The IKEA ASM Dataset: Understanding People Assembling Furniture through Actions, Objects and Pose

Yizhak Ben-Shabat[1,2] Xin Yu[3] Fatemeh Sadat Saleh[1,2] Dylan Campbell[1,2]
Cristian Rodriguez-Opazo[1,2] Hongdong Li[1,2] Stephen Gould[1,2]

[1]Australian National University [2]Australian Centre for Robotic Vision (ACRV)
[3]University of Technology Sydney

https://ikeaasm.github.io/

Abstract

The availability of a large labeled dataset is a key requirement for applying deep learning methods to solve various computer vision tasks. In the context of understanding human activities, existing public datasets, while large in size, are often limited to a single RGB camera and provide only per-frame or per-clip action annotations. To enable richer analysis and understanding of human activities, we introduce IKEA ASM—a three million frame, multi-view, furniture assembly video dataset that includes depth, atomic actions, object segmentation, and human pose. Additionally, we benchmark prominent methods for video action recognition, object segmentation and human pose estimation tasks on this challenging dataset. The dataset enables the development of holistic methods, which integrate multi-modal and multi-view data to better perform on these tasks.

1. Introduction

Furniture assembly understanding is closely related to the broader field of action recognition. The rise of deep learning has rapidly advanced this field [8]. However, deep learning models require vast amounts of training data and are often evaluated on large datasets of short video clips, typically extracted from YouTube [8, 33], that include a set of arbitrary yet highly discriminative actions. Therefore, the research on assembly understanding is far behind generic action recognition due to the insufficient datasets for training such models and other challenges such as the need to understand longer timescale activities. Existing assembly datasets [68] are limited to the classification of very few actions and focus on human pose and color information only.

We aim to enable research of assembly understanding and underlying perception algorithms under real-life conditions by creating diversity in the assembly environment, assemblers, furniture types and color, and body visibility. To this end, we present the novel IKEA ASM dataset, the first publicly available dataset with the following properties:

- **Multi-modality**: Data is captured from multiple sensor modalities including color, depth, and surface normals. It also includes various semantic modalities including human pose and object instance segmentation.
- **Multi-view**: Three calibrated camera views cover the work area to handle body, object and self occlusions.
- **Fine-grained**: There is subtle distinction between objects (such as table top and shelf) and action categories (such as aligning, spinning in, and tightening a leg), which are all visually similar.
- **High diversity**: The same furniture type is assembled in numerous ways and over varying time scales. Moreover, human subjects exhibit natural, yet unusual poses, not typically seen in human pose datasets.
- **Transferability**: The straightforward data collection protocol and readily available furniture makes the dataset easy to reproduce worldwide and link to other tasks such as robotic manipulation of the same objects.

While the task of furniture assembly is simple and well-defined, there are several difficulties that make inferring actions and detecting relevant objects challenging. First, unlike standard activity recognition the background does not provide any information for classifying the action (since all actions take place in the same environment). Second, parts being assembled are symmetric and highly similar requiring understanding of context and the ability to track objects relative to other parts and sub-assemblies. Third, the strong visual similarity between actions and parts requires a higher-level understanding of the assembly process and state information to be retained over long time periods.

On the other hand, the strong interplay between geometry and semantics in furniture assembly provides an opportunity to model and track the process. Moreover, cues

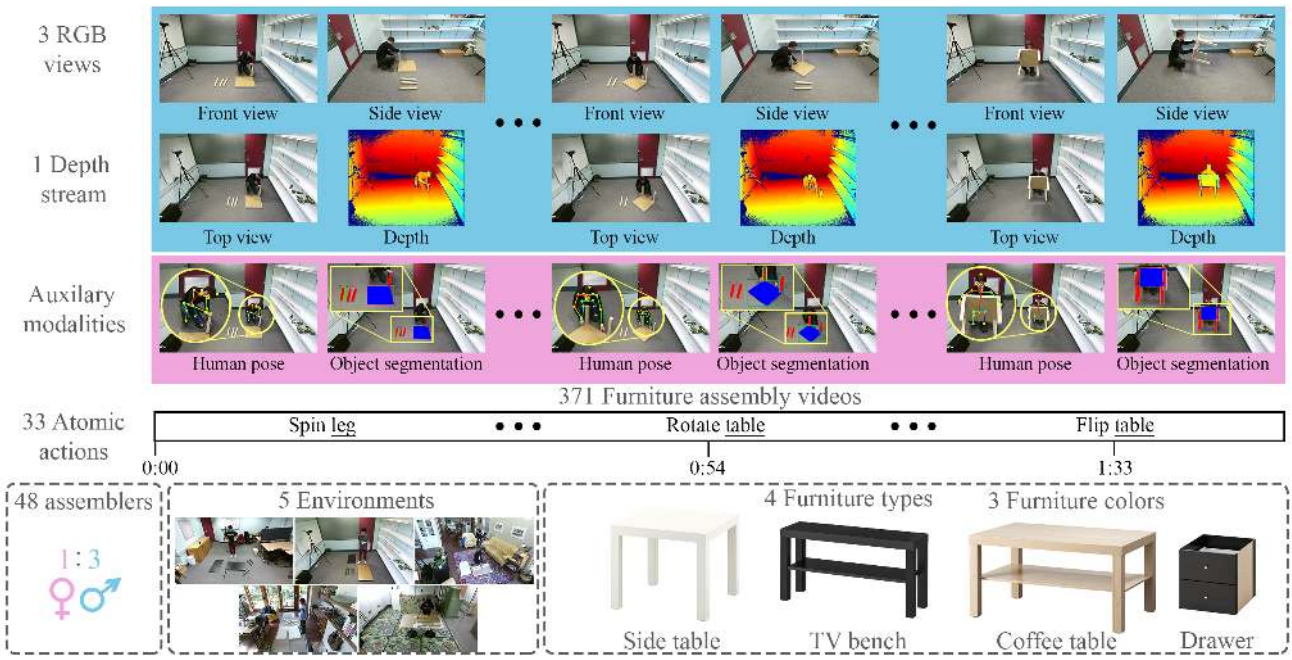

Figure 1: Overview of the IKEA ASM dataset. The dataset contains 371 furniture assembly videos from three camera views, including 3D depth, object segmentation and human pose annotations of 33 atomic actions.

obtained from the different semantic modalities, such as human pose and object types, combine to provide strong evidence for the activity being performed. Our dataset enables research along this direction where both semantics and geometry are important and where short-term feed-forward perception is insufficient to solve the problem.

The main contributions of this paper are: (1) the introduction of a novel furniture assembly dataset that includes multi-view, and multi-modal annotated data; and (2) evaluation of baseline method for different tasks (action recognition, pose estimation, object instance segmentation and tracking) to establish performance benchmarks.

2. Background and related work

Related Datasets. The increasing popularity of action recognition in the computer vision community has led to the emergence of a wide range of action recognition datasets. One of the most prominent datasets for action recognition is the Kinetics [52] dataset—a large-scale human action dataset collected from Youtube videos. It is two orders of magnitude larger than some predecessors, e.g. the UCF101 [65] and HMDB51 [34]. Additional notable datasets in this context are ActivityNet [18] and Charades [63], which include a wide range of human activities in daily life. The aforementioned datasets, while very large in scale, are not domain specific or task-oriented. Additionally, they are mainly centered on single-view RGB data.

