UNIK: A Unified Framework for Real-world Skeleton-based Action Recognition - Supplementary Material

Di Yang*[1]
di.yang@inria.fr
Yaohui Wang*[1]
yaohui.wang@inria.fr
Antitza Dantcheva[1]
antitza.dantcheva@inria.fr
Lorenzo Garattoni[2]
lorenzo.garattoni@toyota-europe.com
Gianpiero Francesca[2]
gianpiero.francesca@toyota-europe.com
François Brémond[1]
francois.bremond@inria.fr

[1] Inria
Université Côte d'Azur
Valbonne, France

[2] Toyota Motor Europe
Brussels, Belgium

Appendix

Overview. In this supplementary material (SM) we provide additional details *w.r.t.*, our experimental analysis provided in the main paper. In section A, we provide details pertaining to the datasets, the implementation of our framework, as well as additional results and analysis, containing 4 parts. (i) An ablation study of the initialization strategy of the proposed *Dependency Matrix*. (ii) *Multi-stream fusion* in UNIK. (iii) An evaluation of pre-training UNIK on Posetics by *Linear transfer learning* (backbone fixed). (iv) The results on NTU-60 with the full model and all provided 25 joints to compare with state-of-the-art. In section B, we provide details on the proposed Posetics dataset (see Fig. 3) including the comparison to other related datasets. Section C provides the visualization of (a) the confusion matrices on Smarthome from our proposed framework, as well as (b) the skeletons we extract for Posetics. Finally, section D shows the list of all the 320 action categorizations together with their number of clips incorporated in the Posetics dataset.

A Experimental Details

A.1 Datasets

Toyota Smarthome. Toyota Smarthome [4] (Smarthome) is a real-world dataset for daily living action classification, recorded in an apartment, where 18 older subjects carry out tasks of daily living during a day. The dataset contains 16,115 videos of 31 action classes, and the videos are taken from 7 different camera viewpoints. All actions are performed in a natural way without strong prior instructions. It provides RGB videos and two versions of skeleton

UNIK	Pre-training	Smarthome (J+B) #Params	CS(%)	CV1(%)	CV2(%)	Penn Action (J+B) #Params	Accuracy (%)
Fine-tuning	Scratch	3.45M	63.1	22.9	61.2	3.45M	94.0
Fine-tuning	Posetics	3.45M	**64.3**	**36.1**	**65.2**	3.45M	**97.9**
Linear classification	Scratch	7.97K	24.6	17.2	20.7	3.85K	29.8
Linear classification	Posetics	7.97K	**51.9**	**35.2**	**52.2**	3.85K	**97.3**

Table 1: Mean per-class accuracy on Smarthome and Top-1 classification accuracy on Penn Action by *Fine-tuning* (Backbone not fixed) and *Linear classification* (Backbone fixed) for evaluation of extracted features by pre-training. "J+B": Joint and Bone two stream fusion.

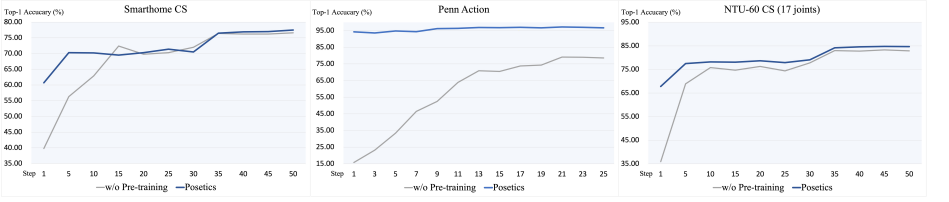

Figure 1: Validation accuracy with the training steps on Smarthome, Penn Action and NTU-60 datasets for demonstrating the impact of Pre-training on Posetics.

data, which is extracted either from LCRNet++ [13] (v1) or from SSTA-PRS [18] (v2). In this work, we use the skeleton-v2 for all experiments and comparisons. Unless stated, we only use 2D data for the experiments. For the evaluation on this dataset, we follow the cross-subject (CS) and cross-view (CV1 and CV2) protocols.

Penn Action. Penn Action dataset [19] contains 2,326 video sequences of 15 different actions and human joint annotations for each sequence. Given that annotated skeletons have a large number of missing joints due to occlusions and truncations, we use LCRNet++ [13] to obtain the full-body 2D skeletons for experiments. We report Top-1 accuracy following the standard train-test split.

NTU-RGB+D 60. NTU-RGB+D 60 [14] (NTU-60) is a large-scale multi-modality dataset which consists of 56,880 sequences of high-quality 2D and 3D skeletons with 25 joints, associated with depth maps, RGB and IR frames captured by the Microsoft Kinect v2 sensor. We only use sequences of 3D skeletons in this work and we follow the cross-subject (CS) and cross-view (CV) evaluation protocol. For ablation study (see A.3 and 4.2 in main paper), we use all provided 25 joints while for evaluation of pre-training (see B and 4.3 in main paper), use only 17 joints to adapt to the pre-trained model on Posetics.

