Published in the European Conference on Computer Vision Workshop:

Observing and Understanding Hands in Action (ECCVW-Hands), September 2024

ChildPlay-Hand: A Dataset of Hand Manipulations in the Wild

Arya Farkhondeh*, Samy Tafasca*, and Jean-Marc Odobez

Idiap Research Institute, Martigny, Switzerland
École Polytechnique Fédérale de Lausanne, Switzerland
{afarkhondeh, stafasca, odobez}@idiap.ch

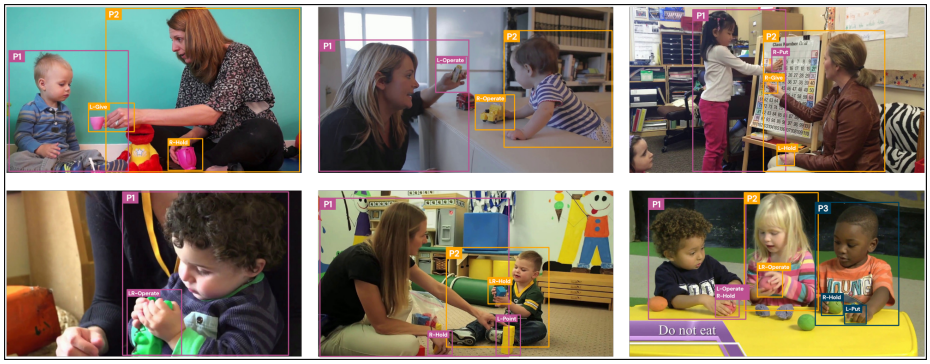

Fig. 1: Sample instances from the ChildPlay-Hand dataset with person bounding boxes and the per-hand object bounding boxes and corresponding action classes.

Abstract. Hand-Object Interaction (HOI) is gaining significant attention, particularly with the creation of numerous egocentric datasets driven by AR/VR applications. However, third-person view HOI has received less attention, especially in terms of datasets. Most third-person view datasets are curated for action recognition tasks and feature pre-segmented clips of high-level daily activities, leaving a gap for in-the-wild datasets. To address this gap, we propose ChildPlay-Hand, a novel dataset that includes person and object bounding boxes, as well as manipulation actions. ChildPlay-Hand is unique in: (1) providing per-hand annotations; (2) featuring videos in uncontrolled settings with natural interactions, involving both adults and children; (3) including gaze labels from the ChildPlay-Gaze dataset for joint modeling of manipulations and gaze. The manipulation actions cover the main stages of an HOI cycle, such as grasping, holding or operating, and different types of releasing. To illustrate the interest of the dataset, we study two tasks: object in hand detection (OiH), i.e. if a person has an object in their hand, and manipulation stages (ManiS), which is more fine-grained and targets the main stages of manipulation. We benchmark various spatio-temporal and segmentation networks, exploring body vs. hand-region information and comparing pose and RGB modalities. Our findings suggest that ChildPlay-Hand is a challenging new benchmark for modeling HOI in the wild.

* equal contribution

1 Introduction

Hands are our primary means of physical interaction with the environment, particularly through object manipulation. This explains why understanding hands in action has been an important topic in computer vision, focusing on various aspects such as recognizing gestures (see the recent review [24]) and hand-object interaction (HOI) [2,8,10,20,28]. This understanding naturally finds applications in various domains, including human-computer interaction [26], robotics [35], and augmented and virtual reality (AR/VR) [12].

Recently, much research has concentrated on HOI from a first-person perspective due to its applications in AR/VR. This has led to the creation of numerous egocentric video datasets, each tailored to specific settings and activities. These range from cooking activities in kitchen environments (FPHA [8] and EPIC-Kitchen [2]), to more diverse scenarios in Ego4D [10]. Also, some other HOI datasets are captured from multiple views and are collected in lab environment and come with rich data like 3D hand pose and/or 6D object poses thanks to the use of dedicated setup, as in H2O [20] and Assembly101 [28].

Despite the extensive efforts in egocentric HOI, the study of hands from a third-person perspective has received less attention. Most third-person view datasets [11,13,15,18,22,29–31] are designed for action recognition tasks, where isolated short clips are annotated with human activities. These datasets primarily focus on high-level human activities in daily life, such as driving a car or washing dishes, serving a different research purpose. As a result, there is a notable lack of datasets specifically concentrating on hands interacting with objects in the wild from a third-person view.

To address this gap, we introduce ChildPlay-Hand. It is derived from the ChildPlay [33] (or ChildPlay-Gaze) video dataset, which was originally introduced for addressing the gaze following task for children and toddlers, and features videos from childcare facilities and school settings. We selected this dataset because it contains both adults and children, and due to the rich diversity of hand manipulation behaviors mixed within children and adults interactions. Building upon this dataset, we provide dense annotations of body bounding boxes, hand manipulation action for each hand, and object bounding boxes of objects involved in such interactions. In terms of actions, we are interested in the main stages of the hand-object manipulation cycle: grasping, holding (passively having an object in hand) or operating (i.e. doing something actively with the object), and releasing (with several variants, see Sec. 3). While this may seem crude, this level of granularity allows to cover the entire hand-object interaction activities without being limited to the vocabulary of a particular application domain, and is anyway a first level of analysis that needs to be performed in HOI. Also, when addressed in the wild, it is actually very challenging.

The above annotation scheme results in a dataset that is unique in several ways. First, it provides **per-hand** annotations as hands can do multi-tasking. This is rare among existing datasets that typically focus on the activities of both hands as a whole. Second, it features videos in **uncontrolled settings** with varying camera views, showcasing scenes with multiple people interacting

naturally with each other and objects in a free manner. This is in stark contrast with datasets that assume a fixed view with a single person performing a given activity, where hands and objects typically stay visible. Third, the decision to use videos from ChildPlay-Gaze [33] will enable in the future the study of the **coordination between manipulations and gaze** from a third-person view, which, to the best of our knowledge, has not been explored before. Lastly, **the inclusion of children** (and not dominantly adults) in the dataset makes it a valuable source for behavioral studies of this demographic group. Such datasets are typically private due to the sensitive nature of the data.

