A Additional details on ChimpACT

A.1 Ethogram

We detail the ethogram definition in Tab. A1, which systematically describes the daily behaviors of chimpanzees.

Table A1: **The ethogram used for the ChimpACT dataset.**

category	definition	subcategory	subcategory definition
locomotion	patterns of self-initiated movement of an individual	0. moving	moving horizontally, *e.g.*, walking, running
		1. climbing	moving vertically, *e.g.*, climbing up or down a structure
		2. resting	remaining stationary, *e.g.*, standing, sitting, or lying
		3. sleeping	resting and keeping eyes closed
object interaction	direct physical interactions with inanimate stationary or movable objects by hands, feet or mouth	4. solitary object playing	non-social and non-goal-directed object interaction and exploration
		5. eating	consuming and processing food
		6. manipulating object	manipulation of any kind of inanimate object excluding eating
social interaction	at least two chimpanzees are interacting in differentiated roles: with one individual initiating the social behavior (initiator) and one individual receiving the social behavior (recipient)	7. grooming	a chimpanzee, the groomer, is cleaning the fur, head, hand, feet, or genitals of another chimpanzee, usually using their hands and/or mouth
		8. being groomed	one chimpanzee, the groomee, is getting their skin or fur cleaned by another chimpanzee
		9. aggressing	a chimpanzee is showing agonistic behavior towards another chimpanzee. This can range from charging and chasing another chimpanzee to direct physical contact such as slapping, hitting, and biting
		10. embracing	a chimpanzee is embracing another chimpanzee with their arms, not to be confused with carrying
		11. begging	a chimpanzee is requesting food or another object from another chimpanzee, oftentimes by extending their arm, reaching, or using an open palm begging gesture
		12. being begged from	a chimpanzee is requested food or another object by another chimpanzee
		13. taking object	taking an object from the possession of another chimpanzee, the transfer might be resisted or not
		14. losing object	the possession is taken by another chimpanzee
		15. carrying	a chimpanzee (usually an adult) carries another chimpanzee (usually an infant or juvenile) on the back, front, side, arm, or leg for more than 2 steps
		16. being carried	a chimpanzee (usually an infant or juvenile) is carried by another chimpanzee (usually an adult) on the back, front, side, arm, or leg for more than 2 steps.
		17. nursing	a female chimpanzee is nursing (breastfed, *i.e.*, making physical contact with the nipple) an infant/juvenile
		18. being nursed	an infant/juvenile is being nursed (breastfed, *i.e.*, making physical contact with the nipple) by a female chimpanzee
		19. playing	a chimpanzee is physically interacting with another individual in a friendly, teasing, or mock fighting way (*e.g.*, play fighting and other behaviors)
		20. touching	a chimpanzee makes body contact with another chimpanzee (*e.g.*, holding hands) and it does not fit with any of the other social interaction categories described above
others	other behaviors	21. erection	a male chimpanzee has an erect penis
		22. displaying	a male chimpanzee, usually with puffed up hair (piloerection) and an erection, performs a dominance display, which includes walking with a swagger, swinging their arms to the sides, and making calls with increasing amplitude, commonly ending by stomping against or slapping objects. Displays can be directed at another chimpanzee or be undirected

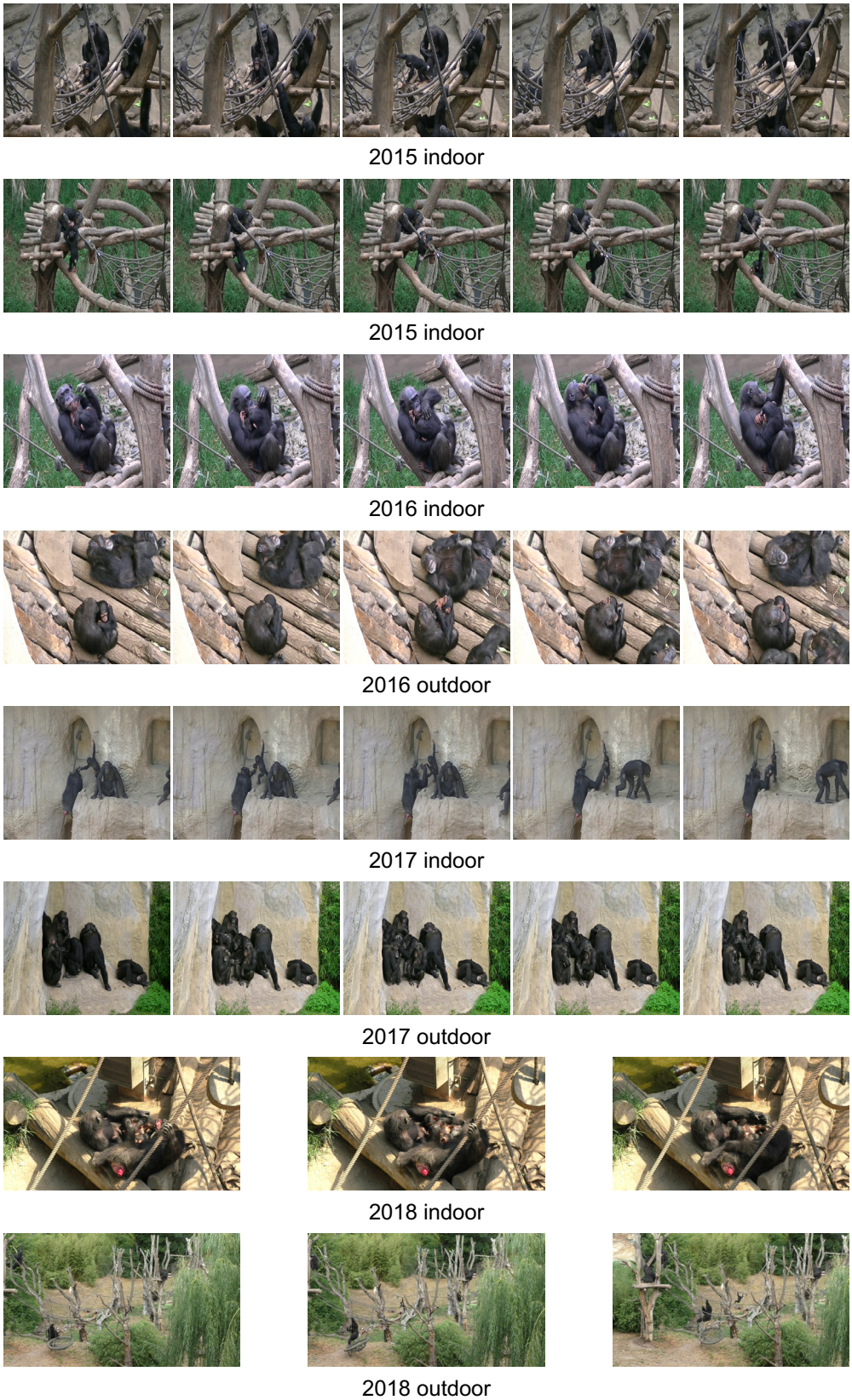


Figure A1: **Example frames from the `ChimpACT` dataset.** `ChimpACT` possesses rich social interactions of the complex everyday life of group-living chimpanzees and contains several environmental enrichment.

A.2 Dataset details

Collection and organization 405 hours of video footage of the Leipzig A-group chimpanzees were collected between 2015 and 2018. To create a representative sample of the footage, 163 video clips were selected, with 15, 35, 86, and 27 clips taken from each year. These video clips cover the four seasons. Each clip is 1000 frames long, with only 3 clips being shorter than 1000 frames. Visual examples from six clips, featuring both indoor and outdoor enclosures, are shown in Fig. A1. The dataset covers a diverse range of physical scenarios, camera views, and social behaviors, as demonstrated in these examples. For instance, in the third row of the figure, an adult chimpanzee is shown grooming an infant chimpanzee in her arms, while later on, the same infant is nursed.

Annotation process and quality The annotation process was conducted using BasicFinder CO., Ltd.'s private labeling platform, which involved a team of 15 annotators and 2 managers. Prior to commencing the annotation work, our team developed comprehensive guidelines that explicitly outlined the requirements for labeling. These guidelines covered several aspects, including:

(i) Assigning a bounding box for each chimpanzee in the image. (ii) Specifying the visibility of the bounding boxes. (iii) Assigning tracking IDs to each bounding box for tracking purposes. (iv) Localizing 2D keypoints within each bounding box. (v) Indicating the visibility of each 2D keypoint. (vi) Assigning behavior labels for each bounding box.