NTU-RGB+D 120. NTU-RGB+D 120 [10] (NTU-120) dataset extends the number of action classes and videos of NTU-RGB+D 60 to 120 classes 114,480 videos. Similarly, we use only sequences of 3D skeletons and we follow the cross-subject (CS) and cross-set (CSet) evaluation protocols. Similarly, for evaluation of pre-training (see 4.3 in main paper), use only 17 joints to adapt to the pre-trained model on Posetics.

A.2 Implementation Details

Implementation of UNIK. Unless otherwise stated in the ablation study, all UNIK models have $N = 3, \tau = 1$ for S-LSU, and $t = 9, d = 1, 3, 3, 3, 3, 1, 1, 1, 1, 1, 1$, in each block respectively for T-LSU. We use SGD for training with momentum 0.9, an initial learning rate of 0.1 for 50, 30, 50, 60, and 65 epochs with step LR decay with a factor of 0.1 at epochs {30, 40}, {10, 20}, {30, 40}, {30, 50}, and {45, 55} for Smarthome, Penn Action, NTU-60, NTU-120, and Posetics, respectively. Weight decay is set to 0.0001 for final models. For NTU-60 and 120, all skeleton sequences are padded to 300 frames by replaying the actions. For

Initialization of DepM	Smarthome CS (%)	NTU-60 CS (%)
Human-intrinsic topology	55.9	86.1
Zeros init.	56.9	86.6
Ones init.	55.5	86.2
Uniform init.		
$a = 1$	58.2	87.1
$a = \sqrt{5}$	**58.5**	**87.3**
$a = \sqrt{10}$	57.9	86.8

Table 2: Mean per-class accuracy on Smarthome CS and Classification accuracy on NTU-60 CS using joint data only. Note that we set $\tau = 1, N = 3$ for this ablation study.

Methods	NTU-RGB+D 60	
	CS(%)	CV(%)
ST-GCN [47]	81.5	88.3
2s-AGCN [45]	88.5	95.1
MS-G3D Net [41]	91.5	96.2
UNIK w/o pt. (Ours)	89.5	95.7
MS-G3D Net pt. on Posetics (Ours)	**91.7**	96.2
UNIK pt. on Posetics (Ours)	89.7	**96.4**

Table 3: Top-1 accuracy comparison with state-of-the-art on the NTU-60 Skeleton dataset.

Smarthome, Penn Action, Posetics, we randomly choose 400, 150, 150 frames respectively for each training epoch and all frames for test. 2D and 3D inputs are pre-processed with normalization and centering following [12], [15] respectively. As we have both 2D and 3D skeleton data on Posetics, we pre-train two models for transferring to benchmarks with different types of skeleton data. Note that for ablation study of UNIK (see 4.2 in main paper) and A.3), we train all models from scratch, without pre-training.

Number of Joints. SSTA-PRS [28] and LCRNet++ [34] provide 13 joints of the main body. We add "hip", "chest", "neck" and "nose" by interpolation and obtain 17 joints for all experiments of real-world datasets (*i.e.*, Posetics, Smarthome, Penn Action). On NTU-60 and 120, we use 3D Kinect skeleton data with 25 joints for ablation study of UNIK (Sec. 5.2) while 17 main body joints for generalizability study (Sec. 5.3) to adapt to the pre-trained model on Posetics.

A.3 Additional Results and Analysis

Ablation Study: Initialization of Dependency Matrix. The proposed UNIK is to learn an optimal Dependency Matrix to extract skeleton features. To explore the impact of initialization of the Dependency Matrix, in this section, we experiment on Smarthome CS and NTU-60 CS with zeros initialization, ones initialization, human topological initialization (*i.e.*, Adjacency Matrix in GCNs) and uniform initialization (proposed in this paper) with different settings of the hyper-parameter a (see Eq. 3). The results in Tab. 2 verify the analysis in 3.2 (main paper): the fully dense uniform initialization is the most effective to reach the global optimal representation of the dependency matrix.

Ablation Study: Joint-Bone-Motion Fusion. In this section, we demonstrate the impact of the two stream fusion of joint and bone data. The model, pre-trained on joint data, can be instrumental in improving performance on bone-stream to obtain better two-stream fusion results. Besides, additional motion data (*i.e.* joint-motion and bone-motion) can also provide minor improvement which is not as significant as the bone data because our proposed Temporal Unit can effectively process the motion information. Related results are summarized in Tab. 6.