Instructional video datasets usually include domain specific videos, e.g., cooking (MPII [57], YouCook [15], YouCook2 [77], EPIC-Kitchens [14]) and furniture assembly (IKEA-FA [68]). These are most often characterized by having fine grained action labels and may include some additional modalities to the RGB stream such as human pose and object bounding boxes. There are also more diverse variants like the recent COIN [67] dataset, which forgoes the additional modalities in favor of a larger scale.

The most closely related to the proposed dataset are the Drive & Act [45] and NTU RGB+D [62, 40] datasets. Drive & Act is specific to the domain of in-car driver activity and contains multi-view, multi-modal data, including IR streams, pose, depth, and RGB. While the actors follow some instructions, their actions are not task-oriented in the traditional sense. Due to the large effort in collecting it, the total number of videos is relatively low (30). Similarly, NTU RGB+D [62] and its recent extension NTU RGB+D 120 [40] contain three different simultaneous RGB views, IR and depth streams as well as 3D skeletons. However, in this case the videos are very short (few seconds), non-instructional and are focused on general activities, some of which are health related or human interaction related. For a detailed quantitative comparison between the proposed and closely-related datasets see Table 1.

Other notable work is the IKEA Furniture Assembly Environment [35], a simulated testbed for studying robotic manipulation. The testbed synthesizes robotic furniture assembly data for imitation learning. Our proposed dataset is

Dataset	Year	Dur.	#Videos	#Frames	Activity type	Source	Views	3D	Human pose	object seg.	bb
MPII Cooking[57]	2012	9h,28m	44	0.88M	cooking	collected	1	✗	✓	✗	
YouCook [15]	2013	2h,20m	88	NA	cooking	YouTube	1	✗	✗	✓ (bb)	
MPII Cooking 2 [58]	2016	8h	273	2.88M	cooking	collected	1	✗	✓	✗	
IKEA-FA [68]	2017	3h,50m	101	0.41M	assembly	collected	1	✗	✓	✗	
YouCook2 [77]	2018	176h	2000	NA	cooking	YoutTube	1	✗	✗	✓ (bb)	
EPIC-Kitchens [14]	2018	55h	432	11.5M	cooking	collected	1	✗	✗	✓ (bb)	
COIN [67]	2019	476h,38m	11827	NA	180 tasks	YouTube	1	✗	✗	✗	
Drive&Act [45]	2019	12h	30	9.6M	driving	collected	6	✓	✓	✗	
IKEA-ASM	2020	35h,16m	371	3M	assembly	collected	3	✓	✓	✓	

Table 1: Instructional video dataset comparison.

complimentary to this work as it captures real-world data of humans that can be used for domain-adaptation.

In this paper we propose a furniture assembly domain-specific, instructional video dataset with multi-view and multi-modal data, which includes fine grained actions, human pose, object instance segmentation and tracking labels.

Related methods. We provide a short summary of methods used as benchmarks in the different dataset tasks including action recognition, instance segmentation, multiple object tracking and human pose estimation. For an extended summary, see the supplementary material.

Action Recognition . Current action recognition architectures for video data are largely image-based. The most prominent approach uses 3D convolutions to extract spatio-temporal features,and includes methods like convolutional 3D (C3D) [69], which was the first to apply 3D convolutions in this context, pseudo-3D residual net (P3D ResNet) [52], which leverages pre-trained 2D CNNs and utilizes residual connections and simulates 3D convolutions, and the two-stream inflated 3D ConvNet (I3D) [8], which uses an inflated inception module architecture and combines RGB and optical flow streams. Other approaches attempt to decouple visual variations by using a mid-level representation like human pose (skeletons). One idea is to use a spatial temporal graph CNN (ST-GCN) [74] to process the skeleton's complex structure. Another is to learn skeleton features combined with global co-occurrence patterns [36].

Instance Segmentation. Early approaches to instance segmentation typically perform segment proposal and classification in two stages [50, 13, 49]. Whereas recent one-stage approaches tend to be faster and more accurate [22, 37]. Most notably, Mask R-CNN [22] combines binary mask prediction with Faster R-CNN [55], showing impressive performance. They predict segmentation masks on a coarse grid, independent of the instance size and aspect ratio which tends to produce coarse segmentation for instances occupying larger part of the image. To alleviate this problem approaches have been proposed to focus on the boundaries of larger instances, e.g., InstanceCut [29], TensorMask [9], and point-based prediction as in PointRend [30].

Multiple Object Tracking (MOT). Tracking-by-detection is a common approach for multiple object tracking. MOT can be considered from different aspects: It can be categorized into online or offline, depending on when the decisions are made. In online tracking [60, 71, 3, 12, 73, 28], the tracker assigns detections to tracklets at every time-step, whereas in offline tracking [66, 44] the decision about the tracklets are made after observing the whole video. Different MOT approaches can also be divided into geometry-based [60, 5] or appearance-based [12, 3, 73]. In our context, an application may be human-robot collaboration during furniture assembly, where the tracking system is required to make real-time online decisions [60, 5]. In this scenario, IKEA furniture parts are almost textureless and of the same color and shape, and thus the appearance information could be misleading. Additionally, IKEA furniture parts are rigid, non-deformable objects, that are moved almost linearly in a short temporal window. As such, a simple, well-designed tracker that models linear motions [5] is a reasonable choice.

Human Pose Estimation. Multi-person 2D pose estimation methods can be divided into bottom-up (predict all joints first) [51, 7, 6, 53] or top-down (detect all person bounding boxes first) [22, 19, 10]. The popular OpenPose detector [7, 6] assembles the skeleton using a joint detector and part affinity fields. This was extended to incorporate temporal multi-frame information in Spatio-Temporal Affinity Fields (STAF) [53]. Mask R-CNN [22] is a notable top-down detection-based approach, where a keypoint regression head can be learned alongside the bounding box and segmentation heads. Monocular 3D human pose estimation methods can be categorized as being model-free [47, 48] or model-based [25, 26, 32, 31]. The former include VideoPose3D [48] which estimates 3D joints via temporal convolutions over 2D joint detections in a video se-

quence. The latter approach predicts the parameters of a body model, often the SMPL model [43], such as the joint angles, shape parameters, and rotation. Some model-based approaches [25, 26, 31] leverage adversarial learning to produce realistic body poses and motions. Therefore, they tend to generalize better to unseen datasets, and so we focus on these methods as benchmarks on our dataset.

3. The IKEA assembly dataset

The IKEA ASM video dataset will be made publicly available for download of all 371 examples and ground-truth annotations. It includes three RGB views, one depth stream, atomic actions, human poses, object segments, and extrinsic camera calibration. Additionally, we provide code for data processing, including depth to point cloud conversion, surface normal estimation, visualization, and evaluation in a designated github repository.