ChildPlay-Hand can serve as a new benchmark for several tasks such as action recognition, action localization, and human-object interaction. In this work, we demonstrate its use and challenge for addressing two temporal action segmentation (TAS) tasks: object in hand (OiH) and manipulation stages (ManiS). The former is a coarse task of predicting whether an object is in a given hand, while the latter is more fine-grained, aiming to predict the main stages of hand-object manipulation interaction. For benchmarking, we explore different spatio-temporal networks like PoseConv3D [4], RGBPoseConv3D [4], and Hiera [27] by fine-tuning them on ChildPlay-Hand. We also explore other aspects such as the use of hand inputs compared to full body inputs. We then use the best spatio-temporal network to extract features for full hand sequences as input to TAS methods such as MS-TCN [5] and report frame-based and segmental metrics. In summary, our **contributions** are as follows:

- We propose ChildPlay-Hand, a unique and novel dataset of hand which provides per-hand activity annotations, features both adults and children, in uncontrolled and natural settings, and nicely complements existing gaze labels from ChildPlay-Gaze [33];
- We benchmark the dataset for two tasks, investigating different state-of-the-art spatio-temporal action recognition networks, as well as Temporal Action segmentation methods. Our work establishes several baselines, comparing pose-only networks to multimodal and visual-only networks, and explores the effectiveness of full-body vs. hand region inputs.

2 Related Datasets

2.1 Hand-Object Interaction (HOI) datasets

Recent video datasets of HOI are predominantly egocentric, each focusing on specific activities and settings. For instance, FPHA [8] and EPIC-Kitchen [2] capture cooking activities in kitchen environments, while Ego4D [10] contains more diverse scenes, environments, and daily activities. In addition to egocentric view, some datasets offer multiple views, typically captured in controlled lab settings with dedicated setups, such as H2O [20] and Assembly101 [28]. These datasets often include 3D hand poses and/or 6D object poses. In terms of activities, Assembly101 [28] contains fine-grained interaction verbs used to describe assembling and disassembling actions, such as "unscrew" and "remove" to achieve broader actions like "detach". Similarly, H2O [20] defines 11 action

verbs, including "open", "apply", and "read", to cover actions performed by participants when interacting with 8 distinct objects.

Despite great contributions to egocentric HOI, there is a notable lack of third-person view datasets capturing hand manipulations in uncontrolled environments. We fill this gap with ChildPlay-Hand. While Assembly101 [28] offers a multi-view setting and is larger in scale, our dataset stands out by moving away from lab environments. It includes scenes from uncontrolled settings captured from a third-person view, marking a significant step towards recognizing and analysing hand-object interactions in the wild.

2.2 Action Recognition

Unlike action recognition datasets, which feature pre-segmented clips of high-level activities typically performed by an individual, our dataset contains frame-wise annotations of per-hand manipulation actions performed by multiple people in the scene.

For instance, ActivityNet [13] or Kinetics [15] contains video clips annotated with high-level activities such as painting furniture, washing dishes, driving a car, and planting trees. In these clips, hands are part of the broader context of these actions, rather than the primary focus. NTU [22, 29] includes subsets of hand-centric actions such as "throw", "pick up", and "drop", performed in a scripted manner by individuals in a lab environment. These actions are isolated and lack the complexity of natural interactions. A subset of actions in AVA-Action [11] are hand-related, such as touching, carrying, and holding, annotated at 1Hz. Hence they lack the start and end of an action. Furthermore, the videos are typically low in quality, and contains scenes from movies in rather non-daily life situations, making it difficult to study hands in detail or spending efforts to provide additional annotations.

Something-Something [9] contains clips of finer-grained hand actions related to manipulating objects, offering a more focused view on hand-object interactions. The Charades [30] dataset contains videos of daily indoor activities where actors were tasked with performing scripted short-term actions, hence lacks the expected natural diversity of hand manipulations found in real situations.

2.3 Temporal Action Segmentation (TAS)

Widely used datasets for TAS typically focus on procedural activities, primarily confined to kitchen settings and cooking-related tasks. GTEA [6] is an egocentric dataset featuring videos recorded in a single kitchen, capturing activities such as making sandwiches and preparing coffee. 50Salads [32] contains videos of salad preparation captured from a top-down perspective, showcasing activities such as cutting vegetables and mixing ingredients. Breakfast [17] increases the diversity of kitchen environments and individuals performing actions related to breakfast preparation. The scenes are captured from a third-person view using 3 to 5 cameras. The dataset includes finer-level actions such as "taking a cup" and "pouring milk" as part of coarser activities like making coffee.

Our dataset can serve as yet another dataset for TAS, but it goes beyond kitchen environments and activities by focusing on per-hand manipulations in uncontrolled, free-play settings where inter-category and intra-category manipulation segments vary in length, making it a challenging dataset for TAS.

3 ChildPlay-Hand

In this section, we present ChildPlay-Hand by first defining the hand interaction class labels (Sec. 3.1), providing the annotation protocol (Sec. 3.2), presenting statistics of the annotations (Sec. 3.3), and comparing it with other datasets (Sec. 3.4).

3.1 Hand Interactions

We are interested in the main stages of a hand manipulation cycle that a person would consider to grossly description interactions with an object: *background* → *grasp* → *hold/operate* → *release* → *background*; these are complemented with other more precise activities (see below). Additionally, we annotated other actions beyond the aforementioned ones: the various pre-release gestures differentiating between different ways of getting rid of an object, such as dropping, putting/placing, throwing, and giving; and pointing. This was done given the nature of our video dataset, which features multi-person scenes, and due to the importance of these gestures for analysing children behaviors and adult-children interaction. In summary, the set of considered categories in ChildPlay-Hand and their definitions are as follows:

- **Background:** Hands Idle, i.e. not holding any object.
- **Grasp:** is the transition moment from idle hands to securing an object. It typically starts with a shift in visual attention, followed by approaching and then securing the object.
- **Hold:** Passively holding an object in the hand.
- **Operate:** Performing an intentional activity with the object (e.g., playing, disassembling, distorting, displacing).
- **Give:** Extend your hand towards a person with the intention of giving them an object (whatever the object is taken or not).
- **Put/Place:** Place an object on a surface.
- **Drop:** Releasing an object from a distance.
- **Throw:** Throw an object.
- **Release:** is the transition moment from having the object in hand to having idle hands by simply letting go, dropping, putting/placing, throwing, or giving it to someone (i.e. is situated at the end of the previous actions).
- **Point:** Pointing at an object, person or area if the object is not countable, e.g., a wall.