To ensure that the annotators followed these guidelines accurately, the project managers provided training based on the guidelines. Following the training, the annotators performed a trial annotation on a small dataset. We actively sought feedback from the annotators during this phase, which allowed us to address any issues and make necessary improvements. We conducted a thorough review of the trial annotations to verify that the quality met our standards.

During the trial labeling phase, we reached out to three labeling companies and ultimately selected BasicFinder CO., Ltd. based on their exceptional labeling quality. It is worth noting that BasicFinder CO., Ltd. has previously led the annotation efforts for the BDD100K (Yu et al., 2020) dataset, which is a substantial dataset used for autonomous driving purposes. This experience demonstrates their ability to maintain high annotation standards for complex and extensive datasets. Consequently, their involvement improves the reliability of our `ChimpACT` dataset annotations as well.

Once we were confident in the quality of the trial annotations, we proceeded with the large-scale annotation process. To manage the annotations efficiently, each video clip was designated as an annotation task, and our managers assigned these tasks to individual annotators using BasicFinder CO., Ltd.'s platform, ensuring that there was no overlap in assignments. BasicFinder CO., Ltd. has implemented rigorous quality management practices throughout the annotation process. These practices include a customized workflow, complete job traceability, precise performance tracking, multiple levels of auditing, and scientific personnel management. By adhering to these practices, we were able to maintain high standards of quality and accuracy while ensuring efficient processing speed. The annotation process followed a sequential workflow of execution, review, and quality control. Experienced annotators were responsible for executing the annotations, while the manager, as well as our team, conducted thorough reviews and quality control checks. Any annotations that did not meet the required standards were sent back to the annotators for corrections. The quality control phase involved a comprehensive review and verification of all data by both the managers and our own team, ensuring the integrity and accuracy of the annotations. Once all the data had been confirmed to meet our standards of quality, we concluded the annotation process.

More specifically, to label chimpanzee identities, annotators only needed to assign a tracking ID to each chimpanzee, which was then reviewed by the primatologist in our team, who assigned the apes' names based on his knowledge of the observed Leipzig A-group chimpanzees. The process of localizing 2D keypoints within each bounding box and assigning behavior labels for each chimpanzee presented bigger challenges than other tasks. To overcome these challenges, we implemented several measures to ensure accuracy and consistency. For the labeling of 2D keypoints, we provided detailed instructions accompanied by visual illustrations, aiming to provide clear guidelines for annotators to precisely identify and mark the keypoints. For labeling of behaviors, we supplied example videos showcasing different chimpanzee behaviors, created by our team's experienced primatologists. These videos served as valuable references, enabling annotators to accurately assign behavior labels based on observed actions. Throughout the annotation process, the primatologists actively participated, offering their expertise and providing valuable feedback to ensure the annotations aligned with

scientific standards. Finally, the behavioral primatologists in our team manually reviewed all labeled frames to ensure data reliability. These measures and the involvement of the primatologists were instrumental in enhancing the overall quality and reliability of the annotations.

For more information on the dataset, including pre-processing scripts, and visualized annotations, please refer to our project website.

B Discussion on ChimpACT

Intended uses The ChimpACT dataset is a versatile resource that can be used for studying algorithms for chimpanzee detection, tracking, identification, pose estimation, and spatiotemporal action detection. Therefore, the dataset is both relevant for questions in computer vision and primate behavior. In the context of computer vision, it lends itself to other research topics, including but not limited to pose tracking, few-shot learning, weakly-supervised learning, and transfer learning. Considering primate behavior, the dataset shares numerous features with other video data commonly collected with captive and wild chimpanzee populations. This makes it an ideal resource for fine-grained investigations of social (*e.g.*, grooming, nursing, aggression) and nonsocial (*e.g.*, locomotion, object interactions) chimpanzee behaviors. We strongly encourage researchers to utilize our dataset solely for research purposes that promote animal welfare and conservation. We firmly discourage any use of the dataset for harmful activities such as poaching, hunting or any other exploitation of primates. It is crucial for researchers to approach the data with a focus on positive societal impacts and to refrain from any potential negative consequences.

Ethics The ChimpACT dataset raises no ethical concerns regarding the privacy information of human subjects, as it solely focuses on chimpanzees. Studying the social behavior of chimpanzees provides an ethical and efficient means to explore aspects of human sociality due to our phylogenetic proximity. By analyzing their behaviors, we can gain insights into the evolution of human social behavior and potentially contribute to both the scientific and ethical understanding of the human condition. The ethics committee of the Wolfgang Köhler Primate Research Center approved the observational data collection for this project.

Maintenance, distribution, and license The ChimpACT dataset will be maintained by the authors and made publicly available with a total of 160,500 frames (around 2 hours) on our project website. The ChimpACT dataset will be distributed under the CC BY-NC 4.0 license.

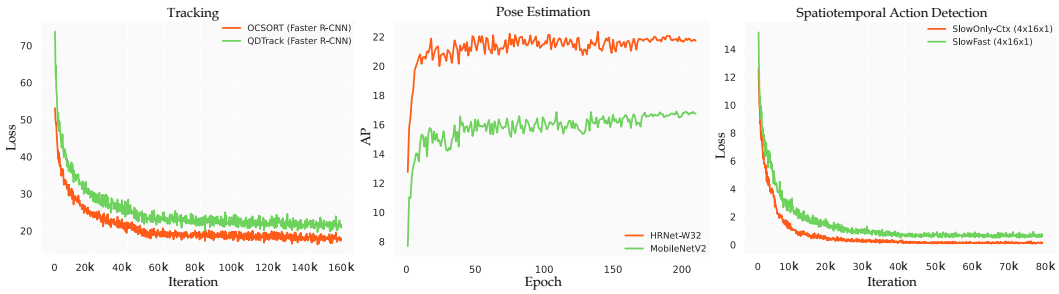
Wage paid to annotators We collaborated with BasicFinder CO., Ltd. for the annotation process. The labeling was carried out by 15 annotators, and they were offered a fair wage as per the prearranged contract. The total expenditure for the labeling process was approximately 70,000 RMB.

C Experiments

We trained all the models with officially-used training configurations for each of the three tracks. Please refer to the code implementation on our Github for details. Although we trained the models for different epochs in experiments conducted on different tracks, these choices were made based on conventional practices. Based on the training loss curves provided in Figs. A2a and A2c, it can be observed that all tracking and spatiotemporal action detection methods have reached convergence within the chosen training epochs. To assess the potential overfitting of the pose estimation models, we have included the validation curve on the AP metric in Fig. A2b. The validation curve demonstrates the performance of the pose models on the validation set, which indicates that the pose estimation models are not exhibiting signs of overfitting. Therefore, based on the training loss curves and the validation curve, it can be concluded that the chosen training epochs are appropriate for both tracking and pose estimation methods.

C.1 Detection, tracking, and ReID

We partitioned the dataset of 163 videos into three sets: 127 videos for training, 17 for validation, and 19 for testing. Of note, all individual chimpanzees are present in both the training and testing sets. In the test set, there are 12 and 7 videos for indoor and outdoor scenes, respectively.



(a) (b) (c)

Figure A2: **Training or validation curves on three tracks of example methods.** (a) Training loss curve of example tracking methods. The training iterations correspond to 10 epochs. (b) Validation curve on the AP metric of example pose estimation methods. (c) Training loss curve of example spatiotemporal action detection methods. The training iterations correspond to 20 epochs.

For the evaluation metrics, MOTA (Multiple Object Tracking Accuracy) takes into account FP (False Positives), FN (False Negatives), and IDs (IDentity switches). Usually, FP and FN are larger than IDs; therefore, MOTA mainly assesses the detection performance. IDF1 evaluates the ability to preserve subject identities to assess identification association performance. HOTA (Higher Order Tracking Accuracy) is a recently proposed metric that considers accurate detection, association, and localization equally important, and balances their effects explicitly.