Feature Evaluation. In this section, we evaluate the features of skeletons extracted by UNIK (fixed) pre-trained on Posetics. Unlike the fine-tuning in the main paper, we freeze

Dataset	RW	2D	3D	#Videos	#Classes	#Joints	Skeleton data	Skeleton quality	Year
Human3.6M [3]	×	✓	✓	209	15	24	Motion Capture system	High	2014
NTU-RGB+D 60 [14]	×	✓	✓	56,880	60	25	Kinect v2.0	High	2016
NTU-RGB+D 120 [10]	×	✓	✓	114,480	120	25	Kinect v2.0	High	2019
Penn Action [19]	✓	✓	×	2,326	15	13	Handcrafted annotation	Medium	2013
UAV-Human [9]	✓	✓	×	21,224	155	17	AlphaPose [5]	Medium	2021
Kinetics-Skeleton [17]	✓	✓	×	260,232	400	18	OpenPose [1]	Low	2018
Toyota Smarthome [4]	✓	✓	✓	16,115	31	13	LCRNet++ [23]	Low	2019
Skeletics-152 [6]	✓	×	✓	125,621	152	25	VIBE [8]	High	2021
Posetics (Ours)	✓	✓	✓	142,000	320	25	SSTA-PRS [23]+Intrpl.	High	2021

Table 4: A survey of recent real-world datasets for skeleton-based action recognition."RW": Real-world. "Intrpl.": Interpolation.

the UNIK backbone pre-trained on Posetics, then retrain linear classifiers on smaller benchmarks, Smarthome and Penn Action. The results in Tab. 1 demonstrate the effectiveness of transfer learning with fewer parameters compared with classification from scratch. Fine-tuning (*i.e.*, UNIK backbone not fixed) results are also shown for reference. In addition, we visualize the training curve for the Top-1 accuracy with training steps during fine-tuning (see Fig. 1). From the curves, we deduce that at the beginning of training steps, the pre-training has a significant boost for all transferred datasets. This suggests that the weights of the model are well pre-trained on Posetics, providing a strong transfer ability *i.e.*, pre-trained on Posetics is generic and can be used for extracting features of skeleton sequences.

Results with 25 Joints on NTU-60. With 25 joints, our model achieves competitive performance on NTU-60 CS (see Tab. 3). We argue that, (i) we simplify our model as generically as possible without data-specific settings that can improve the performance but weaken the transfer behavior. (ii) Unlike the 13-to-17 joints that can be obtained by median interpolation, 17-to-25 interpolation is roughly inserted according to the proportion of the human body. Therefore, the additional 8 joints on fingers and feet are not as exact as the Kinect data and related results are inferior to the ones (see 5.3 in main paper) using NTU 17 joints. In order to achieve a higher performance, we can also pre-train an additional MS-G3D Net [11] on Posetics, with the settings adapted to NTU-60 videos. Fine-tuning minor improves results, which constitutes state-of-the-art.

B Details of Posetics Dataset

In this section, we first review the recent skeleton-based action recognition datasets and then compare the most impressive ones with Posetics by (i) the benefit of pre-training for smaller datasets and (ii) skeleton (*i.e.*, pose) quality obtained by different pose estimators.

Review of Skeleton Datasets. Tab. 4 shows an overview of pertinent skeleton-based action recognition datasets, which we proceed to describe. To evaluate methods in 3D human action recognition, [10, 14, 16] were recorded in laboratory conditions, where acquired actions were performed by actors under strict guidance. In contrast, [4, 6, 9, 17, 19] aimed to explore real-world action recognition using estimated or handcrafted annotated skeleton data. As the largest real-world dataset, Kinetics-Skeleton [17] provides poses extracted using OpenPose [1]. GCNs were applied on the real-world videos [2] using pseudo 3D data *i.e.*, 2D data and confidence [1]. However, the pose quality is limited due to occlusions and truncations. Skeletics-152 [6] addressed this issue by applying VIBE [8] to obtain higher-quality skeleton data from Kinetics-700 [3] and manually scaled down the dataset by omitting the object-oriented action categories. However, there are still videos with missing skeletons. Deviating from the above, the higher-quality skeleton data in our dataset is calculated using

UNIK	Posetics(JB) Top-1(%)	Posetics(JB) Top-5(%)	Smarthome(J) CS(%)
OpenPose [1]	45.9	69.5	-
VIBE-Pose [8]	-	-	42.5
VIBE-Mesh [8]	-	-	43.2
SSTA-PRS [18]	**47.6**	**71.3**	**58.9**

Table 5: Classification accuracy of UNIK using different poses on Posetics and Smarthome.