The above level of granularity has several advantages. It covers key moments of a complete hand-object interaction cycle without being too fine-grained or

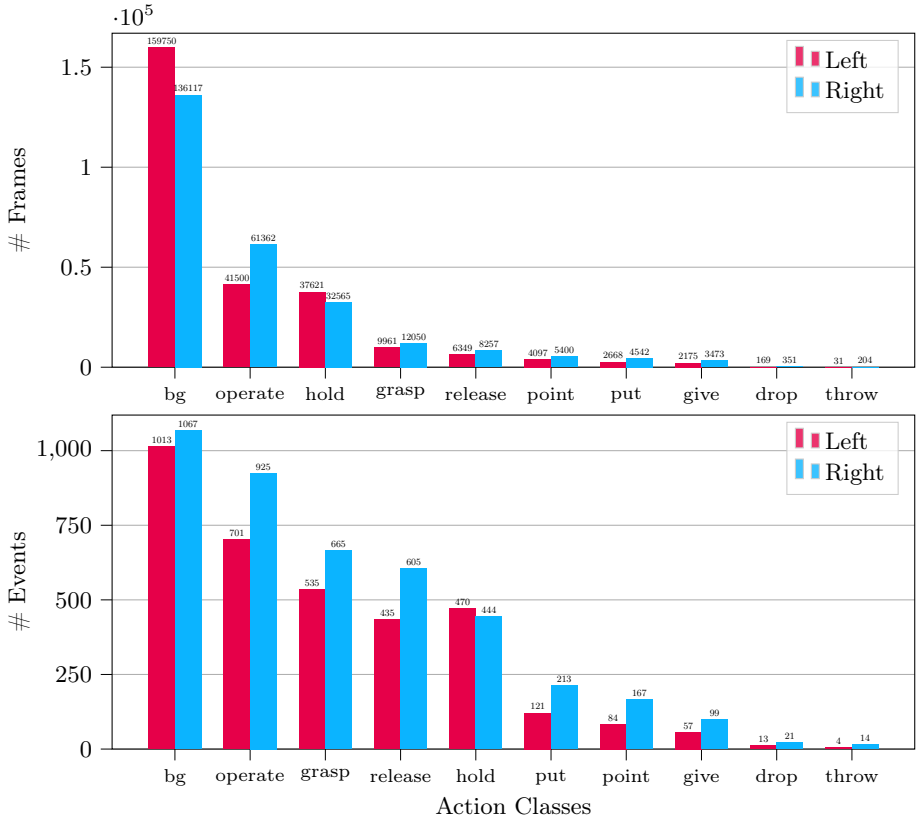

Fig. 2: Distribution of hand action classes in the dataset. We show the distribution in frames (top) and events (bottom).

coarse. For example, grasping could be broken down into micro-events like initiating intention, hand pre-shaping, reach planning, approaching, object contact, and securing the object. Similarly, a cycle or series of cycles, depending on the object and activity, can describe actions like assembling a toy, but this is not our aim here.

3.2 Annotation Protocol

Source of Videos: We borrowed video clips from ChildPlay-Gaze [33] and annotated them with the defined hand actions described in Sec. 3.1. This dataset was originally annotated with the 2D pixel location of gaze targets for the task of gaze following, focusing on children's gaze. The raw videos were downloaded from YouTube and feature children playing and interacting with adults in uncontrolled environments such as childcare facilities, schools, homes, and therapy centers. The dataset contains 401 video clips of high visual quality, mainly in indoor settings, featuring at least one child, often with one or two adults and multiple

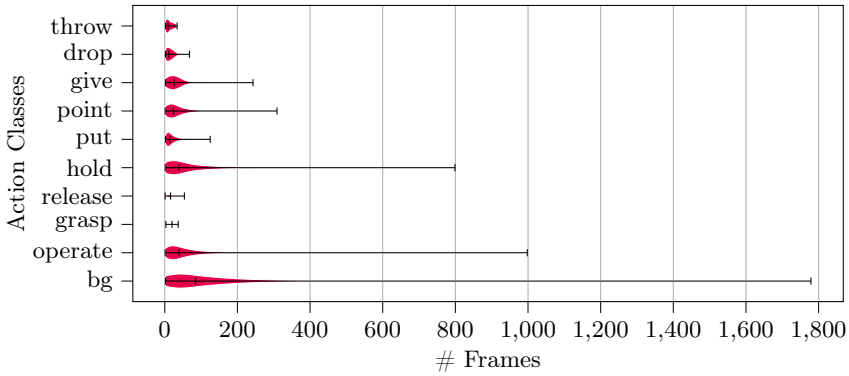

Fig. 3: Distribution of event duration (in frames) per action class. The violin plot shows the min, max and median values of each distribution.

children. Activities are unscripted, with the dominant activity typically being "playing with toys".

Annotation Process: We densely annotated ChildPlay-Gaze clips with body bounding boxes, per-hand interactions with objects, involved object bounding boxes, for up to three people in the scene, the same as those annotated with gaze. Seven annotators were assigned on the annotation process, followed by a review stage done by an expert in the field. The two action classes of `grasp` and `release` were not manually annotated because they can be inferred from the existing annotations at the transition moments (*e.g.* a person's hand changing from `background` to `hold` or `operate` has inevitably gone through a `grasp` event). We empirically set the duration for `grasp` to 20 frames, or approximately $0.6s$, with 4 frames inside the object manipulation segment and 16 frames outside (*i.e.* `background`). For `release`, this duration is split into 8 frames inside and 8 frames outside. When the duration of the preceding (resp. following) event for `grasp` (resp. `release`) is smaller than the outside frames (16 for `grasp` and 8 for `release`), we use the available frames instead. We conduct a manual verification step afterwards to ensure that the `grasp` and `release` instances are valid and acceptable.

3.3 Annotation Statistics

Figure 2 provides the per-hand distribution of action labels by frames and by event segments. Expectedly, the distribution features a long tail with actions like `give`, `drop`, and `throw` having much fewer instances than the rest. It is also interesting to note that the right hand `operate` has a higher frequency than the left counterpart, whereas the opposite can be observed for the `hold` class. This is likely due to the fact that a higher percentage of people is right-handed, so the active manipulation of objects is more often associated with the dominant right hand first, while the more passive `hold` is delegated to the left. We follow the training, validation, and test splits in ChildPlay-Gaze [33]. More information about the dataset splits can be found in the supplementary material.

Dataset	View	# Segments	# Actions	Per-Hands	In-the-Wild	Multi-Person
50Salads [32]	Top-Down	899	6	✗	✗	✗
Breakfast [17]	Exo	11,300	14	✗	✗	✗
Assembly101-Coarse [28]	Ego-Exo	104,759	11	✗	✗	✗
Assembly101-Fine [28]	Ego-Exo	1M	24	✗	✗	✗
ChildPlay-Hand	Exo	7,653	10	✓	✓	✓

Table 1: Comparison of ChildPlay-Hand with other related datasets. Note that # Actions refers to verbs in other datasets.

We also provide the frame-wise event duration distribution of each action class (*cf*. Figure 3). We observe that, aside from the `background` class which tends to be longer, all other actions typically last between 0 and 4 seconds (assuming a frame rate of 30 FPS), with holding and operating being relatively longer compared to the other more short-term and specific labels.

3.4 Comparison to Other Datasets

Tab. 1 compares ChildPlay-Hand across different aspects. While our dataset has fewer segments and actions due to the granularity we focus on, it has other unique aspects. ChildPlay-Hand features per-hand actions, captures in-the-wild scenes of children playing in a free-play manner in various backgrounds, and includes multiple people interacting with the same objects.