Data collection. Our data collection hardware system is composed of three Kinect V2 cameras. These three cameras are oriented to collect front, side and top views of the work area. In particular, the top-view camera is set to acquire the scene structure. The front and side-view cameras are placed at eye-level height (∼1.6m). The three Kinect V2 cameras are triggered to capture the assembly activities simultaneously in real time (∼24 fps). To achieve real-time data acquisition performance, multi-threaded processing is used to capture and save images on an Intel i7 8-core CPU with NVIDIA GTX 2080 Ti GPU used for data encoding.

To collect our IKEA ASM dataset, we ask 48 human subjects to assemble furniture in five different environments, such as offices, labs and family homes. In this way, the backgrounds are diverse in terms of layout, appearance and lighting conditions. The background is dynamic, containing moving people who are not relevant to the assembly process. These environments will force algorithms to focus on human action and furniture parts while ignoring the background clutter and other distractors. Moreover, to allow human pose diversity, we ask participants to conduct assembly either on the floor or on a table work surface. This yields a total of 10 camera configurations (two per environment).

Statistics. The IKEA ASM dataset consists of 371 unique assemblies of four different furniture types (side table, coffee table, TV bench, and drawer) in three different colors (white, oak, and black). There are in total 1113 RGB videos and 371 depth videos (top view). Figure 2 shows the video and individual action length distribution. Overall, the dataset contains 3,046,977 frames (∼35.27h) of footage with an average of 2735.2 frames per video (∼1.89min).

Figure 3 shows the atomic action distribution in the train and test sets. Each action class contains at least 20 clips. Due to the nature of the assemblies, there is a high imbalance (each table assembly contains four instances of leg assembly). The dataset contains a total of 16,764 annotated

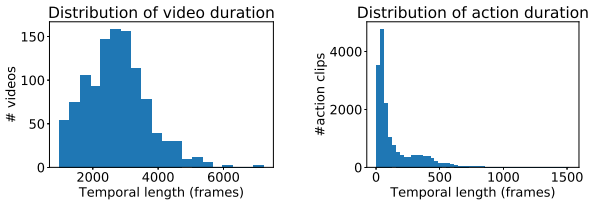

Figure 2: The duration statistics of the videos (left) and actions (right) in the IKEA assembly dataset.

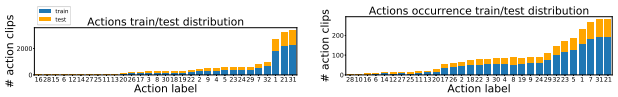

Figure 3: The IKEA ASM dataset action distribution (left) and action instance occurrence distribution (right).

actions with an average of 150 frames per action (∼6sec). For a full list of action names and ids, see supplemental.

Data split. We aim to enable model training that will generalize to previously unseen environments and human subjects. However, there is a great overlap between subjects in the different scenes and creating a split that will hold-out both simultaneously results in discarding a large portion of the data. Therefore, we propose an environment-based train/test split, i.e., test environments do not appear in the trainset and vise-versa. The trainset and testset consist of 254 and 117 scans, respectively. Here, test set includes environments 1 and 2 (family room and office). All benchmarks in Section 4 were conducted using this split. Additionally, we provide scripts to generate alternative data splits to hold out subjects, environments and joint subject-environments.

Data annotation. We annotate our dataset with temporal and spatial information using pre-selected Amazon Mechanical Turk workers to ensure quality. Temporally, we specify the boundaries (start and end frame) of all atomic actions in the video from a pre-defined set. Actions involve interaction with specific object types (e.g., table leg).

Multiple spatial annotations are provided. First, we annotate instance-level segmentation of the objects involved in the assembly. Here an enclosing polygon is drawn around each furniture part. Due to the size of the dataset, we manually annotate only 1% of the video frames which are selected as keyframes that cover diverse object poses and human poses throughout the entire video and provide pseudo ground-truth for the remainder (see §4.3). Visual inspection was used to confirm the quality of the pseudo ground-truth. For the same set of manually annotated frames, we also assign each furniture part with a unique ID, which preserves the identity of that part throughout the entire video.

We also annotated the human skeleton of the subjects involved assembly. Here, we asked workers to annotated

Method	Frame acc.			
	top 1	top 3	macro	mAP
ResNet18 [23]	27.06	55.14	21.95	11.69
ResNet50 [23]	30.38	56.1	20.03	9.47
C3D [69]	45.73	69.56	32.48	21.98
P3D [52]	**60.4**	**81.07**	**45.21**	**29.86**
I3D [8]	57.57	76.55	39.34	28.59

Table 2: Action recognition baseline frame-wise accuracy, macro-recall, and mean average precision results.

12 body joints and five key points related to the face. Due to occlusion with furniture, self-occlusions and uncommon human poses, we include a confidence value between 1 and 3 along with the annotation. Each annotation was then visually inspected and re-worked if deemed to be poor quality.

4. Experiments and benchmarks

We benchmark several state-of-the-art methods for the tasks of frame-wise action recognition, object instance segmentation and tracking, and human pose estimation.

4.1. Action recognition

We use three main metrics for evaluation. First, the frame-wise accuracy (FA) which is the de facto standard for action recognition. We compute it by counting the number of correctly classified frames and divide by the total number of frames in each video and then average over all videos in the test set. Second, since the data is highly imbalanced, we also report the macro-recall by separately computing recall for each category and then averaging. Third, we report the mean average precision (mAP) since all untrimmed videos contain multiple action labels. We compare several state-of-the-art methods for action recognition, including I3D [8], P3D ResNet [52], C3D [69], and frame-wise ResNet [23]. For each we start with a pre-trained model and fine-tune it on the IKEA ASM dataset using parameters provided in the original papers. To handle data imbalance we use a weighted random sampler where each class is weighted inversely proportional to its abundance in the dataset. Results are reported in Table 2 and show that P3D outperforms all other methods, consistent with performance on other datasets. Additionally, the results demonstrate the challenges compared to other datasets where I3D, for example, has an FA score of 57.57% compared to 68.4% on Kinetics and 63.64% on Drive&Act dataset.

4.2. Multi-view and multi-modal action recognition

We further explore the affects of multi-view and multi-modal data using the I3D method. In Table 3 we report performance on different views and different modalities. We also report their combination by averaging softmax output scores. We clearly see that combining views gives a boost in performance compared to the best single view method.

Data type	View	Frame acc.			
		top 1	top 3	macro	mAP
RGB	top view	57.57	76.55	39.34	28.59
	front view	60.75	79.3	42.67	32.73
	side view	52.16	72.21	36.59	26.76
	combined views	63.09	80.54	45.23	32.37
Human pose	HCN [36]	37.75	63.07	26.18	22.14
Human pose	ST-GCN [74]	36.99	59.63	22.77	17.63
	combined RGB+pose	64.15	80.34	46.52	32.99
Depth	top view	35.43	59.48	21.37	14.4
	combined all	63.83	81.08	44.42	31.25

Table 3: Action recognition frame-wise accuracy, macro-recall, and mean average precision results for multi-view/modal inputs.

We also find that combining views and pose gives an additional performance increase. Additionally, combining views, depth and pose in the same manner results a small disadvantage, which is due to the inferior performance of the depth based method. This suggests that exploring action recognition in the 3D domain is an open and challenging problem. The results also suggest that a combined, holistic approach that uses multi-view and multi-modal data, facilitated by our dataset, should be further investigated in future work.