Methods	Pre-training	Smarthome CS (%)	Smarthome CV1 (%)	Smarthome CV2 (%)	Penn Action Top-1 Acc. (%)	*NTU-60 CS(%)
UNIK-J	NTU-RGB+D 120	59.2	28.3	59.7	91.7	-
UNIK-J	Kinetics-Skeleton	58.9	29.5	60.6	95.4	†82.5
UNIK-J	Skeletics-152	59.2	-	-	91.2	-
UNIK-J	Posetics (Ours)	**62.1**	**33.4**	**63.6**	**97.2**	**85.3**
UNIK-B	Posetics (Ours)	61.1	31.3	62.5	97.4	84.9
UNIK-J&B	Posetics (Ours)	64.3	36.1	65.0	97.9	86.8
UNIK-J&B&M	Posetics (Ours)	**64.5**	**36.2**	**65.1**	**98.0**	**87.1**

Table 6: Comparison of datasets by pre-training (top) and impact of two-stream fusion (bottom). "J"/"B"/"M": Joint/Bone/Motion stream. "†": The input data (2D) is different from other competitors (3D) on NTU-60 due to the lack of 3D data on Kinetics-Skeleton. "*": We only use 17 main joints adapted to the pre-trained model on Posetics.

pose refinement method [18] which aims at real-world pose estimation [2]. Instead of omitting the samples in object-oriented action categories, we merge action categories that are incompatible with skeleton-based action recognition to keep the scale and we filter out the non-skeleton videos. Our large-scale dataset can be more effectively used for pre-training and transferring onto other simulated [10, 14] and real-world [4, 19] scenarios.

Selection of Pose Estimators. In this section, we compare the proposed Posetics with Kinetics-Skeleton [17] and Skeletics-152 [6] by the pose (*i.e.*, skeleton) quality. While [17], [6] use OpenPose [1] and VIBE [8] respectively to estimate poses, Posetics uses SSTA-PRS [18] that integrates the advantages of three pose estimators including OpenPose [1]. Hence, Posetics has higher quality poses, in particular in cases of occlusions and truncations (see Fig. 4 and Fig. 5 for qualitative comparison). For quantitative comparison, we lack ground-truth poses, and hence we indirectly evaluate the quality of poses through the performance of action recognition. Towards this, we use all clips of Posetics and Smarthome with different 2D pose data for action recognition. Experimental results in Tab. 5 show that the performance using SSTA-PRS is higher than that using other pose estimators. This motivates us to select SSTA-PRS for pose extraction on Kinetics videos.

Comparison of Pre-training. In this section, we compare Posetics with NTU-120 [10], Kinetics-Skeleton [17] and Skeletics-152 [6] datasets by fine-tuning performances after pre-training. We note that we only have 2D skeletons on Kinetics-Skeleton for pre-training. Consequently, we use 2D data of NTU-60 for fine-tuning. Results in Tab. 6 demonstrate the effectiveness of our Posetics dataset compared to other datasets [6, 10, 17] (*i.e.*, pre-training on Posetics boosts the most on target datasets).

C Visualization

In this section, we visualize the *confusion matrices* of action classification on Smarthome cross-subject to compare the 2s-AGCN [15], UNIK, and UNIK with pre-training on Posetics (see Fig. 2). Next, we present a qualitative comparison of pose quality in Posetics and Kinetics-Skeleton (see Fig. 4, 5).