4 Experiments

In this section, we define the tasks (Sec. 4.1), discuss the selected models, protocol, implementation details for recognition (Sec. 4.2) and segmentation (Sec. 4.3).

4.1 Benchmarked Tasks

The ChildPlay-Hand dataset can be used for various video understanding tasks. Following the standard terminology, we can address Action Recognition by exploiting pre-segmented actions. It can also be used for Spatio-Temporal Action Localization, similar to the AVA-Action dataset [11], which involves not only recognizing but also localizing actions performed by all individuals within a key frame. Alternatively, the task can be formulated as an Human-Object Interaction recognition problem, by predicting all triplets within a key frame, <person bounding box, hand interaction, object bounding box>.

In this work, we focus on **Temporal Action Segmentation (TAS)**, where the goal is to perform frame-wise prediction $Y = \{y_1, \ldots, y_T\}$ on an input video $X = \{x_1, \ldots, x_T\}$ of a person where, $y_t \in \{a_L, a_R\}$ denote the left and right hand actions and can belong to one of C categories.

T1: Object in Hand (OiH): We first investigate the binary task of detecting whether a person has an object in a given hand. We generate labels by categorizing `hold` and `operate` as positive (OiH) and the rest as negative. The main motivation behind this task is that it serves as a prequel to segmenting different stages between non-OiH and OiH and evaluates the ability of a model to detect the presence of an object in hand. This task remains very challenging due to self-occlusions or hands being close (i.e., objects being occluded by hands), the small sizes and great variety of objects.

T2: Manipulation Stages (ManiS): Next, we target the primary stages of

an object manipulation cycle: hands begin in an idle state (`background`), grasp an object (`grasp`), perform a series of holding (`hold`) and operating actions (`operate`), release the object (`release`), and return to the idle state. To distinguish between these categories and segment them, a combination of two primary cues is needed: distinguishing among the hand motions and tracking the presence or absence of an object in the hands, thus making it challenging.

To perform frame-wise prediction of the entire input sequence, methods in TAS usually rely on pre-extracted features for each frame, e.g. using I3D [1]. However, these representations need to be informative and tailored to the task. To this end, we explore different spatio-temporal networks (detailed in Sec. 4.2) and fine-tune them on ChildPlay-Hand under a recognition protocol to measure their ability to recognize object and hand manipulations within a short window. Then, we use the top-performing recognition network to extract frame-wise representations for the TAS methods in Sec. 4.3.

4.2 OiH and ManiS Recognition

Below, we first detail the two types of video volume used as input to the networks and then provide a brief overview of the selected networks.

Inputs: We experiment with two types of inputs for the person of interest (PoI) in the scene: full-body and hand-region. For each type, we use the corresponding per-frame bounding boxes to crop out the body or hand of the PoI according to the Subject-Centered Cropping method from [4]: we compute the body/hand bounding box that encapsulates the PoI across frames in a short temporal window and use it to crop the input modalities, such as RGB frames and keypoints, across all frames in the window. For the body variant, we use the available ground-truth bounding boxes. For the hand, as ground-truth boxes are unavailable, we generate pseudo-hand bounding boxes as described in the supp. materials with illustrations.

Methods: PoseConv3D [4] is a strong state-of-the-art (SoA) model for skeleton-based action recognition. We experiment with PoseConv3D as a baseline to evaluate the performance of pose-only representation. Unlike graph-based methods [21] that use the coordinate-based representation of 3D body joints, PoseConv3D creates heatmap volumes of 2D body joints and has shown to be more robust. Each heatmap has a size of $K \times H \times W$. Here, K denotes the number of joints, represented by Gaussian maps centered at each joint. These heatmaps are then stacked along the temporal dimension to form a volume of size $K \times T \times H \times W$. Then, the input heatmap volumes are encoded using 3D-CNNs. For body keypoints, we extract $K = 17$ keypoints using HRNet [34], and for hand keypoints, we use $K = 21$ per hand, extracted using ZoomNet [14]. Note, to make use of the pre-trained PoseConv3D weights, which are trained on body keypoints, for use with hand keypoint heatmaps, we add an adapter layer to the model that maps the 21 channels of hand keypoints to the 17 channels expected by PoseConv3D, using a 3D-CNN layer.

However, pose alone (esp. body pose) is insufficient to recognize actions such as grasping, holding, and releasing, as these actions require appearance cues

to determine whether objects are in the hands. Therefore, we experiment with the multimodal version of PoseConv3D, RGBPoseConv3D [4], which has two pathways: a pose pathway operating at a higher frame rate and a visual pathway using RGB frames at a lower frame rate, similar to Slow-fast networks [7]. The fast branch for pose captures fine-grained motion of the keypoints, while the slow visual branch can provide appearance context (e.g., objects in the hands) with lateral connections between the two pathways. Following [4], we use 8 frames with a temporal stride of 4 for the visual pathway and 32 frames with no stride for the pose pathway.

Additionally, given the success of transformer-based architectures in video understanding, we experiment with a recent hierarchical transformer, Hiera [27]. This network has demonstrated strong performance, particularly on the AVA-Action [11] dataset, as also shown in LART [25]. Please refer to the supp. materials for the implementation details of these networks.

Recognition Metrics: For the recognition part, we compare models based on frame-based metrics only. For the OiH task, we use accuracy as well as precision, recall, and F1 of the OiH state. For the ManiS task, we compute accuracy as well as macro precision, recall, and F1 by averaging each of these metrics across the categories. The choice of macro metrics is due to the imbalanced nature of the dataset, treating all classes equally since they all represent main stages of hand-object manipulation.

4.3 OiH and ManiS Segmentation

Input: TAS methods process entire sequences as input and make frame-wise predictions. They rely on extracted features from spatio-temporal networks trained for recognition to form a feature grid sized $L \times D$, where L represents the sequence length and D denotes the feature dimension.

Methods: As a first baseline, we use the best frame-recognition network identified in Sec. 4.2 to produce frame-wise prediction via a standard Sliding-Window strategy. However, to account for a larger temporal context, smooth predictions and avoid over-segmentation, and learn the plausibility of manipulation activity sequence and ordering, we experiment with MS-TCN [5], a multi-stage convolutional-based network specifically designed for TAS. It relies on 1D convolutional kernels to process the feature grid along time, maintaining full temporal resolution across its multi-stage architecture. Each stage (SS-TCN) of MS-TCN comprises multiple layers of dilated convolutions, which are used to progressively increase the receptive field and to refine the predictions made by the previous stage.