Results We additionally evaluated the performance on the **indoor** and **outdoor** test set in Tabs. A2 and A3, respectively. Notably, the results indicate that these approaches achieve consistently better performance on the indoor test set compared to the outdoor test set. This may be attributed to the greater complexity of outdoor scenarios and the presence of varying camera views, which can significantly increase the difficulty of detecting and tracking chimpanzees. Furthermore, the presence of occlusions, similar appearances, and other environmental factors can further exacerbate the challenges of chimpanzee tracking in outdoor settings.

Table A2: **Results of the detection, tracking, and ReID track on `ChimpACT` *indoor* test set.**

Method	Detector	ReID	HOTA ↑	MOTA ↑	MOTP ↑	IDF1 ↑	mAP ↑	FP ↓	FN ↓	IDs ↓
SORT (Bewley et al., 2016)	Faster R-CNN	ResNet-50	49.1	52.7	21.0	49.2	76.2	8275	9396	731
	YOLOX		41.3	46.7	18.9	38.4	77.2	6440	13163	1105
DeepSORT (Wojke et al., 2017)	Faster R-CNN	ResNet-50	53.2	51.6	21.0	58.3	76.2	8277	9398	1144
	YOLOX		43.3	46.8	18.9	40.8	77.2	6440	13163	1092
Tracktor (Bergmann et al., 2019)	Faster R-CNN	ResNet-50	53.6	54.5	20.6	58.3	76.2	6575	10966	146
QDTrack (Pang et al., 2021)	Faster R-CNN	–	53.6	53.6	20.9	58.5	76.7	8121	9591	332
ByteTrack (Zhang et al., 2022)	Faster R-CNN	–	48.8	38.9	22.1	52.3	72.7	11599	11799	372
	YOLOX	–	51.0	48.0	17.7	55.6	76.2	5080	14893	245
OC-SORT (Cao et al., 2023)	Faster R-CNN	–	48.6	40.5	21.6	52.5	71.8	10022	12693	431
	YOLOX	–	49.8	47.9	19.3	53.6	75.9	7550	12292	422

Table A3: **Results of the detection, tracking, and ReID track on `ChimpACT` *outdoor* test set.**

Method	Detector	ReID	HOTA ↑	MOTA ↑	MOTP ↑	IDF1 ↑	mAP ↑	FP ↓	FN ↓	IDs ↓
SORT (Bewley et al., 2016)	Faster R-CNN	ResNet-50	31.3	43.1	25.2	35.0	63.3	3288	8142	422
	YOLOX		34.8	31.9	22.9	35.0	61.5	3649	9786	751
DeepSORT (Wojke et al., 2017)	Faster R-CNN	ResNet-50	39.4	41.7	25.2	47.8	63.3	3280	8134	726
	YOLOX		36.6	31.8	22.9	37.0	61.5	3649	9786	788
Tracktor (Bergmann et al., 2019)	Faster R-CNN	ResNet-50	38.8	42.5	24.5	45.0	63.3	2734	9146	94
QDTrack (Pang et al., 2021)	Faster R-CNN	–	40.0	50.5	27.1	49.6	73.3	3705	6067	534
ByteTrack (Zhang et al., 2022)	Faster R-CNN	–	32.5	32.7	30.1	42.9	62.1	5346	8375	312
	YOLOX	–	44.2	37.5	23.1	51.5	60.4	1842	11042	139
OC-SORT (Cao et al., 2023)	Faster R-CNN	–	28.3	27.4	29.8	39.6	60.5	5342	9341	448
	YOLOX	–	42.7	31.6	22.6	47.8	60.5	4298	9695	252

We visualize the tracking results in Figs. A3 and A4, with the ground-truth bounding boxes and chimpanzee identities shown in the last row. We visualized the confidence scores of the estimated bounding boxes and their associated IDs in each frame obtained by the evaluated methods. It is worth noting that we do not require individual identification of each chimpanzee, but rather assign the same ID to the same animal across frames, following the common practice in multi-human tracking (Milan

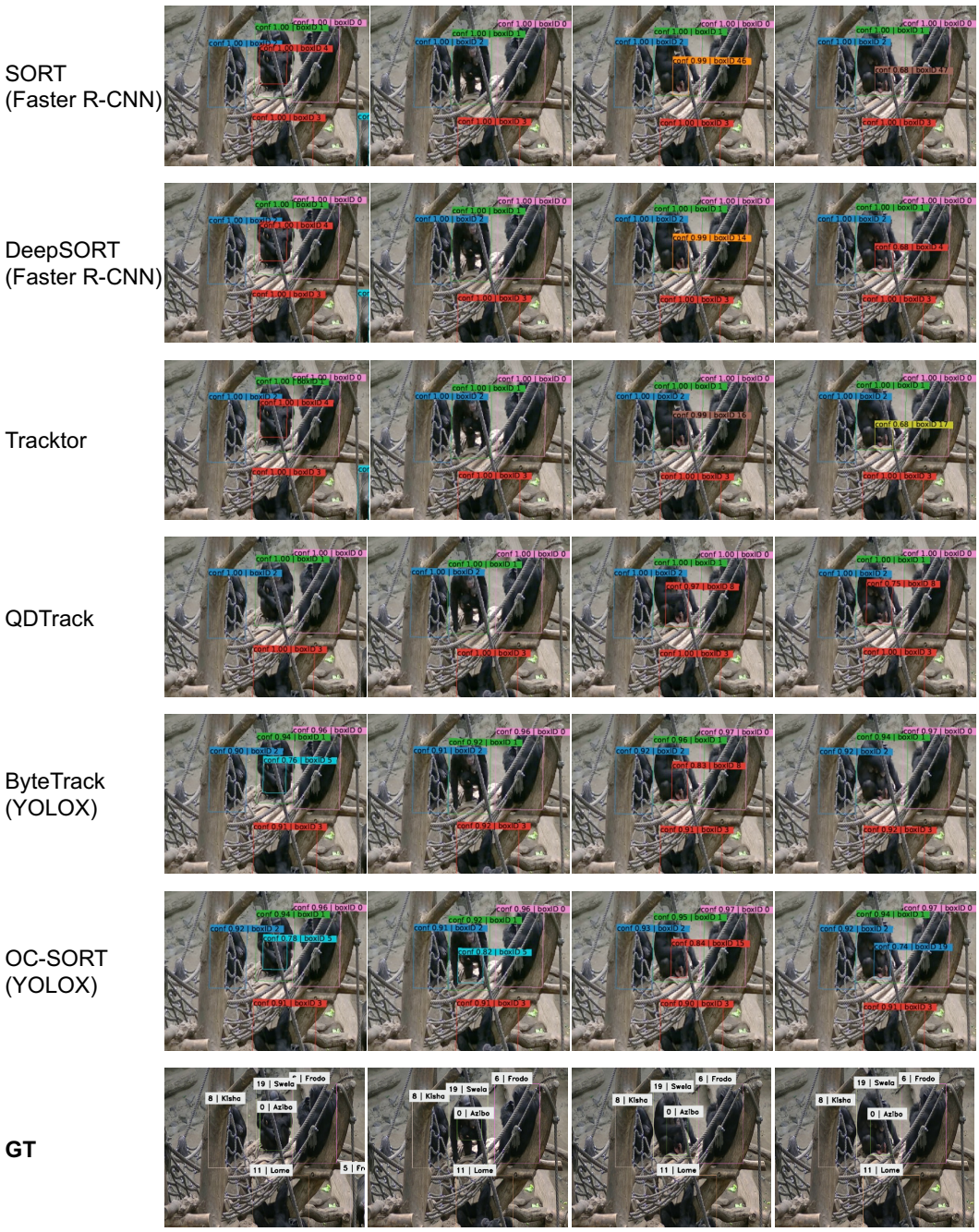


Figure A3: **Qualitative results of representative methods on the `ChimpACT` test set on the tracking task.** For each method, we visualize the estimated confidence score ("conf") and the associated IDs ("boxID") of each bounding box in each frame. The ground-truth bounding boxes and chimpanzee names are shown in the last row, and we add a number left to the name to make it easier to track. Please zoom in for details.