4.3. Instance segmentation

As discussed in Section 3, the dataset comes with manual instance segmentation annotation for 1% of the frames (manually selected keyframes that cover diverse object poses and human poses throughout the entire video). To evaluate the performance of existing instance segmentation methods on almost texture-less IKEA furniture, we train Mask R-CNN [22] with ResNet50, ResNet101, and ResNeXt101, all with feature pyramid networks structure (FPN) on our dataset. We train each network using the implementation provided by the Detectron2 framework [72]. Table 4 shows the instance segmentation accuracy for the aforementioned baselines. As expected, the best performing model corresponds to the Mask R-CNN with ResNeXt101-FPN, outperforming ResNet101-FPN and ResNet50-FPN with 3.8% AP and 7.8% AP, respectively.

Since the manual annotation only covers 1% of the whole dataset, we propose to extract pseudo-ground-truth automatically. To this end, we train 12 different Mask R-CNNs with a ResNet50-FPN backbone to overfit on subsets of the training set that cover similar environments and furniture. We show that to achieve manual-like annotations with more accurate part boundaries, training the models with PointRend [30] as an additional head is essential. Figure 4 compares the automatically generated pseudo-ground-truth with and without the PointRend head. To evaluate the effectiveness of adding pseudo-ground-truth, we compare the Mask R-CNN trained with ResNet50-FPN with 1% annotated data (i.e., manual annotations) and 20% annotated data

Figure 4: Comparison between pseudo ground-truth without PointRend head (left) and with PointRend head (right).

(combination of manual and automatically generated annotations) illustrated in Table 5(a) and see a slight improvement. Note that any backbone architecture can benefit from the automatically generated pseudo-ground-truth.

We also investigate the contribution of adding a PointRend head to the Mask R-CNN with ResNet-50-FPN when training on 1% of manually annotated data. Table 5(b) shows that boundary refinement through point-based classification improves the overall instance segmentation performance. This table also clearly shows the effect of PointRend on estimating tighter bounding boxes.

Additionally, to evaluate the effect of furniture color and environment complexity, we report the instance segmentation results partitioned by color (Table 5(c)) and by environment (Table 5(d)). Note that, for both of these experiment we use the same model trained on all furniture colors and environments available in the training set. Table 5(c) shows that oak furniture parts are easier to segment. On the other hand, white furniture parts are the hardest to segment as they reflect the light in the scene more intensely. Another reason is that white parts might be missed due to poor contrast against the white work surfaces.

Although Mask R-CNN shows promising results in many scenarios, there are also failure cases, reflecting the real-world challenges introduced by our dataset. These failures are often due to (1) relatively high similarities between different furniture parts, e.g., front panel and rear panel of drawers illustrated in Figure 5(top row) and (2) relatively high similarities between furniture parts of interest and other parts of the environment which introduces false positives. An example of the latter can be seen in Figure 5(bottom row) where Mask R-CNN segmented part of the working surface as the shelf.

4.4. Multiple furniture part tracking

As motivated in Section 3, we utilize SORT [5] as a fast online multiple object tracking algorithm that only relies on geometric information in a class-agnostic manner. Given the detections predicted by the Mask R-CNN, SORT assigns IDs to each detected furniture part at each time-step.

To evaluate the MOT performance, we use standard metrics [56, 4]. The main metric is MOTA, which combines three error sources: false positives (FP), false negatives (FN) and identity switches (IDs). A higher MOTA score implies better performance. Another important metric is IDF1, i.e., the ratio of correctly identified detections over the average number of ground-truth and computed detections. The number of identity switches (IDs), FP and FN are also frequently reported. Furthermore, mostly tracked (MT) and mostly lost (ML), that are respectively the ratio of ground-truth trajectories that are covered/lost by the tracker for at least 80% of their respective life span, provide finer details on the performance of a tracking system. All metrics were computed using the official evaluation code provided by the MOTChallenge benchmark[1].

Figure 5: Illustration of part instance segmentation failure cases. (Top row) Mask R-CNN fails to correctly classify different panels of the drawer due to high similarity. (Bottom row) Mask R-CNN incorrectly segments part of the working surface as a furniture part (e.g., shelf) leading to considerable false positives.

Table 6 shows the performance of SORT on each test environment as well as the entire test set. The results reflect the challenges introduced by each environment in the test set. For instance, in Env1 (Family Room) provides a side view of the assembler and thus introduces many occlusions. This can be clearly seen in the number of FN. Moreover, since the tracker may lose an occluded object for a reasonably long time, it may assign new IDs after occlusion, thus affecting the mostly tracked parts and IDF1. On the other hand, the front view provided in Env2 (Office) leads to less occlusions, and thus better identity preservation reflected in IDF1 and MT. However, since the office environment contains irrelevant but similar parts, e.g., the desk partition or the work surface illustrated in Fig. 5(bottom row), we observed considerably higher FP which further affects MOTA.

4.5. Human pose

The dataset contains 2D human joint annotations in the COCO format [39] for 1% of frames, the same keyframes selected for instance segmentation, which cover a diverse range of human poses across each video. As shown in Figure 6, there are many highly challenging and unusual poses in the dataset, due to the nature of furniture assembly, particularly when performed on the floor. There are also many other factors that reduce the accuracy of pose estimation ap-

[1] https://motchallenge.net/

Feature Extractor	Annotation Type	AP	AP50	AP75	table-t	leg	shelf	side-p	front-p	bottom-p	rear-p
ResNet-50-FPN	mask	58.1	77.2	64.2	80.8	59.8	68.9	32.8	50.0	66.0	48.3
ResNet-101-FPN	mask	62.1	82.0	68.0	84.4	71.6	67.5	33.5	53.7	70.2	54.0
ResNeXt-101-FPN	mask	65.9	85.3	73.2	87.6	71.2	76.0	44.3	52.6	73.4	56.2
ResNet-50-FPN	bbox	59.5	77.7	68.9	77.3	63.5	64.7	41.0	60.1	61.8	48.5
ResNet-101-FPN	bbox	64.6	81.8	72.8	84.9	75.6	66.0	42.4	61.6	68.1	53.3
ResNeXt-101-FPN	bbox	69.5	86.4	78.9	89.4	76.8	73.7	53.3	65.8	68.7	59.0

Table 4: Evaluating the effect of backbone architecture of Mask R-CNN in furniture part instance segmentation.

(a) Influence of adding Pseudo GT

Setting	AP_{segm}	$AP50_{segm}$	$AP75_{segm}$	AP_{box}	$AP50_{box}$	$AP75_{box}$
Manual GT	58.1	77.2	64.2	59.5	77.7	68.9
Manual + Pseudo GT	60.1	77.7	66.1	62.6	77.8	69.9

(b) Influence of PointRend head

Setting	AP_{segm}	$AP50_{segm}$	$AP75_{segm}$	AP_{box}	$AP50_{box}$	$AP75_{box}$
Without PointRend	58.1	77.2	64.2	59.5	77.7	68.9
With PointRend	61.4	80.9	67.0	63.9	82.2	73.2

(c) Color-based Evaluation

Colors	AP_{segm}	$AP50_{segm}$	$AP75_{segm}$	AP_{box}	$AP50_{box}$	$AP75_{box}$
White	55.5	76.2	60.6	57.3	76.5	65.5
Black	57.8	74.8	64.4	58.5	75.4	67.6
Oak	62.9	82.1	69.5	64.5	82.3	75.8

(d) Environment-based Evaluation

Environments	AP_{segm}	$AP50_{segm}$	$AP75_{segm}$	AP_{box}	$AP50_{box}$	$AP75_{box}$
Env1 (Family Room)	47.1	63.0	53.5	49.9	65.6	58.4
Env2 (Office)	64.4	85.0	70.7	64.8	84.3	74.6

Table 5: Ablation study on furniture part instance segmentation. Note, all experiments are conducted with Mask R-CNN with ResNet-50-FPN as the backbone and tested on the same manually annotated data.