Segmentation Metrics: In addition to the frame-based metrics, we compute segmental metrics. Given the predicted and ground-truth segments, we first find the optimal matching between them using Bipartite matching [19]. The cost function used is as follows:

$$C(i,j) = \begin{cases} 1 - O(D_i, G_j) & \text{if } label(D_i) = label(G_j) \\ 2 & \text{if } label(D_i) \neq label(G_j) \end{cases}$$

Method	Input	N. Param.	Pre-train	Window	T1: OiH				T2: ManiS			
					Acc.	F1	Prec.	Rec.	Acc.	m-F1	m-Prec.	m-Rec.
Dummy Classifier	-	-	-	-	53.6	35.3	31.4	40.3	40.0	19.4	19.8	19.7
PoseConv3D [4]	body	2M	NTU-60	48×1	66.5	32.6	44.3	25.8	63.0	30.1	36.0	28.9
PoseConv3D [4]	hand	2M	NTU-60	48×1	62.6	51.5	43.5	63.2	59.3	32.9	34.8	31.7
RGBPoseConv3D [4]	body	36M	NTU-60	8×4,32×1	72.4	51.5	57.7	46.6	66.0	33.8	44.9	31.2
RGBPoseConv3D [4]	hand	36M	NTU-60	8×4,32×1	75.1	53.0	65.2	44.7	67.7	36.8	**51.1**	33.4
Hiera [27]	body	52M	K400	16×2	79.5	65.4	69.8	61.5	66.5	45.0	45.2	45.7
Hiera [27]	hand	52M	K400	16×2	**85.4**	**74.3**	**83.4**	**67.0**	**73.2**	**51.2**	50.6	**53.1**

Table 2: Comparison of different networks. All metrics are frame-based. m-{metric} stands for macro. Stratified sampling is used as dummy classifier. Window of e.g., 16×2 refers to 16 frames with a temporal stride of 2.

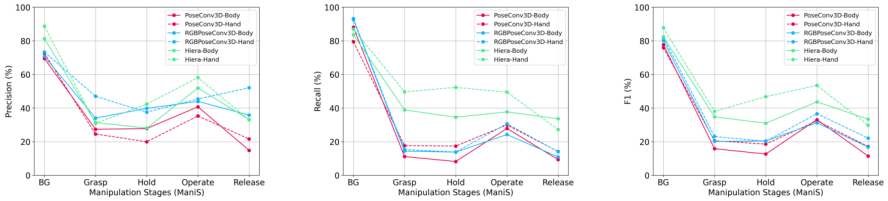

Fig. 4: Class-wise frame-based: Precision, Recall, and F1.

where $O(D_i, G_j)$ is the overlap between a predicted segment D_i and a ground-truth segment G_j, defined as:

$$O(D_i, G_j) = 2 \times \frac{\text{Prec.} \times \text{Rec.}}{\text{Prec.} + \text{Rec.}}, \quad \text{Prec.} = \frac{|D_i \cap G_j|}{|D_i|}, \quad \text{Rec.} = \frac{|D_i \cap G_j|}{|G_j|}$$

After optimal matching, a matched segment is said to be a true positive if its overlap is above 0 (and of the same class); otherwise, it is considered a false positive. Unmatched recognized segments become false positives, and unmatched ground-truth segments become false negatives. Using these counts, we compute precision, recall, and F1. We also report Edit score as in TAS methods [3].

5 Results

In this section, we first present the results of recognition experiments in Sec. 5.1, followed by segmentation results in Sec. 5.2.

5.1 Recognition Results

OiH Task: As shown in Tab. 2, the top-performing network is Hiera with hand inputs, achieving an accuracy of 85.4% and an F1 score of 74.3%, and outperforming other methods by a large margin in terms of F1. Beyond architectural differences, this performance may also be attributed to the pre-training on Kinetics-400 (K400) [15] which offers more diverse scenes than NTU-60 [29].

ManiS Task: Hiera applied on hand crops also outperforms the other approaches in this task, with the hand version leading at 73.2% accuracy and a 51.2% macro F1 score. However, RGBConvPose3D with hand inputs shows slightly better m-Prec. (+0.5%). Looking at class-wise performance in Fig. 4, the Hiera family generally delivers the best results across manipulation stages,

particularly in the `hold` and `operate` stages, where the hand version significantly outperforms others. However, other methods like RGBPoseConv3D-Hand exhibit better precision in the `grasp` and `release` stages. Despite Hiera-Hand achieving relatively better performances in the `hold` and `operate` stages with F1 scores above 40%, the transitional stages of `grasp` and `release` remain challenging. Confusion matrices can be found in the supp. materials.

Comparing full-body and hand-region inputs: In both tasks, hand variants consistently outperform their body counterparts in terms of F1 score. Looking at Fig. 4, this difference is more pronounced in certain classes depending on the method. For instance, Hiera-Hand outperforms its body variant more notably in the `hold` and `operate` stages, while the body version performs better on `release` with higher recall. In the case of RGBPoseConv3D, the hand variant shows notable improvements in the `operate` and `release` stages, while performing on par with the body variant on `hold`.

Indeed, hand-regions allow models to focus more directly on the main area of interest, which likely contributes to their improved performance. However, the body input also contains additional contextual information that could be valuable. For instance, the orientation and posture of the body, along with attentional cues, could help distinguish between intentional and unintentional hand movements, which is crucial for differentiating between `hold` and `operate`. It would be interesting to explore ways to leverage both hand-focused and full-body information together. We leave this for future work.

On the effectiveness of pose-only recognition: In OiH task, using pose-alone especially body-pose is indeed insufficient to recognize OiH state. This is reflected in Tab. 2, where RGBPoseConv3D-Body outperforms PoseConv3D-Body by a large margin in terms of F1 score. The same applies to RGBPoseConv3D-Hand and PoseConv3D-Hand, though with a smaller difference, likely due to the fact that hand pose implicitly contains information about objects in hand to some extent, whereas body-pose (with only the wrist keypoint as hand information) does not. Interestingly, PoseConv3D-Hand performs on par with RGBPoseConv3D-Body, indicating the effectiveness of hand-pose keypoints. In ManiS task, we can see a similar trend with RGBPoseConv3D outperforming PoseConv3D.

Looking at F1 scores in Fig. 4, with the body inputs, having appearance cues helps more notably with `release`, `hold`, and `grasp` while pose-only performs slightly better on `operate`. This shows that while body-pose can be effective when manipulating an object, other stages indeed need appearance cues. With hand inputs, appearance cues help more with `operate` and `release` while performing more closely to each other on `hold` and `grasp`.