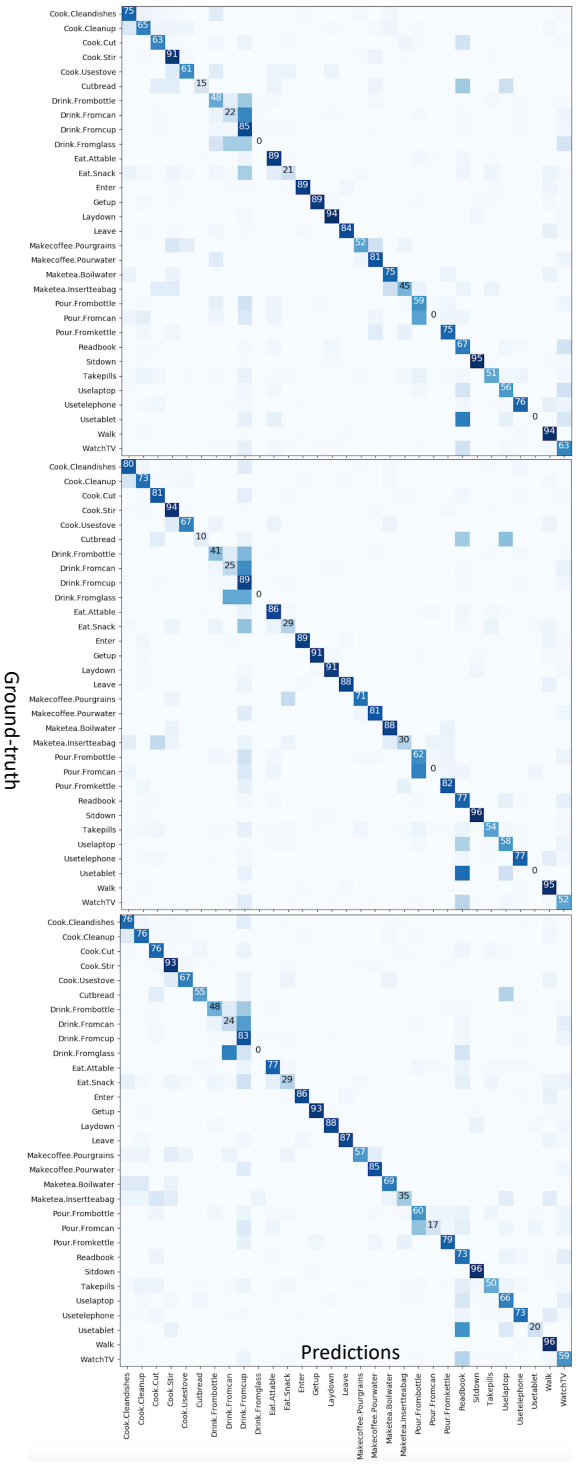

Figure 2: **Confusion matrices of action classification (%) on Smarthome-CS.** Comparison of 2s-AGCN (top), UNIK (center) and UNIK with pre-training on Posetics (bottom).

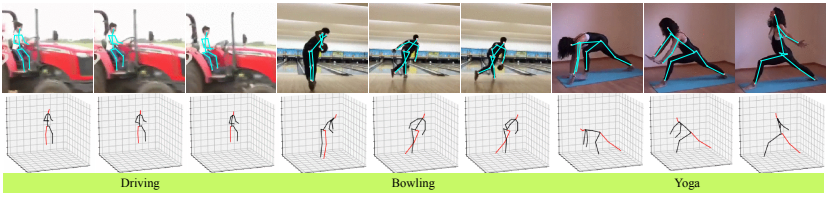

Figure 3: **Posetics**: the proposed large-scale, real-world skeleton dataset.

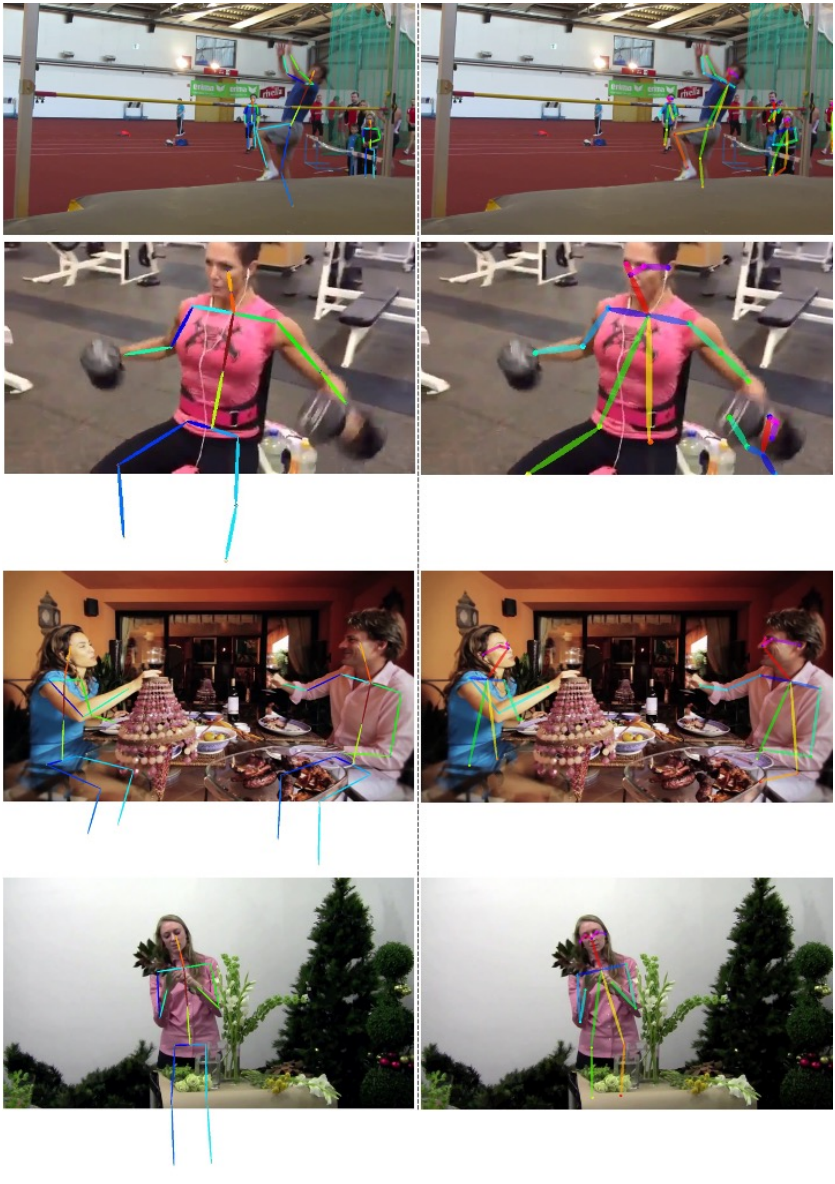

Figure 4: **Visualization of pose data-I** in Posetics (left) and Kinetics-Skeleton (right) in the case of occlusions and truncations.