et al., 2016). The estimated box ID is therefore used solely for evaluating the tracking performance. We observed that the evaluated methods performed well in scenarios with minimal occlusion, but struggled to detect and associate the same individual chimpanzee when heavy occlusion occurred. For instance, in Fig. A3, the infant chimpanzee's bounding box is lost in some frames, and its identity is erroneously switched later due to heavy occlusion. This is a challenging task in chimpanzee detection and tracking, as occlusions frequently occur in group-living habitats. Please refer to the supplementary video for more experimental results. In conclusion, the experimental results reveal the limitations of existing methods for chimpanzee detection and tracking, underscoring the need for more robust algorithms to be developed. We believe that our dataset can make a valuable contribution

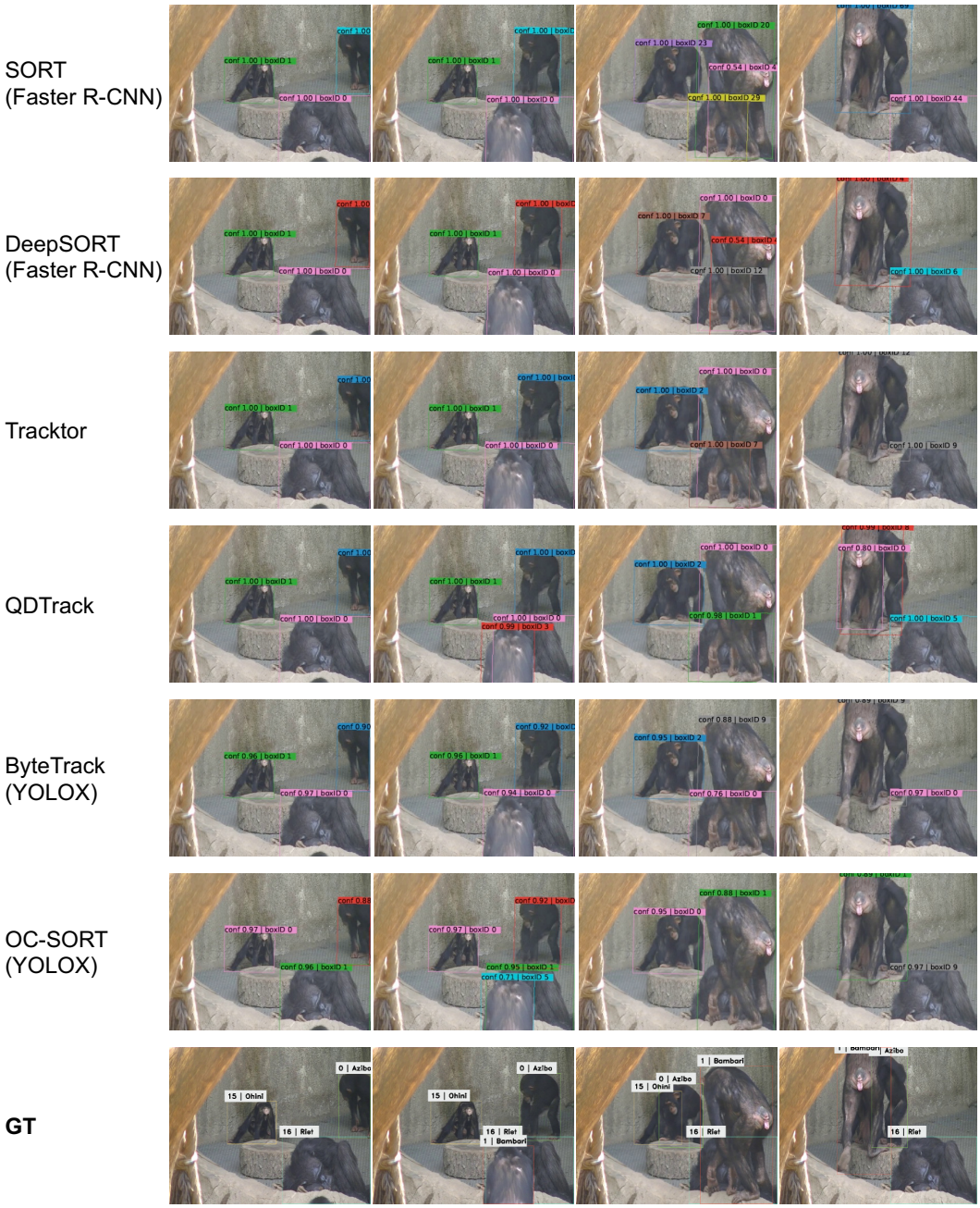


Figure A4: **More qualitative results of representative methods on the `ChimpACT` test set on the tracking task.** For each method, we visualize the estimated confidence score ("conf") and the associated IDs ("boxID") of each bounding box in each frame. The ground-truth bounding boxes and chimpanzee names are shown in the last row, and we add a number left to the name to make it easier to track. Please zoom in for details.

to the advancement of this field, by providing a challenging benchmark for evaluating and comparing different methods.

C.2 Pose estimation

We followed the partition of the dataset as the first track to train and evaluate the methods.

Results We report the PCK@0.1 for the 16 keypoints in Tab. A4. The results reveal that the keypoints on the face, such as the eyes and lips, exhibited better estimation compared to the arms and legs. This could be attributed to the fact that eyes and lips have more distinctive visual patterns than

Table A4: **Results of the pose estimation track for each keypoint on `ChimpACT` test set.** We report PCK@0.1 metric. We abbreviate the keypoint names. Please refer to Sec. 3.3 for the keypoint definition.

	Method	Backbone	0.hip	1.rknee	2.rankle	3.lknee	4.lankle	5.neck	6.ulip	7.llip	8.reye	9.leye	10.rshoul	11.relbow	12.rwrist	13.lshoul	14.lelbow	15.lwrist
Regression	SimpleBaseline (Xiao et al., 2018)	ResNet-50	51.1	45.8	52.3	44.7	48.8	56.4	76.2	77.9	85.7	85.2	54.7	46.1	29.2	60.5	48.5	31.5
		ResNet-101	51.3	49.0	53.3	47.3	50.2	58.2	77.1	78.7	86.4	86.4	57.9	46.8	32.5	60.2	51.8	35.4
		ResNet-152	50.6	50.5	56.8	47.4	45.3	58.3	76.4	77.4	86.8	86.0	55.8	45.1	35.2	58.2	51.3	35.8
	RLE (Li et al., 2021)	MobileNetV2	53.1	46.9	53.8	49.0	48.7	61.4	77.1	78.7	86.3	85.1	59.2	41.6	33.5	59.0	48.2	31.9
		ResNet-50	47.7	42.6	46.6	42.7	46.2	57.7	75.9	77.4	81.4	79.3	59.0	44.0	30.3	58.9	48.5	30.5
		ResNet-101	51.9	49.4	52.8	55.4	49.1	61.0	79.5	80.4	87.2	86.6	60.3	46.4	40.2	62.0	53.6	39.0
		ResNet-152	54.1	50.2	52.8	53.1	49.3	60.6	79.8	80.7	88.0	85.4	63.1	50.7	42.5	61.1	53.5	38.6
Heatmap-based	CPM (Wei et al., 2016)	CPM	61.1	65.9	71.7	59.7	68.7	67.3	85.5	87.2	91.1	90.5	67.0	60.1	59.6	67.8	66.3	53.4
	Hourglass (Newell et al., 2016)	Hourglass-4	62.4	65.3	70.8	65.2	67.8	66.4	84.0	85.9	86.9	87.3	68.5	61.7	60.4	67.7	66.0	56.0
	MobileNetV2 (Sandler et al., 2018)	MobileNetV2	58.8	64.8	71.2	61.3	64.8	67.0	83.8	85.4	91.0	89.1	69.3	58.6	56.9	67.4	64.8	52.1
	SimpleBaseline (Xiao et al., 2018)	ResNet-50	63.2	67.9	70.7	64.4	67.5	66.8	85.1	86.3	92.8	90.5	70.6	59.1	57.7	67.6	65.0	54.4
		ResNet-101	62.0	64.6	69.6	61.4	68.4	67.4	85.1	87.4	91.9	89.6	70.1	61.2	56.3	66.7	63.5	54.1
		ResNet-152	64.5	64.6	69.2	62.5	69.9	67.3	86.5	88.5	91.1	89.7	72.4	62.4	58.5	69.9	66.0	55.3
	HRNet (Sun et al., 2019)	HRNet-W32	65.8	69.5	74.5	66.1	69.2	70.5	88.2	90.4	92.6	92.1	76.1	67.7	64.4	72.4	69.5	62.8
		HRNet-W48	61.5	69.1	74.6	65.1	70.4	70.7	87.5	88.9	93.8	92.2	75.3	64.7	61.1	72.1	70.6	58.9
	DarkPose (Zhang et al., 2020)	ResNet-50	62.6	64.4	68.6	63.2	66.6	69.9	86.3	87.7	91.7	90.5	73.8	61.5	59.1	69.6	66.9	58.0
		ResNet-101	61.7	62.9	70.5	62.6	65.7	67.0	86.3	87.7	92.0	89.7	70.1	59.7	55.4	68.6	62.8	54.4
		ResNet-152	63.3	68.6	69.1	62.5	66.0	67.7	86.5	88.0	92.6	89.5	71.8	61.9	56.1	69.5	63.3	53.4
		HRNet-W32	63.5	67.3	74.0	67.2	71.6	70.0	88.3	89.5	93.4	92.1	75.6	65.3	64.3	73.1	69.2	62.6
		HRNet-W48	65.9	69.7	73.5	67.1	72.8	72.0	89.6	91.3	94.5	91.8	73.3	62.6	61.2	71.6	70.8	62.8
	HRFormer (Yuan et al., 2021)	HRFormer-S	63.0	66.5	70.7	64.2	68.5	67.5	84.5	85.6	91.0	89.1	71.3	61.0	59.1	68.3	64.9	56.0
		HRFormer-B	61.4	67.2	71.9	66.3	70.9	67.7	84.9	86.2	93.6	90.6	71.9	66.3	62.3	70.8	67.2	58.0