5.2 Segmentation Results

OiH Task: Tab. 3 reports the segmentation results obtained with the methods discussed in Sec. 4.3. In terms of segmental F1 (S-F1), both SS-TCN and MS-TCN outperform the sliding-window approach, which is expected since the

Method	T1: OiH						T2: ManiS					
	S-Prec.	S-Rec.	S-F1	F-Acc.	F-Prec.	F-Rec.	m-S-Prec.	m-S-Rec.	m-S-F1	Edit	F-Acc.	m-F-F1
Sliding-Window	18.5	**87.3**	30.5	85.4	83.4	67.0	13.1	**78.6**	22.1	34.2	73.2	51.2
SS-TCN [5]	41.2	81.0	54.6	**86.5**	**84.2**	70.3	28.9	64.2	38.9	61.6	77.7	53.9
MS-TCN [5]	**54.7**	81.0	**65.3**	86.2	83.2	**70.5**	**41.7**	64.5	**49.6**	**66.1**	**78.1**	**54.0**

Table 3: Comparison of TAS methods. S-{metric} refers to segmental, F-{metric} to frame-based, and m-{metric} to macro.

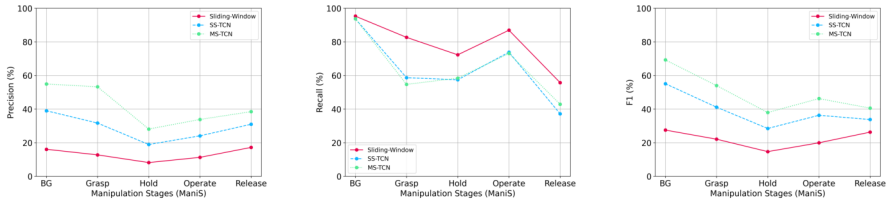

Fig. 5: Class-wise segmental: Precision, Recall, and F1.

sliding-window method cannot leverage long-range temporal context. These methods improve segmental precision (S-Prec.) by removing spurious short term predictions which strongly affect this metric. Additionally, they also slightly improve frame-based metrics. When comparing SS-TCN and MS-TCN, MS-TCN further improves the segmental metrics while maintaining the same frame accuracy.

ManiS Task: Here, MS-TCN also outperforms the other models largely due to the improvement on the precision metric as in the OiH task, at the cost of only a slight degradation of the recall (esp. for `grasp`). We can also notice that the frame-based metrics improves, showing that some regularization helps here as well. In general, these results highlight that multiple stages are are beneficial for handling the task's complexity, unlike in OiH where more stages did not improve frame-based metrics. Fig. 5 shows that MS-TCN consistently achieves the highest F1 scores across all classes.

Qualitative Examples: Fig. 6 shows qualitative examples (QEs) of predicted segments. In QE1, all methods perform well on both tasks, although MS-TCN fails to predict the first `grasp` segment, likely due to over-smoothing. In QE2, while all methods perform well on the OiH task, they struggle to accurately predict the `hold` segment in the ManiS task. This is due to slight hand motion made by the person while passively holding the object, leading the methods to predict `operate` instead. In this case, the hand-region alone does not provide enough context to determine whether the person is intentionally moving the object. In QE3, the predictions are of lower quality. For example, in the ManiS task, `background` is incorrectly predicted as `grasp`, due to confusion between the motions of pointing and grasping. Also, the last `release` mostly is predicted as `operate` while in this stage there is no object in hand. In the OiH task, all methods struggle to accurately predict the boundaries of the OiH state, especially when the hands are close to the object during the `release`.

These observations again highlight the difficulty of modeling hands in action in the wild, where hand movements can extend beyond object manipulation (e.g., pointing), objects come in varying shapes and sizes, and multiple people

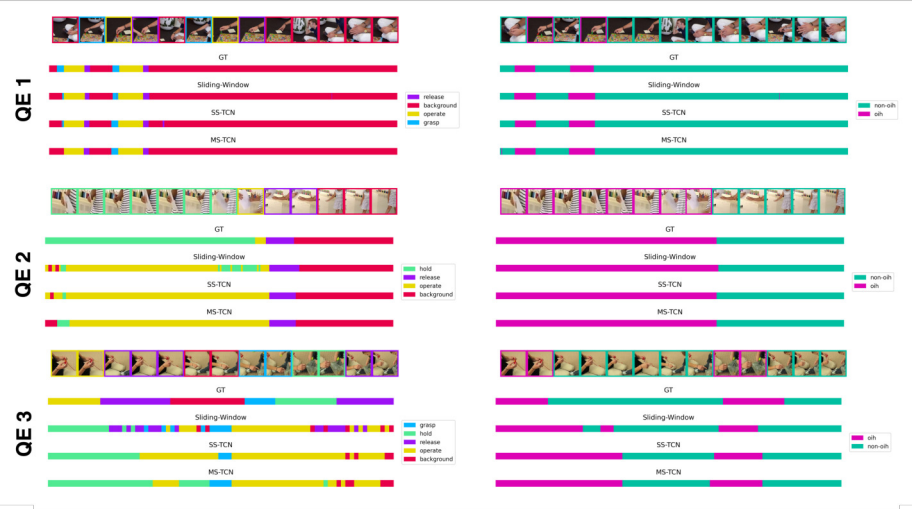

Fig. 6: Qualitative results of predicted segments for hand interaction sequences in ChildPlay-Hand. The left figure corresponds to the ManiS task, and the right figure to the OiH task. Hand-region frames for each segment are highlighted using the same color as their label, representing the key moments within each segment. Note that the hand regions are cropped using the procedure discussed in Sec. 4.2.

may interact with the same object or be in close proximity, making accurate prediction of hand manipulations challenging.

6 Conclusion

In this work, we introduced ChildPlay-Hand, a novel dataset of hand manipulations in the wild. This dataset is unique for its rich annotations and scenes featuring multiple people interacting naturally with objects and each other in uncontrolled settings. These annotations enable the use of ChildPlay-Hand for a variety of tasks and protocols, including spatio-temporal action localization, pre-segmented action recognition, and human-object interaction. Future research can also leverage the accompanying gaze labels to explore the coordination between manipulations and visual attention. In this work, we proposed two specific tasks under recognition and temporal segmentation protocols: object in hand (OiH) and manipulation stages (ManiS). We benchmarked various spatio-temporal models with varied modalities and input types as well as segmentation models, on these tasks. Our findings indicate that ChildPlay-Hand can serve as a challenging benchmark for understanding hands in action from a third-person view in uncontrolled settings.

Acknowledgement. This research was supported by the AI4Autism project (digital phenotyping of autism spectrum disorders in children, grant agreement number CRSII5_202235 / 1) of the Sinergia interdisciplinary program of the SNSF.