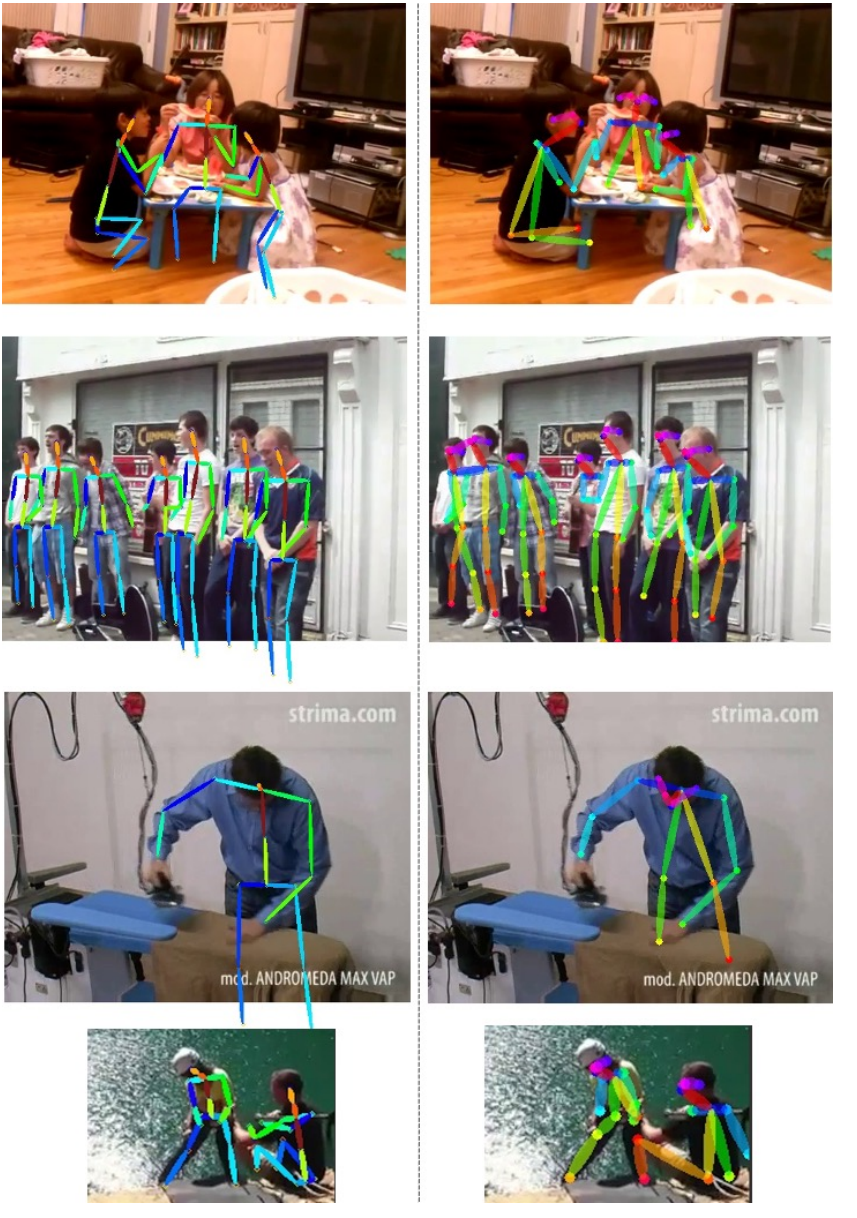

Figure 5: **Visualization of pose data-II** in Posetics (left) and Kinetics-Skeleton (right) in the case of occlusions and truncations.

D List of Posetics Human Action Classes

This is the list of classes included in the Posetics dataset. The number of clips for each action class is given by the number in brackets following each class name.

1. abseiling (493)
2. air drumming (567)
3. answering questions (250)
4. applauding (189)
5. applying cream (221)
6. archery (620)
7. arm wrestling (751)
8. arranging flowers (311)
9. assembling computer (54)
10. auctioning (294)
11. baby waking up (234)
12. baking cookies (255)
13. balloon blowing (467)
14. bandaging (168)
15. barbequing (301)
16. bartending (270)
17. beatboxing (614)
18. bee keeping (130)
19. belly dancing (395)
20. bench pressing (734)
21. bending back (344)
22. bending metal (99)
23. bik (719)
24. blasting sand (227)
25. blowing (1906)
26. bobsledding (177)
27. bookbinding (115)
28. bouncing on trampoline (402)
29. bowling (543)
30. breading or breadcrumbing (55)
31. breakdancing (482)
32. brush painting (104)
33. brushing teeth (815)
34. building (266)
35. bungee jumping (172)
36. busking (609)
37. canoeing or kayaking (414)
38. capoeira (790)
39. carrying baby (376)
40. cartwheeling (341)
41. carving pumpkin (218)
42. catching fish (270)
43. celebrating (281)
44. changing oil (100)
45. changing wheel (130)
46. checking tires (171)
47. cheerleading (717)
48. chopping wood (590)
49. clapping (322)
50. clay pottery making (120)
51. clean (1798)
52. climbing (1053)
53. contact juggling (479)
54. cooking (315)
55. counting money (196)