limbs, which are often surrounded by heavy fur. Tab. A5 further reports the PCK@0.1 for each action category on the test set. We observe that different action types exhibit variations in pose accuracy, for example, with climbing generally achieving slightly higher accuracy compared to resting in most methods. This observation can be attributed to the higher potential for self-occlusion during resting, as chimpanzees tend to exhibit significant self-occlusion due to their flexible joints. This is evident in the visualized examples in Fig. A5, where (a) and (c) depict resting poses with pronounced self-occlusion. In contrast, during climbing, the body is mostly in an extended state, as shown in (b) and (d). Consequently, the PCK tends to be slightly higher for climbing compared to resting as shown in Tab. A6. To validate this assumption, we further evaluate the performance of all the methods for non-occluded poses in Tab. A7. It is interesting to note that all the methods achieve high PCK accuracy when all the keypoints are visible. This demonstrates their effectiveness in accurately estimating poses when occlusions are minimal or absent.

These observations highlight the unique and intricate nature of chimpanzee pose estimation, which is complicated by their flexible joint articulations and extended range of motion, as well as the dissimilar physical appearances of their fur in comparison to that of humans. Consequently, developing accurate pose estimation algorithms for chimpanzees requires careful consideration and specialized techniques that account for their unique characteristics.

Figs. A6 and A7 present the qualitative results of several models on the `ChimpACT` test split, with the ground-truth poses displayed in the last row. It is promising to observe that directly transferring human pose estimation algorithms to chimpanzees yielded decent performance. However, due to self-occlusions and different physical appearance and joint articulations, these models are susceptible to errors in estimating the positions of limbs, as seen in the misaligned right arm and leg of the young chimpanzee in the first column of Fig. A6 and the third column of Fig. A7.

Table A6: **Results of the pose estimation by HRNet-W32 model.** We report PCK@0.05 and PCK@0.1 metrics.

No.	Action	PCK@0.05	PCK@0.1
(a)	resting	43.8	62.5
(b)	climbing	81.2	93.8
(c)	resting	68.8	93.8
(d)	climbing	75.0	100.0

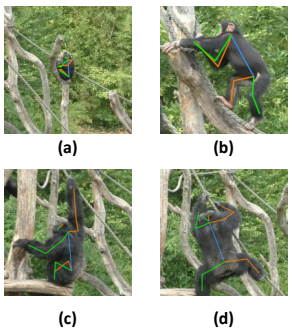


Figure A5: **Visualization of predicted pose by HRNet-W32 (Sun et al., 2019).**

Table A5: **Results of the pose estimation track for each action category on `ChimpACT` test set.** We report PCK@0.1 metric. The action category number is consistent with Tab. A1.

	Method	Backbone	0	1	2	3	4	5	6	7	8	9	10	11	12	15	16	17	18	19	20	21	22
Regression	SimpleBaseline (Xiao et al., 2018)	ResNet-50	39.8	47.7	48.0	56.1	44.4	64.2	56.7	35.7	26.1	81.3	67.0	51.3	45.0	26.5	34.5	3.7	5.5	29.9	26.9	87.5	56.6
		ResNet-101	39.3	49.0	48.1	59.5	46.3	60.9	57.5	38.9	20.7	75.0	69.5	57.5	45.0	32.1	35.5	2.9	6.6	28.8	26.6	62.5	52.5
		ResNet-152	40.8	45.6	49.9	56.2	46.5	63.9	54.8	37.7	23.6	68.8	67.4	62.5	51.3	30.6	34.3	6.6	5.3	29.5	26.1	75.0	56.6
	RLE (Li et al., 2021)	MobileNetV2	40.8	48.0	50.0	52.9	47.1	63.5	57.7	38.4	18.1	62.5	67.6	53.8	48.8	28.8	36.5	5.7	8.5	29.4	29.8	62.5	62.5
		ResNet-50	42.0	52.1	51.5	57.6	50.0	65.2	57.8	41.0	20.2	75.0	69.0	50.0	55.0	31.9	35.4	4.6	3.2	29.0	31.5	81.3	61.3
		ResNet-101	43.3	46.4	51.8	58.3	46.6	68.0	55.8	34.1	18.7	75.0	68.5	48.8	43.8	31.9	35.2	6.0	5.6	29.7	31.6	75.0	62.6
		ResNet-152	41.4	52.9	50.7	57.2	49.3	64.2	56.3	34.1	17.7	75.0	72.5	48.8	46.3	35.6	31.8	5.4	6.1	30.2	30.1	75.0	59.8
Heatmap-based	CPM (Wei et al., 2016)	CPM	49.4	59.4	59.0	60.9	53.9	73.1	65.2	46.6	28.4	81.3	66.6	52.5	60.0	41.6	34.8	10.6	3.8	36.1	41.8	68.8	66.2
	Hourglass (Newell et al., 2016)	Hourglass-4	48.1	66.5	55.3	63.2	58.5	71.9	67.4	50.3	27.8	81.3	72.8	47.5	65.0	44.0	38.2	14.1	1.8	40.6	35.3	81.3	63.6
	MobileNetV2 (Sandler et al., 2018)	MobileNetV2	49.8	58.4	56.1	59.3	54.8	71.2	65.1	52.8	25.8	75.0	72.8	60.0	48.8	39.8	35.8	11.7	1.5	35.0	36.2	81.3	61.4
	SimpleBaseline (Xiao et al., 2018)	ResNet-50	52.3	60.0	57.2	60.9	56.3	73.9	66.0	53.3	25.2	81.3	72.0	62.5	65.0	45.0	35.4	8.4	2.4	39.1	37.6	75.0	65.7
		ResNet-101	52.0	60.9	57.5	60.8	56.6	71.9	66.4	52.6	28.2	93.8	72.2	71.3	61.3	42.0	34.2	6.4	1.9	39.6	36.9	68.8	67.0
		ResNet-152	51.4	60.0	57.8	60.0	57.4	71.6	66.3	55.4	27.8	81.3	79.1	58.8	52.5	45.1	32.3	5.3	0.5	38.1	38.0	87.5	67.6
	HRNet (Sun et al., 2019)	HRNet-W32	56.7	66.1	60.8	60.9	60.2	76.3	69.3	54.8	27.2	87.5	74.8	61.3	63.8	50.6	38.7	9.1	2.4	40.2	41.4	81.3	65.6
		HRNet-W48	56.9	65.9	59.3	60.9	60.3	75.7	70.3	53.2	30.2	87.5	74.2	66.3	67.5	52.9	37.6	13.9	2.9	41.3	39.7	87.5	65.9
	DarkPose (Zhang et al., 2020)	ResNet-50	52.1	60.9	57.5	62.1	57.4	72.6	66.1	56.0	26.8	75.0	72.9	56.3	61.3	42.1	31.7	9.6	0.4	35.3	39.0	75.0	66.4
		ResNet-101	52.6	62.6	57.6	61.4	56.1	71.8	67.7	51.7	26.3	81.3	73.4	65.0	62.5	44.0	38.2	6.3	2.6	36.7	35.7	87.5	61.1
		ResNet-152	52.6	63.3	57.8	59.4	57.9	73.2	67.9	53.0	25.7	81.3	76.3	57.5	65.0	45.0	35.0	8.7	1.7	35.9	37.1	87.5	68.3
		HRNet-W32	56.9	68.9	62.6	62.5	61.5	74.0	69.7	56.5	26.0	81.3	81.2	58.8	72.5	52.2	42.3	9.8	2.1	41.6	44.5	81.3	67.3
		HRNet-W48	57.6	67.7	60.3	59.2	60.3	73.7	69.6	56.3	28.5	93.8	77.2	53.8	67.5	52.8	36.7	4.2	1.4	39.7	40.4	75.0	63.9
	HRFormer (Yuan et al., 2021)	HRFormer-S	52.9	62.3	55.7	59.6	56.8	72.3	68.2	54.2	23.7	93.8	75.1	68.8	52.5	45.1	33.5	2.5	1.1	40.6	34.5	62.5	66.9
		HRFormer-B	54.2	63.4	58.0	61.3	58.8	72.4	68.2	52.4	25.4	81.3	77.5	55.0	67.5	46.8	37.5	12.6	0.7	40.8	37.7	75.0	63.7