References

1. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 4724–4733 (2017)
2. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Scaling egocentric vision: The epic-kitchens dataset. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)
3. Ding, G., Sener, F., Yao, A.: Temporal action segmentation: An analysis of modern techniques. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**, 1011–1030 (2022)
4. Duan, H., Zhao, Y., Chen, K., Lin, D., Dai, B.: Revisiting skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2969–2978 (June 2022)
5. Farha, Y.A., Gall, J.: Ms-tcn: Multi-stage temporal convolutional network for action segmentation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3570–3579 (2019)
6. Fathi, A., Ren, X., Rehg, J.M.: Learning to recognize objects in egocentric activities. CVPR 2011 pp. 3281–3288 (2011)
7. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 6201–6210 (2018)
8. Garcia-Hernando, G., Yuan, S., Baek, S., Kim, T.K.: First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In: Proceedings of Computer Vision and Pattern Recognition (CVPR) (2018)
9. Goyal, R., Kahou, S.E., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fründ, I., Yianilos, P.N., Mueller-Freitag, M., Hoppe, F., Thurau, C., Bax, I., Memisevic, R.: The "something something" video database for learning and evaluating visual common sense. 2017 IEEE International Conference on Computer Vision (ICCV) pp. 5843–5851 (2017)
10. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., Martin, M., Nagarajan, T., Radosavovic, I., Ramakrishnan, S.K., Ryan, F., Sharma, J., Wray, M., Xu, M., Xu, E.Z., Zhao, C., Bansal, S., Batra, D., Cartillier, V., Crane, S., Do, T., Doulaty, M., Erapalli, A., Feichtenhofer, C., Fragomeni, A., Fu, Q., Gebreselasie, A., González, C., Hillis, J., Huang, X., Huang, Y., Jia, W., Khoo, W., Kolář, J., Kottur, S., Kumar, A., Landini, F., Li, C., Li, Y., Li, Z., Mangalam, K., Modhugu, R., Munro, J., Murrell, T., Nishiyasu, T., Price, W., Ruiz, P., Ramazanova, M., Sari, L., Somasundaram, K., Southerland, A., Sugano, Y., Tao, R., Vo, M., Wang, Y., Wu, X., Yagi, T., Zhao, Z., Zhu, Y., Arbeláez, P., Crandall, D., Damen, D., Farinella, G.M., Fuegen, C., Ghanem, B., Ithapu, V.K., Jawahar, C.V., Joo, H., Kitani, K., Li, H., Newcombe, R., Oliva, A., Park, H.S., Rehg, J.M., Sato, Y., Shi, J., Shou, M.Z., Torralba, A., Torresani, L., Yan, M., Malik, J.: Ego4d: Around the world in 3,000 hours of egocentric video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18995–19012 (June 2022)
11. Gu, C., Sun, C., Ross, D.A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., Schmid, C., Malik, J.: Ava: A video dataset of spatio-temporally localized atomic visual actions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)

12. Hasson, Y., Varol, G., Tzionas, D., Kalevatykh, I., Black, M.J., Laptev, I., Schmid, C.: Learning joint reconstruction of hands and manipulated objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
13. Heilbron, F.C., Escorcia, V., Ghanem, B., Niebles, J.C.: Activitynet: A large-scale video benchmark for human activity understanding. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 961–970 (2015)
14. Jin, S., Xu, L., Xu, J., Wang, C., Liu, W., Qian, C., Ouyang, W., Luo, P.: Whole-body human pose estimation in the wild. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020)
15. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, A., Suleyman, M., Zisserman, A.: The kinetics human action video dataset. ArXiv **abs/1705.06950** (2017)
16. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR **abs/1412.6980** (2014)
17. Kuehne, H., Arslan, A.B., Serre, T.: The language of actions: Recovering the syntax and semantics of goal-directed human activities. 2014 IEEE Conference on Computer Vision and Pattern Recognition pp. 780–787 (2014)
18. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: a large video database for human motion recognition. In: 2011 International conference on computer vision. pp. 2556–2563. IEEE (2011)
19. Kuhn, H.W.: The hungarian method for the assignment problem. Naval Research Logistics (NRL) **52** (1955)
20. Kwon, T., Tekin, B., Stühmer, J., Bogo, F., Pollefeys, M.: H2o: Two hands manipulating objects for first person interaction recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10138–10148 (2021)
21. Li, C., Cui, Z., Zheng, W., Xu, C., Yang, J.: Spatio-temporal graph convolution for skeleton based action recognition. In: AAAI Conference on Artificial Intelligence (2018)
22. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.Y., Kot, A.C.: Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. IEEE Trans. Pattern Anal. Mach. Intell. **42**(10), 2684âĂŞ2701 (2020)
23. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2017)
24. Mohamed, N., Mustafa, M.B., Jomhari, N.: A review of the hand gesture recognition system: Current progress and future directions. IEEE Access **PP**, 1–1 (2021)
25. Rajasegaran, J., Pavlakos, G., Kanazawa, A., Feichtenhofer, C., Malik, J.: On the benefits of 3d pose and tracking for human action recognition. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 640–649 (2023)
26. Rautaray, S.S., Agrawal, A.: Vision based hand gesture recognition for human computer interaction: a survey. Artificial Intelligence Review **43**, 1 – 54 (2012)
27. Ryali, C., Hu, Y.T., Bolya, D., Wei, C., Fan, H., Huang, P.Y., Aggarwal, V., Chowdhury, A., Poursaeed, O., Hoffman, J., Malik, J., Li, Y., Feichtenhofer, C.: Hiera: A hierarchical vision transformer without the bells-and-whistles. ICML (2023)
28. Sener, F., Chatterjee, D., Shelepov, D., He, K., Singhania, D., Wang, R., Yao, A.: Assembly101: A large-scale multi-view video dataset for understanding procedural activities. CVPR (2022)

29. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+d: A large scale dataset for 3d human activity analysis. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1010–1019 (2016)
30. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.K.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: European Conference on Computer Vision (2016)
31. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
32. Stein, S., McKenna, S.J.: Combining embedded accelerometers with computer vision for recognizing food preparation activities. Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing (2013)
33. Tafasca, S., Gupta, A., Odobez, J.M.: Childplay: A new benchmark for understanding children's gaze behaviour. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20935–20946 (2023)
34. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., Xiao, B.: Deep high-resolution representation learning for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**, 3349–3364 (2019)
35. Yang, Y., Li, Y., Fermüller, C., Aloimonos, Y.: Robot learning manipulation action plans by "watching" unconstrained videos from the world wide web. In: AAAI Conference on Artificial Intelligence (2015)

Supplementary Material

7 More Statistics from the ChildPlay-Hand Dataset

The dataset partitioning for training, validation and testing follows the same splits introduced in [33], as we found them to be well balanced for the new annotation task already. Specifically, the class distributions across these splits were fairly similar, as can be shown in Figure 7. However, it is worth noting that the classes with a low frequency evidently feature even fewer instances after splitting (*e.g.* the `throw` action has only 36 and 21 frames in the validation and test sets, respectively).

