxiv.org/abs/2304.05088

[5] Alexander Hoelzemann, Julia L. Romero, Marius Bock, Kristof Van Laerhoven, and Qin Lv. 2023. Hang-Time HAR: A Benchmark Dataset for Basketball Activity Recognition Using Wrist-Worn Inertial Sensors. *MDPI Sensors* 23, 13 (2023). https://doi.org/10.3390/s23135879

[6] Francisco Javier Ordóñez and Daniel Roggen. 2016. Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. *MDPI Sensors* 16, 1 (2016). https://doi.org/10.3390/s16010115

[7] Jorge-L. Reyes-Ortiz, Luca Oneto, Albert Samà, Xavier Parra, and Davide Anguita. 2016. Transition-Aware Human Activity Recognition Using Smartphones. *Neurocomputing* 171 (2016). https://doi.org/10.1016/j.neucom.2015.07.085

[8] Daniel Roggen, Alberto Calatroni, Mirco Rossi, Thomas Holleczek, Kilian Förster, Gerhard Tröster, Paul Lukowicz, David Bannach, Gerald Pirkl, Alois Ferscha, Jakob Doppler, Clemens Holzmann, Marc Kurz, Gerald Holl, Ricardo Chavarriaga, Hesam Sagha, Hamidreza Bayati, Marco Creatura, and José del R. Millàn. 2010. Collecting Complex Activity Datasets in Highly Rich Networked Sensor Environments. In *IEEE Seventh International Conference on Networked Sensing Systems*. https://doi.org/10.1109/INSS.2010.5573462

[9] Philipp M. Scholl, Matthias Wille, and Kristof Van Laerhoven. 2015. Wearables in the Wet Lab: A Laboratory System for Capturing and Guiding Experiments. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing*. https://doi.org/10.1145/2750858.2807547

[10] Dingfeng Shi, Yujie Zhong, Qiong Cao, Lin Ma, Jia Li, and Dacheng Tao. 2023. TriDet: Temporal Action Detection With Relative Boundary Modeling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/cvpr52729.2023.01808

[11] Timo Sztyler and Heiner Stuckenschmidt. 2016. On-Body Localization of Wearable Devices: An Investigation of Position-Aware Activity Recognition. In *IEEE International Conference on Pervasive Computing and Communications*. https://doi.org/10.1109/PERCOM.2016.7456521

[12] Tuan N. Tang, Kwonyoung Kim, and Kwanghoon Sohn. 2023. TemporalMaxer: Maximize Temporal Context With Only Max Pooling for Temporal Action Localization. *CoRR* abs/2303.09055 (2023). https://arxiv.org/abs/2303.09055

[13] Jamie A. Ward, Paul Lukowicz, and Hans W. Gellersen. 2011. Performance Metrics for Activity Recognition. *ACM Transactions on Intelligent Systems and Technology* 2, 1 (2011). https://doi.org/10.1145/1889681.1889687

[14] Jamie A. Ward, Paul Lukowicz, and Gerhard Tröster. 2006. Evaluating Performance in Continuous Context Recognition Using Event-Driven Error Characterisation. In *Location- and Context-Awareness*. https://doi.org/10.1007/11752967_16

[15] Chen-Lin Zhang, Jianxin Wu, and Yin Li. 2022. Actionformer: Localizing Moments of Actions With Transformers. In *European Conference on Computer Vision*. https://doi.org/10.1007/978-3-031-19772-7_29

[16] Yexu Zhou, Haibin Zhao, Yiran Huang, Till Riedel, Michael Hefenbrock, and Michael Beigl. 2022. TinyHAR: A Lightweight Deep Learning Model Designed for Human Activity Recognition. In *ACM International Symposium on Wearable Computers*. https://doi.org/10.1145/3544794.3558467