Table A7: **Results of the pose estimation track for non-occluded poses on `ChimpACT` test set.** We report the PCK metrics. The non-occluded poses denote those with all keypoints visible.

	Method	Backbone	PCK@0.05	PCK@0.1
Regression	SimpleBaseline (Xiao et al., 2018)	ResNet-50	47.6	80.6
		ResNet-101	47.2	77.9
		ResNet-152	54.5	83.0
	RLE (Li et al., 2021)	MobileNetV2	47.7	82.4
		ResNet-50	52.9	82.4
		ResNet-101	28.4	55.1
		ResNet-152	60.0	85.5
Heatmap-based	CPM (Wei et al., 2016)	CPM	74.0	89.4
	Hourglass (Newell et al., 2016)	Hourglass-4	77.6	88.5
	MobileNetV2 (Sandler et al., 2018)	MobileNetV2	67.4	89.0
	SimpleBaseline (Xiao et al., 2018)	ResNet-50	75.2	89.5
		ResNet-101	68.7	84.0
		ResNet-152	71.4	87.1
	HRNet (Sun et al., 2019)	HRNet-W32	77.6	92.1
		HRNet-W48	79.4	90.2
	DarkPose (Zhang et al., 2020)	ResNet-50	74.6	87.1
		ResNet-101	74.6	88.7
		ResNet-152	73.0	89.1
		HRNet-W32	80.7	93.0
		HRNet-W48	78.4	87.1
	HRFormer (Yuan et al., 2021)	HRFormer-S	70.9	88.5
		HRFormer-B	75.6	88.4

C.3 Spatiotemporal action detection

We adopted the same dataset partition as the first track. The frame sampling strategy was defined as $T \times I \times N$. We ablated two strategies that continuously sample one frame every I frames and finally get an input clip with T frames by setting $T \neq 1$. N denotes the number of clips which is used only when $T = 1$. For the four representative methods, we ablated different modules. For LFB (Wu et al., 2019), we ablated different ways of the feature bank operator instantiations, by using non-local (NL) blocks (Wang et al., 2018) or average (Avg) or max (Max) pooling. For SlowFast (Feichtenhofer et al., 2019) and the variant SlowOnly, we ablated the context module (Ctx), which indicates that using both the RoI feature and the global pooled feature for the action classification.

Results We report the mAP for each model's best configuration on several subcategory behaviors in Tab. A8. The models exhibit better performance in detecting locomotion and solitary object interactions, possibly because these actions are relatively simple and involve less interaction between individuals, making it easier for the model to distinguish between action patterns. However, there is still considerable room for improvement in existing models for action categories with higher levels of interaction, such as social interactions.

We provide qualitative results in Figs. A8 and A9. All methods recognized the playing action of the two chimpanzees in Fig. A8, but incorrectly classified the touching actions as grooming in Fig. A9. These two action patterns exhibit subtle differences that significantly challenge the models

Figure A6: **Qualitative results of representative methods on the `ChimpACT` test set on the pose estimation task.** The ground-truth poses are shown in the last row.

to distinguish them accurately. We recommend referring to the supplementary video for the video results to observe the difference. The challenges of such distinctions highlight the need for stronger algorithms to address these issues effectively.

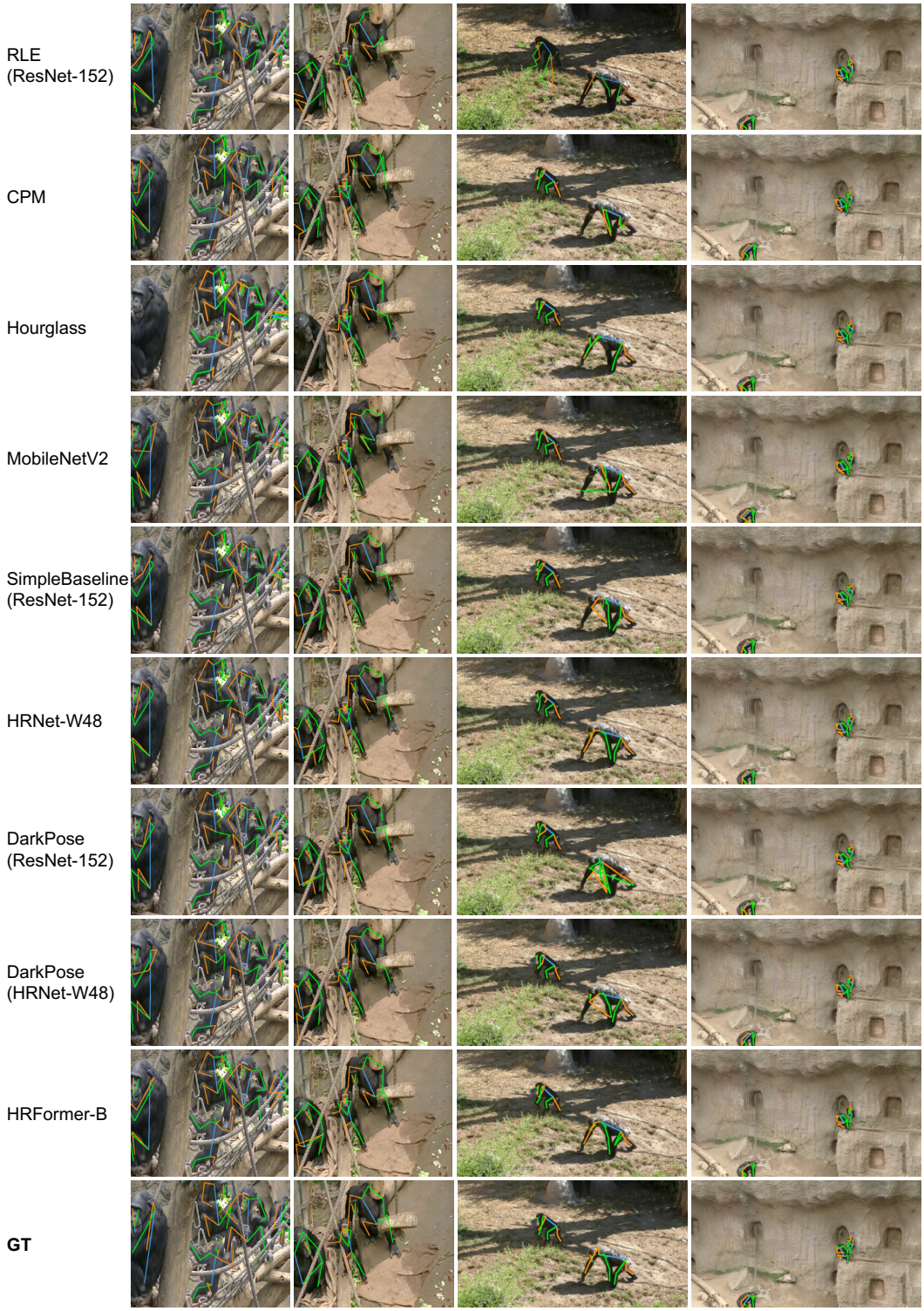


Figure A7: **More qualitative results of representative methods on the `ChimpACT` test set on the pose estimation task.** The ground-truth poses are shown in the last row.