Test Env.	IDF1↑	MOTA↑	MT↑	PT	ML↓	FP↓	FN↓	IDS↓
Env1 (Family Room)	63.7	69.6	60.1	35.9	4.0	92	1152	382
Env2 (Office)	72.0	59.1	94.8	5.2	0.0	4426	681	370
All	70.0	62.1	84.1	14.6	1.2	4518	1833	752

Table 6: Evaluating the performance of SORT [5] in multiple furniture part tracking given the detections computed via MASK R-CNN with ResNeXt-101-FPN backbone.

proaches, including self-occlusions, occlusions from furniture, baggy clothing, long hair, and human distractors in the background. We also obtain pseudo-ground-truth 3D annotations by fine-tuning a Mask R-CNN [22] 2D joint detector on the labeled data, and triangulating the detections of the model from the three calibrated camera views. As a verification step, the 3D points are backprojected to 2D and are discarded if more than 30 pixels from the most confident ground-truth annotations. The reprojection error of the true and pseudo ground-truth annotations is 7.12 pixels on the train set (83% of ground-truth joints detected) and 9.14 pixels on the test set (53% of ground-truth joints detected).

To evaluate the performance of benchmark 2D human pose approaches, we perform inference with existing state-

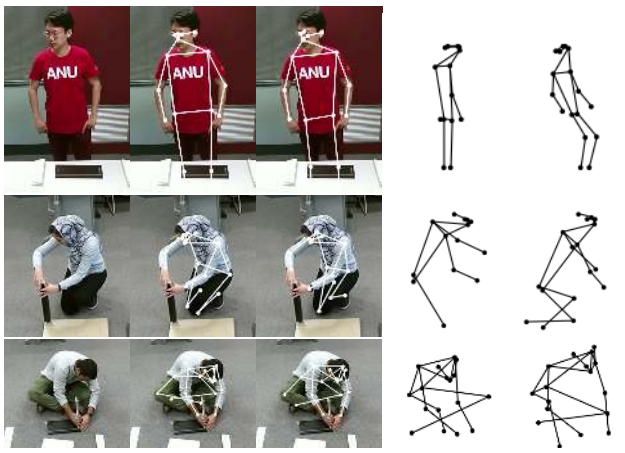

Figure 6: Qualitative human pose results. From left to right: sample image, 2D ground-truth, 2D Mask R-CNN prediction, 3D pseudo-ground-truth (novel view), and 3D VIBE prediction (novel view). The middle row shows an example where the 3D pseudo-ground-truth is incomplete, and the bottom row, shows a partial failure case for the predictions.

of-the-art models, pre-trained by the authors on the large COCO [39] and MPII [1] datasets and fine-tuned on our annotated data. We compare OpenPose [7, 6], Mask R-CNN [22] (with a ResNet-50-FPN backbone [38]), and Spatio-Temporal Affinity Fields (STAF) [53]. The first two operate on images, while the last one operates on videos, and all are multi-person pose estimation methods. We require this since our videos sometimes have multiple people in a frame with only the single assembler annotated. For fine-tuning, we trained the models for ten epochs with learning rates of 1 and 0.001 for OpenPose and Mask R-CNN, respectively. We report results with respect to the best detected person per frame, that is, the one that is closest to the ground-truth keypoints, since multiple people may be validly detected in many frames. We use standard error measures to evaluate the performance of 2D human pose methods: the 2D Mean Per Joint Position Error (MPJPE) in pixels, the Percentage of Correct Keypoints (PCK) [75], and the Area Under the Curve (AUC) as the PCK threshold varies to a maximum of 100 pixels. A joint is considered correct if it is located within a threshold of 10 pixels from the

Method	Input	Train set MPJPE↓	PCK↑	AUC↑	Test set MPJPE↓	PCK↑	AUC↑
OpenPose-pt [7]	Image	17.3	46.9	78.1	16.5	46.7	77.8
OpenPose-ft [7]	Image	11.8	57.8	87.7	13.9	52.6	85.6
MaskRCNN-pt [22]	Image	15.5	51.9	78.2	16.1	51.5	79.2
MaskRCNN-ft [22]	Image	**7.6**	**77.6**	**92.1**	**11.5**	**64.3**	**87.8**
STAF-pt [53]	Video	21.4	41.8	75.3	19.7	41.1	75.4

Table 7: 2D human pose results. The Mean Per Joint Position Error (MPJPE) in pixels and the Percentage of Correct Keypoints (PCK) @ 10 pixels (0.5% image width) are reported. Pretrained models are denoted 'pt' and models fine-tuned on the training data are denoted 'ft'.

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	All
OpenPose-pt [7]	65.0	34.6	59.7	68.2	12.4	29.1	24.7	46.7
OpenPose-ft [7]	71.0	47.9	60.4	68.8	22.7	37.7	26.3	52.6
MaskRCNN-pt [22]	72.5	39.0	65.8	74.1	11.3	33.9	26.1	51.5
MaskRCNN-ft [22]	86.5	56.4	69.4	81.9	26.9	53.4	38.0	64.3
STAF-pt [53]	54.9	35.1	55.9	61.0	11.6	23.4	18.7	41.1

Table 8: 2D human pose test set results per joint group. The Percentage of Correct Keypoints @ 10 pixels is reported.

Method	Male / Female MPJPE↓	PCK↑	Miss↓	Floor / Table MPJPE↓	PCK↑	Miss↓
OpenPose-pt [7]	15.3 / 19.2	46.6 / 46.9	0 / 1	16.9 / 15.8	47.1 / 45.9	1 / 0
OpenPose-ft [7]	13.8 / 14.0	52.4 / 53.0	0 / 6	14.3 / 13.0	52.5 / 52.8	0 / 6
MaskRCNN-pt [22]	15.6 / 17.5	51.9 / 50.5	0 / 1	16.8 / 14.8	53.1 / 48.3	0 / 1
MaskRCNN-ft [22]	11.2 / 11.9	64.6 / 63.8	0 / 0	11.4 / 11.5	65.4 / 62.3	0 / 0
STAF-pt [53]	17.6 / 24.1	40.7 / 42.1	1 / 1	19.3 / 20.3	39.3 / 44.6	1 / 1

Table 9: Evaluating the impact of gender and work surface on 2D human pose test set results. 'Miss' refers to the number of frames in which no joints were detected.

ground-truth position, which corresponds to 0.5% of the image width (1080×1920). Absolute measures in pixel space are appropriate for this dataset because the subjects are positioned at an approximately fixed distance from the camera in all scenes. In computing these metrics, only confident ground-truth annotations are used and only detected joints contribute to the mean error (for MPJPE). The results for 2D human pose baselines on the IKEA ASM train and test sets are reported in Tables 7, 8, and 9. The best performing model is the fine-tuned Mask R-CNN model, with an MPJPE of 11.5 pixels, a PCK @ 10 pixels of 64.3% and an AUC of 87.8, revealing considerable room for improvement on this challenging data. The error analysis shows that upper body joints were detected accurately more often than lower body joints, likely due to the occluding table work surface in half the videos. In addition, female subjects were detected considerably less accurately than male subjects and account for almost all entirely missed detections.