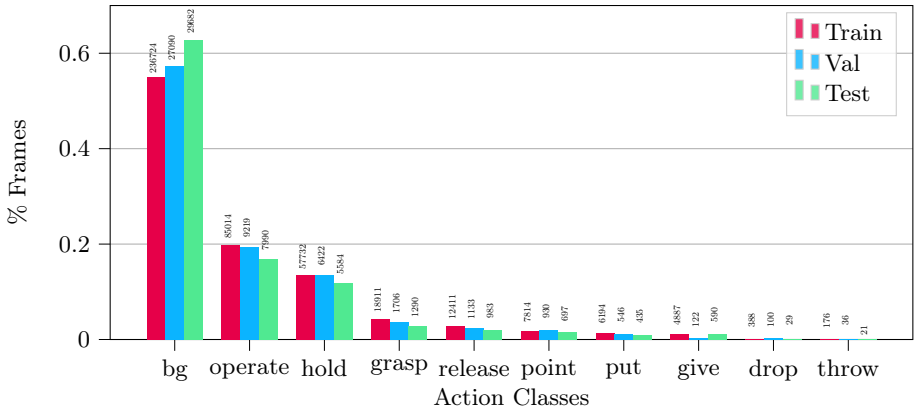

Fig. 7: Distribution of annotations per train/val/test splits. We show the distribution in frames (on top of each bar) and in percentage (y-axis).

8 Implementation Details

Training instances: To create training instances for recognition methods, we sample key frames to serve as the target class for classification and take frames around them to create a short segment. We use every frame in the training set as the middle frame for classification, except for the background class, where we sample every third frame. Frames where the person is occluded are excluded from both training and evaluation.

Training details of recognition: When full-body information is used as input, an instance is a person, so we use two classification heads, one for each hand (left and right). During training, we apply horizontal flip as the only augmentation, flipping the ground-truth hand labels accordingly. However, when the hand is used as the input, an instance is a single hand of a person, so we

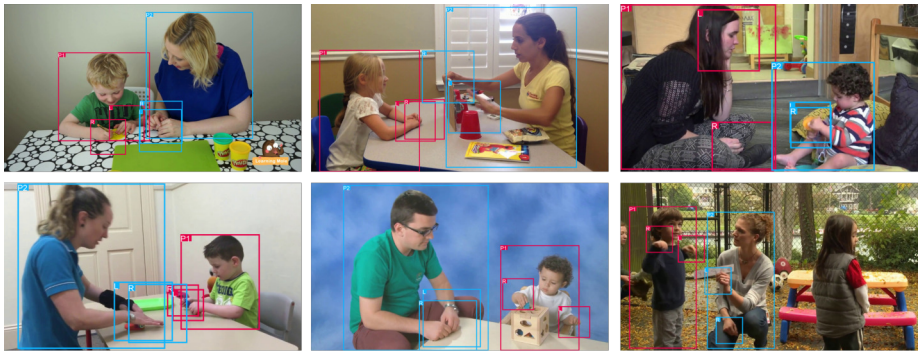

Fig. 8: Examples of the ground-truth body and pseudo-hand bounding boxes in ChildPlay-Hand. These bounding boxes are used in the cropping strategy described in Sec. 4.2 of the main submission to create input volumes for the networks.

use only one classification head and do not flip the labels when horizontal flip is applied. The classification head consists of two MLPs with a ReLU activation in between, which map the temporally mean-pooled features to one of the C categories. We fine-tune all recognition networks for 15 epochs. For PoseConv3D and RGBPoseConv3D models, we use SGD optimizer and set the learning rate of the backbone to 0.0018 and the MLP heads to 0.018, with linear decay at epochs 9 and 13. For Hiera models, we use Hiera-B version and finetune it using the AdamW [23] optimizer with a learning rate of 2e-5 for the backbone and 2e-4 for the MLP heads, applying cosine decay as the scheduler.

Training details of segmentation: We train MS-TCN for 50 epochs using the Adam [16] optimizer with a learning rate of 0.0005, selecting the best checkpoint based on the validation set. The number of layers is set to 10 for both SS-TCN and MS-TCN, with the number of stages set to 4 for MS-TCN.

9 Generating Pseudo-Hand Bounding Boxes

We generate pseudo-hand bounding boxes as follows: (1) we define the center of the box at the wrist keypoint, extended by 50% of the elbow-to-wrist limb, and (2) set the box size to 40% of the smallest side of the body bounding box. Note, while a hand detector could typically be used, detecting and associating hands with body bounding boxes in our unconstrained setting is challenging. Using hand keypoints from a whole-body pose estimator is another option, but these keypoints can be noisy due to self/object occlusions, leading to inaccurate hand bounding boxes. Instead, we find that the wrist and elbow keypoints from 2D body pose estimators like HRNet [34] is robust enough for generating pseudo-hand bounding boxes. See illustrations in Figure 8.

10 On confusion among manipulation categories

After looking at the confusion matrices in Fig. 9 and Fig. 10, we can see a common trend that the major confusion, especially for classes with fewer instances,

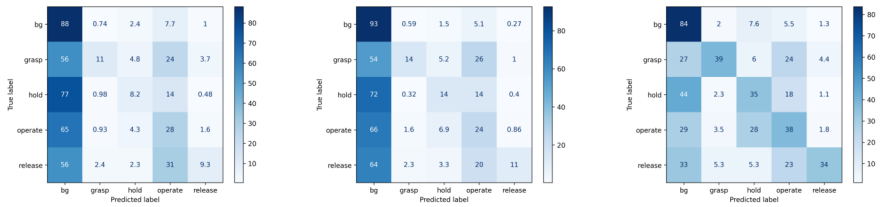

Fig. 9: Confusion Matrices of PoseConv3D, RGBPoseConv3D, and Hiera networks with `body` input.

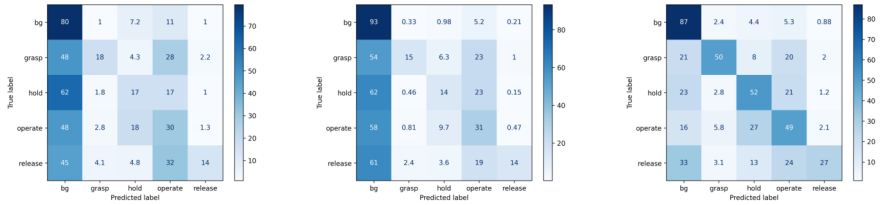

Fig. 10: Confusion Matrices of PoseConv3D, RGBPoseConv3D, and Hiera networks with `hand` input.

namely `grasp` and `release`, occurs with the majority classes, `background` and `operate`. One way to mitigate this is to perform balanced sampling to create balanced mini-batches or apply different weights for the classes at the loss level. Moreover, there is an inherent confusion between `hold` and `operate` since these two classes share the same OiH state, but what differentiates them is not only motion but also intention, making it challenging to distinguish between them, even during annotation. Gaze can potentially help in such cases by acting as a cue to infer intention.