Overall, we hope that our work will inspire further research and development in the area of chimpanzee behavior recognition, with the ultimate goal of improving our understanding of chimpanzee and primate behaviors and ecology.

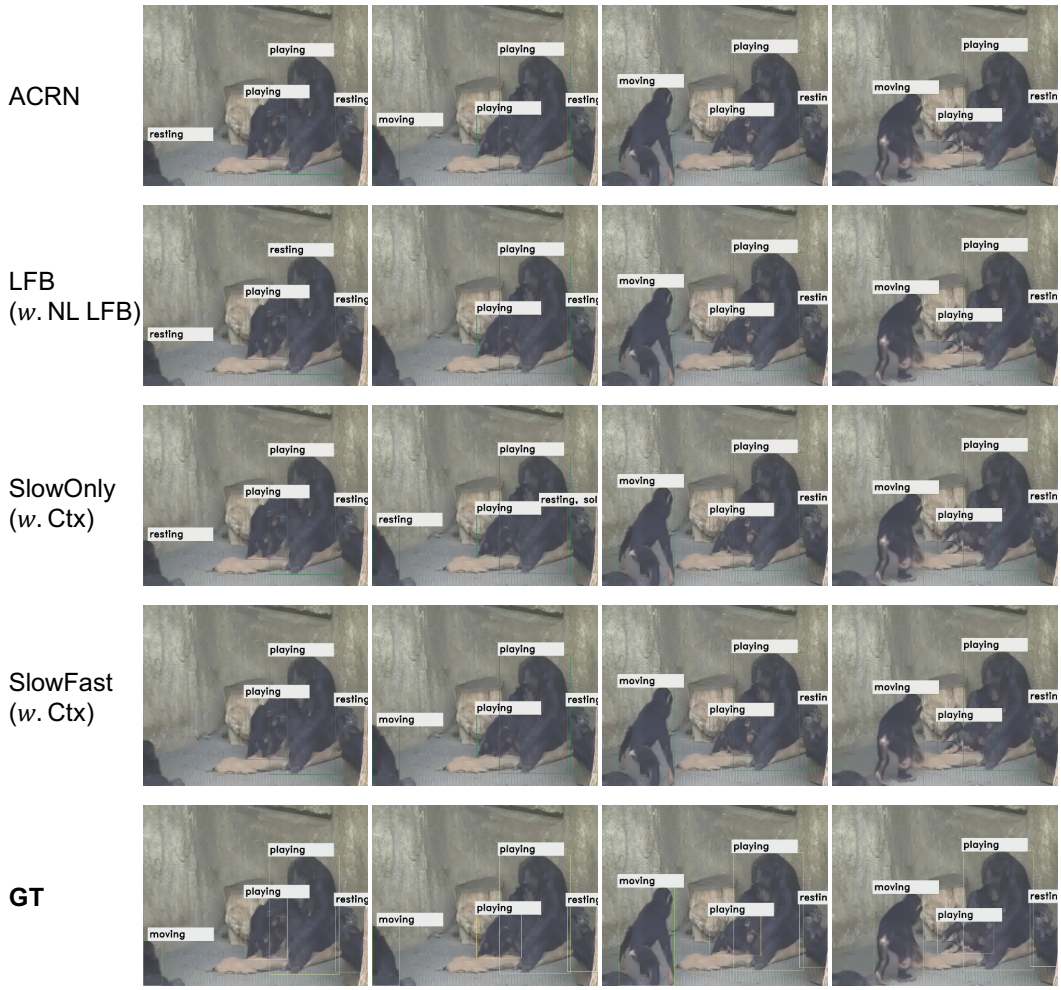


Figure A8: **Qualitative results of representative methods on the `ChimpACT` test set on the spatiotemporal action detection task.** The ground-truth actions are shown in the last row.

Table A8: **Results of spatiotemporal action detection track on `ChimpACT` test set.**

Method	mAP	moving	climbing	sol. obj. playing	eating	grooming	playing	being begged from	aggressing	being nursed
ACRN (Sun et al., 2018)	24.4	60.2	23.2	38.2	54.3	7.7	42.9	0.0	0.0	4.4
LFB (Wu et al., 2019)	22.4	45.3	10.0	34.4	56.3	8.7	51.0	0.4	0.0	32.1
SlowOnly (Feichtenhofer et al., 2019)	24.5	56.1	31.6	41.0	45.4	10.4	43.0	0.0	0.0	7.5
SlowFast (Feichtenhofer et al., 2019)	24.5	60.9	37.2	47.3	35.3	10.4	49.2	0.0	0.0	7.5

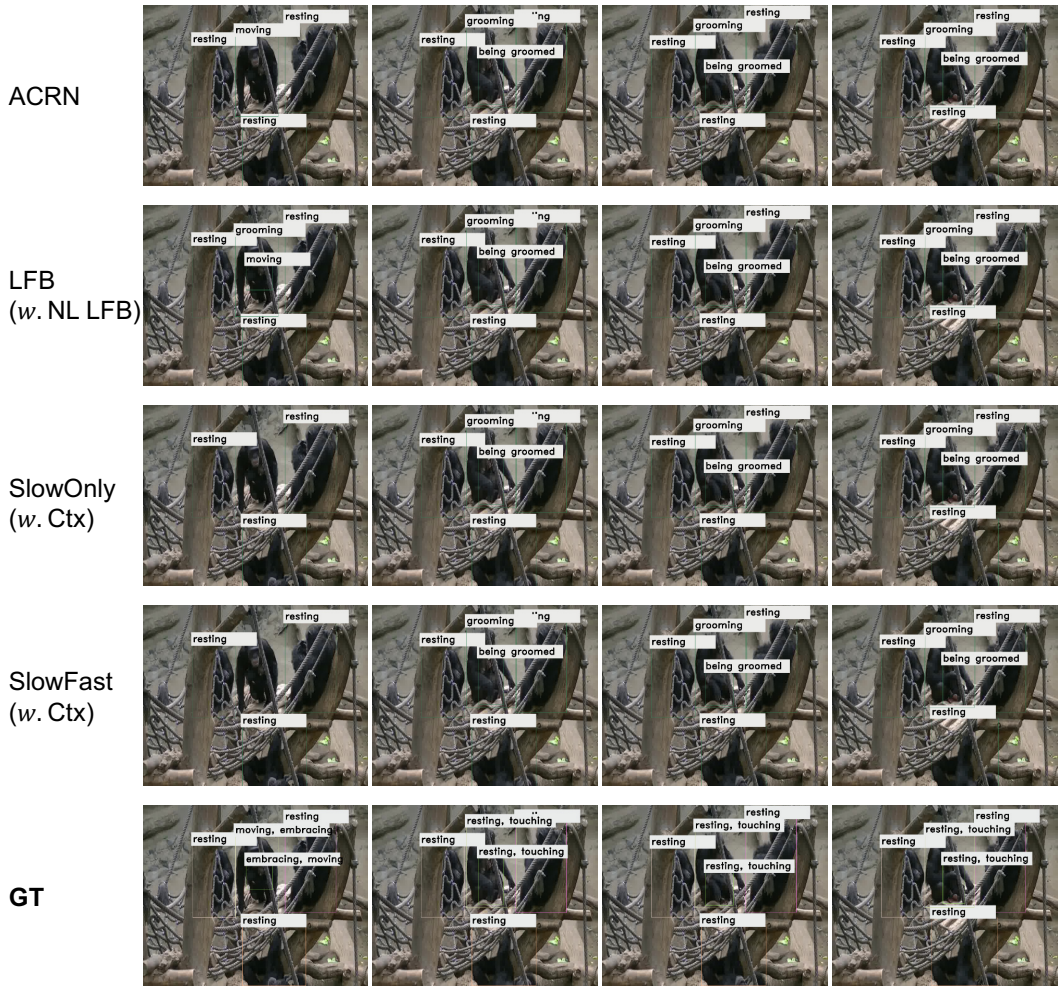


Figure A9: **More qualitative results of representative methods on the `ChimpACT` test set on the spatiotemporal action detection task.** The ground-truth actions are shown in the last row.