56. country line dancing (301)
57. cracking neck (158)
58. crawling baby (840)
59. crossing river (259)
60. crying (530)
61. cutting (342)
62. dancing ballet (493)
63. dancing charleston (376)
64. dancing gangnam style (247)
65. dancing macarena (312)
66. deadlifting (600)
67. decorating the christmas tree (287)
68. digging (140)
69. dining (362)
70. disc golfing (253)
71. diving cliff (206)
72. dodgeball (365)
73. doing aerobics (232)
74. doing laundry (237)
75. drawing (22)
76. dribbling basketball (677)
77. drinking (858)
78. driving (576)
79. drop kicking (367)
80. drumming fingers (71)
81. dunking basketball (530)
82. eating (1891)
83. egg hunting (337)
84. exercising arm (274)
85. exercising with an exercise ball (262)
86. extinguishing fire (190)
87. faceplanting (161)
88. feeding (1265)
89. filling eyebrows (257)
90. finger snapping (456)
91. flipping pancake (345)
92. flying kite (258)
93. folding (541)
94. front raises (750)
95. frying vegetables (43)
96. garbage collecting (125)
97. gargling (223)
98. getting a tattoo (243)
99. giving or receiving award (720)
100. golf chipping (493)
101. golf driving (608)
102. golf putting (688)
103. grinding meat (63)
104. grooming (493)
105. guitar (1204)
106. gymnastics tumbling (613)
107. hair (2506)
108. headbanging (520)
109. headbutting (315)
110. high jump (534)
111. high kick (517)
112. hitting baseball (797)
113. hockey stop (253)
114. holding snake (183)
115. hopscotch (534)
116. hoverboarding (193)

117. hugging (292)
118. hula hooping (687)
119. hurdling (285)
120. hurling (sport) (441)
121. ice climbing (313)
122. ice fishing (203)
123. ice skating (695)
124. ironing (178)
125. jetskiing (244)
126. jogging (185)
127. juggling (1064)
128. jumping into pool (486)
129. jumpstyle dancing (235)
130. kicking (702)
131. kissing (204)
132. kitesurfing (117)
133. knitting (89)
134. krumping (412)
135. laughing (589)
136. laying bricks (176)
137. long jump (614)
138. lunge (556)
139. making (889)
140. making bed (340)
141. making jewelry (76)
142. making snowman (427)
143. marching (647)
144. massaging (808)
145. milking cow (508)
146. mopping floor (308)
147. motorcycling (330)
148. moving furniture (230)
149. mowing lawn (537)
150. nails (313)
151. news anchoring (197)
152. opening bottle (307)
153. opening present (615)
154. paragliding (118)
155. parasailing (138)
156. parkour (137)
157. passing American football (1238)
158. peeling (318)
159. petting (449)
160. picking fruit (395)
161. planting trees (275)
162. plastering (162)
163. playing accordion (647)
164. playing badminton (685)
165. playing bagpipes (615)
166. playing basketball (722)
167. playing cards (75)
168. playing cello (755)
169. playing chess (423)
170. playing controller (37)
171. playing cricket (517)
172. playing cymbals (302)
173. playing didgeridoo (589)
174. playing drums (534)
175. playing harmonica (733)
176. playing harp (850)
177. playing ice hockey (566)

178. playing keyboard (372)
179. playing kickball (289)
180. playing monopoly (222)
181. playing organ (360)
182. playing paintball (345)
183. playing piano (375)
184. playing poker (488)
185. playing recorder (785)
186. playing squash or racquetball (727)
187. playing tennis (723)
188. playing trombone (802)
189. playing ukulele (784)
190. playing violin (769)
191. playing volleyball (552)
192. playing wind music (2277)
193. playing xylophone (541)
194. pole vault (561)
195. presenting weather forecast (714)
196. pull ups (797)
197. pumping fist (509)
198. pumping gas (196)
199. punching bag (667)
200. punching person (boxing) (258)
201. push up (403)
202. pushing (1508)
203. reading (938)
204. recording music (123)
205. riding (2231)
206. riding scooter (378)
207. riding unicycle (540)
208. ripping paper (343)
209. robot dancing (457)
210. rock climbing (590)
211. rock scissors paper (194)
212. roller skating (440)
213. running on treadmill (199)
214. sailing (162)
215. salsa dancing (480)
216. sanding floor (187)
217. scrambling eggs (82)
218. scuba diving (165)
219. setting table (167)
220. shaking hands (350)
221. shaking head (615)
222. sharpening knives (84)
223. sharpening pencil (149)
224. shaving head (481)
225. shaving legs (125)
226. shearing sheep (615)
227. shining shoes (181)
228. shooting basketball (337)
229. shooting goal (soccer) (135)
230. shot put (796)
231. shoveling snow (503)
232. shredding paper (56)
233. shuffling cards (194)
234. side kick (734)
235. sign language interpreting (193)
236. singing (514)
237. situp (556)
238. skateboarding (440)