To evaluate the performance of benchmark 3D human pose approaches, we perform inference with existing state-of-the-art models, pre-trained by the authors on large 3D pose datasets, including Human Mesh and Motion Recovery (HMMR) [26], VideoPose3D (VP3D) [48], and VIBE [31]. All are video-based methods. To measure the performance of the different methods, we use the 3D Mean/median Per Joint Position Error (M/mPJPE), which computes the Euclidean distance between the estimated and ground-truth 3D joints in millimeters, averaged over all joints and frames, Procrustes Aligned (PA) M/mPJPE, where the estimated and ground-truth skeletons are rigidly aligned and scaled before evaluation, and the Percentage of Correct Keypoints (PCK) [46]. As in the Humans 3.6M dataset [24], the MPJPE measure is calculated after aligning the centroids of the 3D points in common. The PCK threshold is set to 150mm, approximately half a head. The results for 3D human pose baselines on the IKEA ASM dataset are reported in Table 10. The best performing model is VIBE, with a median Procrustes-aligned PJPE of 153mm, and a PA-PCK @ 150mm of 50%. The baseline methods perform significantly worse on our dataset than standard human pose datasets, demonstrating its difficulty. For example, OpenPose's joint detector [70] achieves a PCK of 88.5% on the MPII dataset [1], compared to 52.6% on our dataset, and VIBE has a PA-MPJPE error of 41.4mm on the H36M dataset [24], compared to 940mm on our dataset.

Method	PA	Train set MPJPE ↓ ✗	✓	mPJPE ↓ ✗	✓	PCK ↑ ✗	✓	Test set MPJPE ↓ ✗	✓	mPJPE ↓ ✗	✓	PCK ↑ ✗	✓
HMMR [26]	Vid	589	**501**	189	96	32	54	1012	951	369	196	25	40
VP3D [48]	Vid	**546**	518	**111**	87	**63**	70	**930**	**913**	212	179	**44**	47
VIBE [31]	Vid	568	517	139	**81**	55	**74**	963	940	**199**	**153**	43	**50**

Table 10: 3D human pose results. The Mean Per Joint Position Error (MPJPE) in millimeters, the median PJPE (mPJPE), and the Percentage of Correct Keypoints (PCK) @ 150mm are reported, with and without Procrustes alignment (PA). Only confident ground-truth annotations are used and only detected joints contribute to the errors.

5. Conclusion

In this paper, we introduce a large-scale comprehensively labeled furniture assembly dataset for understanding task-oriented human activities with fine-grained actions and common parts. The proposed dataset can also be used as a challenging test-bed for underlying computer vision algorithms such as textureless object segmentation/tracking and human pose estimations in multiple views. Furthermore, we report benchmark results of strong baseline methods on those tasks for ease of research comparison. Notably, since our dataset contains multi-view and multi-modal data, it enables the development and analysis of algorithms that use this data, further improving performance on these tasks. Through recognizing human actions, poses and object posi-

tions, we believe this dataset will also facilitate understanding of human-object-interactions and lay the groundwork for the perceptual understanding required for long time-scale structured activities in real-world environments.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3686–3693, 2014. 7, 8

[2] Anurag Arnab and Philip HS Torr. Pixelwise instance segmentation with a dynamically instantiated network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 441–450, 2017. 15

[3] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 941–951, 2019. 3, 15

[4] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *Journal on Image and Video Processing*, 2008:1, 2008. 6

[5] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3464–3468, 2016. 3, 6, 7, 15

[6] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 3, 7, 16

[7] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7291–7299, 2017. 3, 7, 8, 16

[8] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 1, 3, 5, 15

[9] Xinlei Chen, Ross Girshick, Kaiming He, and Piotr Dollár. Tensormask: A foundation for dense object segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2061–2069, 2019. 3, 15

[10] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7103–7112, 2018. 3, 16

[11] Yucheng Chen, Yingli Tian, and Mingyi He. Monocular human pose estimation: A survey of deep learning-based methods. *Computer Vision and Image Understanding*, page 102897, 2020. 15

[12] Peng Chu and Haibin Ling. Famnet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6172–6181, 2019. 3, 15

[13] Jifeng Dai, Kaiming He, Yi Li, Shaoqing Ren, and Jian Sun. Instance-sensitive fully convolutional networks. In *European Conference on Computer Vision*, pages 534–549. Springer, 2016. 3, 15

[14] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018. 2, 3

[15] Pradipto Das, Chenliang Xu, Richard F Doell, and Jason J Corso. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2634–2641, 2013. 2, 3

[16] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015. 15

[17] Yong Du, Yun Fu, and Liang Wang. Skeleton based action recognition with convolutional neural network. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 579–583. IEEE, 2015. 15

[18] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015. 2, 17

[19] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation.

In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2334–2343, 2017. 3, 16

[20] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6202–6211, 2019. 15

[21] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 15

[22] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3, 5, 7, 8, 15, 16

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[24] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014. 8

[25] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. 3, 4, 16

[26] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3D human dynamics from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5614–5623, 2019. 3, 4, 8, 16

[27] Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. A new representation of skeleton sequences for 3d action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3288–3297, 2017. 15

[28] Chanho Kim, Fuxin Li, and James M Rehg. Multi-object tracking with neural gating using bilinear lstm. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 200–215, 2018. 3, 15

[29] Alexander Kirillov, Evgeny Levinkov, Bjoern Andres, Bogdan Savchynskyy, and Carsten Rother. Instancecut: from edges to instances with multicut. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5008–5017, 2017. 3, 15

[30] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. *arXiv preprint arXiv:1912.08193*, 2019. 3, 5, 15

[31] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. VIBE: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5253–5263, 2020. 3, 4, 8, 16

[32] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4501–4510, 2019. 3, 16

[33] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011. 1

[34] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International Conference on Computer Vision*, pages 2556–2563. IEEE, 2011. 2

[35] Youngwoon Lee, Edward S Hu, Zhengyu Yang, Alex Yin, and Joseph J Lim. Ikea furniture assembly environment for long-horizon complex manipulation tasks. *arXiv preprint arXiv:1911.07246*, 2019. 2

[36] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. *arXiv preprint arXiv:1804.06055*, 2018. 3, 5, 15

[37] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. Fully convolutional instance-aware semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2359–2367, 2017. 3, 15

[38] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017. 7

[39] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755. Springer, 2014. 6, 7

[40] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 2

[41] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European conference on computer vision*, pages 816–833. Springer, 2016. 15