D Data documentation

We follow the datasheet proposed in Gebru et al. (2021) for documenting our ChimpACT and associated benchmarks:

1. **Motivation**
   (a) **For what purpose was the dataset created?**
   This dataset was created to facilitate the study of chimpanzee behaviors, and ultimately advance our understanding of communication and sociality in non-human primates.
   (b) **Who created the dataset and on behalf of which entity?**
   This dataset was created by Xiaoxuan Ma, Stephan P. Kaufhold, Jiajun Su, Wentao Zhu, Jack Terwilliger, Andres Meza, Yixin Zhu, Federico Rossano, and Yizhou Wang. Xiaoxuan Ma, Jiajun Su, Wentao Zhu, Yixin Zhu, and Yizhou Wang are with Peking University. Stephan P. Kaufhold, Jack Terwilliger, Andres Meza, and Federico Rossano are with the University of California, San Diego.
   (c) **Who funded the creation of the dataset?**
   The creation of this dataset was funded by Peking University and the University of California, San Diego.
   (d) **Any other Comments?**
   None.
2. **Composition**
   (a) **What do the instances that comprise the dataset represent?**
   For video data, each instance is a video clip regularized from the raw video. Each instance contains video footage focusing on a group of chimpanzees collected in Leipzig Zoo, Germany. For benchmarking, each instance has rich annotations of chimpanzee identities, poses, and actions. See Sec. 3 and Appx. A.
   (b) **How many instances are there in total?**
   We have 163 video instances in total.
   (c) **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?**
   No, this is a brand-new dataset.
   (d) **What data does each instance consist of?**
   See Sec. 3 and Appx. A.
   (e) **Is there a label or target associated with each instance?**
   Yes. See Sec. 3 and Appx. A.
   (f) **Is any information missing from individual instances?**
   No.
   (g) **Are relationships between individual instances made explicit?**
   Yes.
   (h) **Are there recommended data splits?**
   Yes, we have separated the whole dataset into training, validation, and test set. See Sec. 4.1, Appx. C and the project website for details.
   (i) **Are there any errors, sources of noise, or redundancies in the dataset?**
   There are almost certainly some errors in video annotations. We did our best to minimize these, but some certainly remain.
   (j) **Is the dataset self-contained, or does it link to or otherwise rely on external resources (*e.g.*, websites, tweets, other datasets)?**
   The dataset is self-contained.
   (k) **Does the dataset contain data that might be considered confidential (*e.g.*, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?**
   No.
   (l) **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?**
   No.

(m) **Does the dataset relate to people?**
No.

(n) **Does the dataset identify any subpopulations (*e.g.*, by age, gender)?**
No.

(o) **Is it possible to identify individuals (*i.e.*, one or more natural persons), either directly or indirectly (*i.e.*, in combination with other data) from the dataset?**
Not applicable. Our dataset only contains chimpanzees.

(p) **Does the dataset contain data that might be considered sensitive in any way (*e.g.*, data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?**
No.

(q) **Any other comments?**
None.

3. **Collection Process**

(a) **How was the data associated with each instance acquired?**
See Sec. 3.2 and Appx. A for details.

(b) **What mechanisms or procedures were used to collect the data (*e.g.*, hardware apparatus or sensor, manual human curation, software program, software API)?**
We used JVC Everio cameras to collect video footage (Codec H.264). See Sec. 3.2 for details.

(c) **If the dataset is a sample from a larger set, what was the sampling strategy (*e.g.*, deterministic, probabilistic with specific sampling probabilities)?**
See Sec. 3.3 and Appx. A for details.

(d) **Who was involved in the data collection process (*e.g.*, students, crowdworkers, contractors) and how were they compensated (*e.g.*, how much were crowdworkers paid)?**
The video data was collected by the authors. The annotations were performed by the workers in BasicFinder CO., Ltd., and the workers were offered a fair wage as per the prearranged contract. See Sec. 3 and Appx. B for details.

(e) **Over what timeframe was the data collected?**
The data were collected from 2015 to 2018, and labeled in 2022.

(f) **Were any ethical review processes conducted (*e.g.*, by an institutional review board)?**
Not applicable. The `ChimpACT` dataset raises no ethical concerns regarding the privacy information of human subjects, as it solely focuses on chimpanzees.

(g) **Does the dataset relate to people?**
No.

(h) **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (*e.g.*, websites)?**
Not applicable.

(i) **Were the individuals in question notified about the data collection?**
Not applicable.

(j) **Did the individuals in question consent to the collection and use of their data?**
Not applicable.

(k) **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?**
Not applicable.

(l) **Has an analysis of the potential impact of the dataset and its use on data subjects (*e.g.*, a data protection impact analysis) been conducted?**
Yes, see Appx. B.

(m) **Any other comments?**
None.

4. **Preprocessing, Cleaning and Labeling**

(a) **Was any preprocessing/cleaning/labeling of the data done (*e.g.*, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?**
Yes, see Sec. 3.

(b) **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (*e.g.*, to support unanticipated future uses)?**
Yes, we provide the raw data on our project website.

(c) **Is the software used to preprocess/clean/label the instances available?**
No. The annotation software is the private labeling platform provided by BasicFinder CO., Ltd. However, existing open-source annotation software such as DeepLabCut (Mathis et al., 2018) could also be used to preprocess/clean/label the instances.

(d) **Any other comments?**
None.

5. **Uses**

(a) **Has the dataset been used for any tasks already?**
No, the dataset is newly proposed by us.

(b) **Is there a repository that links to any or all papers or systems that use the dataset?**
Yes, we provide the link to all related information on our project website.

(c) **What (other) tasks could the dataset be used for?**
This dataset could be used for other research topics, including but not limited to pose tracking, few-shot learning, and transfer learning.

(d) **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?**
We propose to annotate the keyframe every 10 frames for the pose track and action detection track. For tracking track, we label all the frames.

(e) **Are there tasks for which the dataset should not be used?**
The usage of this dataset should be limited to the scope of understanding chimpanzee/non-human primate behaviors.

(f) **Any other comments?**
None.

6. **Distribution**

(a) **Will the dataset be distributed to third parties outside of the entity (*e.g.*, company, institution, organization) on behalf of which the dataset was created?**
Yes, the dataset will be made publicly available.

(b) **How will the dataset be distributed (*e.g.*, tarball on website, API, GitHub)?**
The dataset could be accessed on our project website.

(c) **When will the dataset be distributed?**
The dataset will be released by the end of 2023 on our project website.

(d) **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?**
We release our benchmark under CC BY-NC 4.0 [1] license.

(e) **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?**
No.

(f) **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?**
No.

(g) **Any other comments?**
None.

7. **Maintenance**

(a) **Who is supporting/hosting/maintaining the dataset?**
Xiaoxuan Ma is maintaining.

[1] https://paperswithcode.com/datasets/license

(b) **How can the owner/curator/manager of the dataset be contacted (*e.g.*, email address)?**
maxiaoxuan@pku.edu.cn

(c) **Is there an erratum?**
Currently, no. As errors are encountered, future versions of the dataset may be released and updated on our website.

(d) **Will the dataset be updated (*e.g.*, to correct labeling errors, add new instances, delete instances')?**
Yes, if applicable.

(e) **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (*e.g.*, were individuals in question told that their data would be retained for a fixed period of time and then deleted)?**
Not applicable. The dataset does not relate to people.

(f) **Will older versions of the dataset continue to be supported/hosted/maintained?**
Yes, older versions of the benchmark will be maintained on our website.

(g) **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?**
Yes, please get in touch with us by email.

(h) **Any other comments?**
None.