239. ski (1060)
240. skipping rope (219)
241. skydiving (57)
242. slacklining (385)
243. slapping (202)
244. sled dog racing (395)
245. smoking (681)
246. snatch weight lifting (736)
247. sneezing (272)
248. sniffing (150)
249. snorkeling (176)
250. snowboarding (224)
251. snowkiting (147)
252. snowmobiling (129)
253. somersaulting (606)
254. spinning poi (331)
255. spray (358)
256. springboard diving (126)
257. squat (852)
258. sticking tongue out (451)
259. stomping grapes (225)
260. stretching arm (504)
261. stretching leg (579)
262. surfing crowd (128)
263. surfing water (208)
264. sweeping floor (353)
265. swimming backstroke (316)
266. swimming breast stroke (272)
267. swimming butterfly stroke (191)
268. swing dancing (232)
269. swinging legs (259)
270. swinging on something (223)
271. sword fighting (269)
272. tai chi (681)
273. taking a shower (107)
274. tango dancing (563)
275. tap dancing (506)
276. tapping guitar (420)
277. tapping pen (103)
278. tasting (745)
279. testifying (336)
280. texting (212)
281. throw (1652)
282. tickling (244)
283. tobogganing (492)
284. tossing coin (209)
285. tossing salad (106)
286. training dog (241)
287. trapezing (227)
288. trimming or shaving beard (534)
289. trimming trees (187)
290. triple jump (563)
291. tying (628)
292. unboxing (94)
293. unloading truck (145)
294. using computer (176)
295. using remote controller (not gaming) (66)
296. using segway (242)
297. vault (380)
298. waiting in line (233)
299. walking the dog (463)

300. washing dishes (596)
301. washing feet (475)
302. washing hair (124)
303. washing hands (375)
304. water skiing (220)
305. water sliding (141)
306. watering plants (371)
307. waxing back (151)
308. waxing chest (318)
309. waxing eyebrows (227)
310. waxing legs (300)
311. weaving basket (172)
312. welding (147)
313. whistling (184)
314. windsurfing (223)
315. wrapping present (319)
316. wrestling (292)
317. writing (112)
318. yawning (174)
319. yoga (593)
320. zumba (485)

References

[1] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE TPAMI*, 2019.

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

[3] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv:1907.06987*, 2019.

[4] Srijan Das, Rui Dai, Michal Koperski, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota smarthome: Real-world activities of daily living. In *ICCV*, 2019.

[5] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017.

[6] Pranay Gupta, Anirudh Thatipelli, Aditya Aggarwal, Shubh Maheshwari, Neel Trivedi, Sourav Das, and Ravi Kiran Sarvadevabhatla. Quo vadis, skeleton action recognition ? *IJCV*, 2021.

[7] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE TPAMI*, 2014.

[8] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, June 2020.

[9] Tianjiao Li, Jun Liu, Wei Zhang, Yun Ni, Wenqian Wang, and Zhiheng Li. Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. In *CVPR*, 2021.

[10] J. Liu, A. Shahroudy, M. Perez, G. Wang, L. Y. Duan, and A. C. Kot. Ntu rgb+d 120: A large-scale benchmark for 3D human activity understanding. *IEEE TPAMI*, 2020.

[11] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *CVPR*, 2020.

[12] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3D human pose estimation in video with temporal convolutions and semi-supervised training. In *CVPR*, 2019.

[13] Grégory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. LCR-Net++: Multi-person 2D and 3D Pose Detection in Natural Images. *IEEE TPAMI*, 2019.

[14] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3D human activity analysis. *CVPR*, 2016.

[15] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *CVPR*, 2019.

[16] Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. Cross-view action modeling, learning and recognition. In *CVPR*, 2014.

[17] S. Yan, Yuanjun Xiong, and D. Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. *AAAI*, 2018.

[18] Di Yang, Rui Dai, Yaohui Wang, Rupayan Mallick, Luca Minciullo, Gianpiero Francesca, and Francois Bremond. Selective spatio-temporal aggregation based pose refinement system: Towards understanding human activities in real-world videos. *WACV*, 2021.

[19] W. Zhang, M. Zhu, and K. G. Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *ICCV*, 2013.