[42] Shu Liu, Jiaya Jia, Sanja Fidler, and Raquel Urtasun. Sgn: Sequential grouping networks for instance segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3496–3504, 2017. 15

[43] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):1–16, 2015. 4, 16

[44] Andrii Maksai and Pascal Fua. Eliminating exposure bias and loss-evaluation mismatch in multiple object tracking. *arXiv preprint arXiv:1811.10984*, 2018. 3, 15

[45] Manuel Martin, Alina Roitberg, Monica Haurilet, Matthias Horne, Simon Rei, Michael Voit, and Rainer Stiefelhagen. Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2019. 2, 3

[46] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3D human pose estimation in the wild using improved CNN supervision. In *2017 International Conference on 3D Vision (3DV)*, pages 506–516. IEEE, 2017. 8

[47] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Ordinal depth supervision for 3D human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7307–7316, 2018. 3, 16

[48] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3D human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7753–7762, 2019. 3, 8, 16

[49] Pedro O Pinheiro, Tsung-Yi Lin, Ronan Collobert, and Piotr Dollár. Learning to refine object segments. In *European Conference on Computer Vision*, pages 75–91. Springer, 2016. 3, 15

[50] Pedro OO Pinheiro, Ronan Collobert, and Piotr Dollár. Learning to segment object candidates. In *Advances in Neural Information Processing Systems*, pages 1990–1998, 2015. 3, 15

[51] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V Gehler, and Bernt Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4929–4937, 2016. 3, 16

[52] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2, 3, 5, 15

[53] Yaadhav Raaj, Haroon Idrees, Gines Hidalgo, and Yaser Sheikh. Efficient online multi-person 2D pose tracking with recurrent spatio-temporal affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4620–4628, 2019. 3, 7, 8, 16

[54] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 15

[55] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 3, 15

[56] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision*, pages 17–35. Springer, 2016. 6

[57] Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. A database for fine grained activity detection of cooking activities. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1194–1201. IEEE, 2012. 2, 3

[58] Marcus Rohrbach, Anna Rohrbach, Michaela Regneri, Sikandar Amin, Mykhaylo Andriluka, Manfred Pinkal, and Bernt Schiele. Recognizing fine-grained and composite activities using hand-centric features and script data. *International Journal of Computer Vision*, 119(3):346–373, 2016. 3

[59] Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Tracking the untrackable: Learning to track multiple cues with long-term dependencies. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 300–311, 2017. 15

[60] Fatemeh Saleh, Sadegh Aliakbarian, Mathieu Salzmann, and Stephen Gould. Artist: Autoregressive trajectory inpainting and scoring for tracking. *arXiv preprint arXiv:2004.07482*, 2020. 3, 15

[61] Nikolaos Sarafianos, Bogdan Boteanu, Bogdan Ionescu, and Ioannis A Kakadiaris. 3D human pose estimation: A review of the literature and analysis of covariates. *Computer Vision and Image Understanding*, 152:1–20, 2016. 15

[62] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016. 2

[63] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer, 2016. 2

[64] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014. 15

[65] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2

[66] Siyu Tang, Mykhaylo Andriluka, Bjoern Andres, and Bernt Schiele. Multiple people tracking by lifted multicut and person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3539–3548, 2017. 3, 15

[67] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1207–1216, 2019. 2, 3

[68] Sam Toyer, Anoop Cherian, Tengda Han, and Stephen Gould. Human pose forecasting via deep markov models. In *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8. IEEE, 2017. 1, 2, 3

[69] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 3, 5, 15

[70] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016. 8, 16

[71] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017. 3, 15

[72] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 5

[73] Jiarui Xu, Yue Cao, Zheng Zhang, and Han Hu. Spatial-temporal relation networks for multi-object tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3988–3998, 2019. 3, 15

[74] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*, 2018. 3, 5, 15

[75] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2878–2890, 2012. 7

[76] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702, 2015. 15

[77] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2, 3

6. Appendix

6.1. Extended related work

In this section we provide an extended summary of related work for each of the tasks presented in the paper: action recognition, human pose estimation, object instance segmentation and multi-object tracking.

Action Recognition Methods: Current action recognition architectures for video data are largely based on image-based models. These methods employ several strategies for utilizing the additional (temporal) dimension. One approach is to process the images separately using 2D CNNs and then average the classification results across the temporal domain [64]. Another approach includes using an RNN instead [76, 16]. The most recent and most prominent approach uses 3D convolutions to extract spatio-temporal features, this approach includes the convolutional 3D (C3D) method [69] which was the first to apply 3D convolutions in this context, Pseudo-3D Residual Net (P3D ResNet) [52] which leverages pretrained 2D CNNs, utilizes residual connections and simulates 3D convolutions, and the two-stream Inflated 3D ConvNet (I3D) [8] which uses an inflated inception module architecture and combines RGB and optical flow streams. Most recently, the slow-fast method [20] builds on top of the CNN and processes videos using two frame rates separately to obtain a unified representation.

Another approach for action recognition is to decouple the visual variations and use a mid-level representation like human pose (skeletons). Several different approaches were proposed to process the skeleton's complex structure. One approach is to use an LSTM [41], another approach is to use a spatial temporal graph CNN (ST-GCN) [74]. An alternative approach is to encode the skeleton joints and the temporal dynamics in a matrix and process it like an image using a CNN [17, 27]. Similarly, Hierarchical Co-occurrence Network (HCN) [36], adopts a CNN to learn skeleton features while leveraging it to learn global co-occurrence patterns.

Instance Segmentation. Early approaches to instance segmentation usually combined segment proposal classification in a two-stage framework. For instance, given a number of instance proposal, DeepMask [50] and closely related works [13, 49] learn to propose instance segment candidates which are then passed through a classifier (e.g., Fast R-CNN). These approaches are usually slower due to the architecture design and tend to be less accurate compared to one-stage counterparts [22]. To form a single stage instance segmentation Li et al. [37] merged segment proposal and object detection to form a fully convolutional instance segmentation framework. Following this trend, Mask R-CNN [22] combines binary mask prediction with Faster R-CNN, showing impressive performance compared to its prior work.

Mask R-CNN and other similar region-based approaches to instance segmentation [22] usually predict segmentation masks on a coarse grid, independent of the instance size and aspect ratio. While this leads to reasonable performance on small objects, around the size of the grid, it tends to produce coarse segmentation for instances occupying larger part of the image. To alleviate the problem of coarse segmentation of large instances, approaches have been proposed to focus on the boundaries of larger instances, e.g., through pixel grouping to form larger masks [2, 42, 29] as in InstanceCut [29], utilizing sliding windows on the boundaries or complex networks for high-resolution mask prediction as in TensorMask [9], and point-based segmentation prediction as in PointRend [30].

Multiple Object Tracking. With the advances in object detection [54, 21, 55], tracking-by-detection is now a common approach for multiple object tracking (MOT). Mostly studied in the context of multiple person tracking, MOT can be considered from different aspects. It can be categorized into online or offline, depending on when the decisions are made. In online tracking [60, 71, 3, 12, 73, 28], the tracker assigns detections to tracklets at every time-step, whereas in offline tracking [66, 44] the decision about the tracklets are made after observing the whole context of the video. Different MOT approaches can also be divided into geometry-based [60, 5], appearance-based [12, 3, 73], and a combination of appearance and geometry information with social information [44, 59]. The choice of information to represent each object highly depends on the context and scenario. For instance, for general multiple person tracking, social information and appearance information could be helpful, but, in sport scenarios, appearance information could be misleading. In our context for instance, one common application is human-robot collaboration in IKEA furniture assembly, where the tracking system should be able to make its decisions in real-time in an online fashion [60, 5]. Moreover, we know that IKEA furniture parts are almost textureless and of the same color and shape, and thus the appearance information could be misleading. Therefore, one may need to employ a completely geometry-based approach. Additionally, we know that IKEA furniture parts are rigid, non-deformable object, that are moved almost linearly in a short temporal window. Therefore, a simple, well-designed MOT that models linear motions [5] is a reasonable choice.

Human Pose Estimation. The large volume of work on human pose estimation precludes a comprehensive list; the reader is referred to two recent surveys on 2D and 3D human pose estimation [11, 61] and the references therein. Here, we will briefly discuss recent state-of-the-art approaches, including the baselines selected for our experiments. Multi-person 2D pose estimation methods can be

divided into bottom-up (predict all joints first) [51, 7, 6, 53] or top-down (detect all person bounding boxes first) [22, 19, 10] approaches, with the former reaching real-time processing speeds and the latter having better performance. OpenPose [7, 6] uses the CPM joint detector [70] to predict candidate joint heatmaps and part affinity fields, encoding limb orientation, from which the skeletons can be assembled. This was extended to incorporate temporal multi-frame information in Spatio-Temporal Affinity Fields (STAF) [53]. Mask R-CNN [22] is a notable top-down detection-based approach, where a keypoint regression head can be learned alongside the bounding box and segmentation heads. More recently, Cascade Pyramid Networks (CPN) [10] were proposed, which use multi-scale feature maps and hard keypoint mining to improve position accuracy. Monocular 3D human pose estimation methods can be categorized as being model-free [47, 48] or model-based [25, 26, 32, 31]. The former include VideoPose3D [48] which estimates 3D joints via temporal convolutions over 2D joint detections in a video sequence. The latter approach predicts the parameters of a body model, often the SMPL model [43], such as the joint angles, shape parameters, and rotation. For example, Kanazawa *et al*. [25] trained an encoder to predict the SMPL parameters, using adversarial learning to encourage realistic body poses and shapes, and later extended this to video input [26]. Instead of estimating SMPL parameters, Kolotouros *et al*. [32] directly regress the location of mesh vertices using graph convolutions. Finally, Kocabas *et al*. [31] proposed a video-based approach that uses adversarial learning to generate kinematically plausible motions. These model-based approaches tend to generalize better to unseen datasets, and so we focus on these methods as benchmarks on our dataset.

6.2. Dataset auxiliary data

In Table 11 we provide the full atmoic action list along with the action identifier, action verb, object name and a short action description. The action class ids correspond to the ids in Figure 3.

ID	Verb	Object	Description
0	-	-	No Annotation (NA)
1	align	leg	align leg screw with table thread
2	align	side panel	align side panel holes with front panel dowels
3	attach	back panel	attach drawer back panel
4	attach	side panel	attach drawer side panel
5	attach	shelf	attach shelf to table
6	flip	shelf	flip shelf
7	flip	table	flip table
8	flip	table top	flip table top
9	insert	pin	insert drawer pin
10	lay down	back panel	lay down back panel
11	lay down	bottom panel	lay down bottom panel
12	lay down	front panel	lay down front panel
13	lay down	leg	lay down leg
14	lay down	shelf	lay down shelf
15	lay down	side panel	lay down side panel
16	lay down	table top	lay down table top
17	-	-	other (unavailable action class)
18	pick up	back panel	pick up back panel
19	pick up	bottom panel	pick up bottom panel
20	pick up	front panel	pick up front panel
21	pick up	leg	pick up leg
22	pick up	pin	pick up pin
23	pick up	shelf	pick up shelf
24	pick up	side panel	pick up side panel
25	pick up	table top	pick up table top
26	position	drawer	position the drawer right side up
27	push	table	push table
28	push	table top	push table top
29	rotate	table	rotate table
30	slide	bottom panel	slide bottom of drawer
31	spin	leg	pin leg
32	tighten	leg	tighten leg

Table 11: Atomic action class id, action verb, object name and action descrtiption.

6.3. Additional results

6.3.1 Action recognition results

We provide additional visualization of the results for the baseline methods and multi-view multi-modal action recognition. In Fig. 7 and Fig. 8 we visualize the per-class accuracy results of multi-view/multi-modal and action recognition baseline methods respectively.

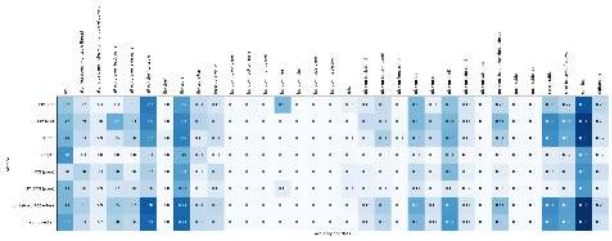

Figure 7: Action recognition accuracy for each class for multi-view/multi-modal baselines.

6.3.2 Action localization results

In this section we provide additional baseline results for the task of action localization. The goal in this task is to find and recognize all action instances within an untrimmed test video. The desired output here is a start and end frame for each action appearing in the video sequence. In order

Method	mAP @ α													
	0.1	0.2	0.3	0.4	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95
C3D	0.23	0.18	0.14	0.11	0.09	0.09	0.08	0.06	0.06	0.05	0.03	0.03	0.02	0.01
P3D	0.38	0.34	0.28	0.23	0.17	0.15	0.13	0.11	0.09	0.08	0.06	0.05	0.02	0.01
I3D	0.3	0.27	0.24	0.19	0.15	0.12	0.11	0.09	0.08	0.06	0.05	0.04	0.02	0
I3D combined views	0.44	0.39	0.33	0.27	0.2	0.18	0.16	0.13	0.12	0.1	0.09	0.06	0.03	0.01
I3D combined all	0.38	0.33	0.28	0.21	0.18	0.15	0.14	0.12	0.09	0.08	0.07	0.04	0.02	0.01

Table 12: Comparison of action localization baselines on the IKEA ASM dataset.

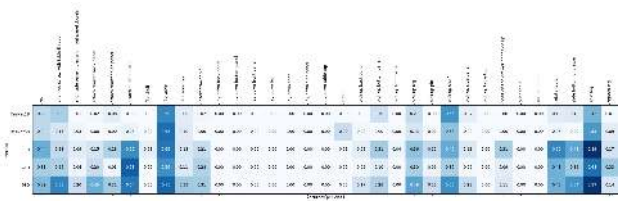

Figure 8: Action recognition accuracy for each class for the baseline methods.

to evaluate performance on this task, we follow [18] and compute the mean average precision (mAP) over all action classes. We set an example as true positive by computing the intersection over union score between the predicted and ground truth temporal segments and checking if it is greater than a threshold $\alpha \in [0.1, 1]$.