IKEA Ego 3D Dataset: Understanding furniture assembly actions from ego-view 3D Point Clouds

Yizhak Ben-Shabat[1,2] Jonathan Paul[2] Eviatar Segev[2] Oren Shrout[2] Stephen Gould[1]

[1]Australian National University [2]Technion, Israel Institute of Technology

sitzikbs@gmail.com, Stephen.Gould@anu.edu.au

https://sitzikbs.github.io/IKEAEgo3D.github.io/

Abstract

We propose a novel dataset for ego-view 3D point cloud action recognition. While there has been extensive research on understanding human actions in RGB videos in recent years, the exploration of its 3D point cloud counterpart has been relatively limited. Furthermore, RGB ego-view datasets are rapidly growing, however, 3D point cloud ego-view datasets are scarce at best. Existing 3D datasets are limited in several ways, some include actions that are distinguishable by full-body motion while others use a distant static sensor that hinders the recognition of small objects. We introduce a new point cloud action recognition dataset—the IKEA Ego 3D *dataset. It includes sequences of point clouds captured from an ego-view using a HoloLens 2 device. The dataset consists of approximately 493k frames and 56 classes of intricate furniture assembly actions of four different furniture types. We evaluate the performance of various state-of-the-art 3D action recognition methods on the proposed dataset and show that it is very challenging.*

1. Introduction

In this paper, we address the task of action recognition from ego-view 3D point cloud sequences. We introduce a novel 3D point cloud dataset of humans performing furniture assembly actions, captured from an ego-view using a HoloLens 2 device. Our research is driven by the remarkable proliferation of online media, mobile devices, surveillance cameras, and the emergence of accessible commodity 3D sensors. These technological advancements have opened up new avenues for the computer vision community to explore and develop data-driven action recognition methods [6, 16]. However, despite these advancements, the potential of the 3D point cloud modality for action recognition has remained largely untapped. This is primarily due to the limited availability of annotated 3D action data.

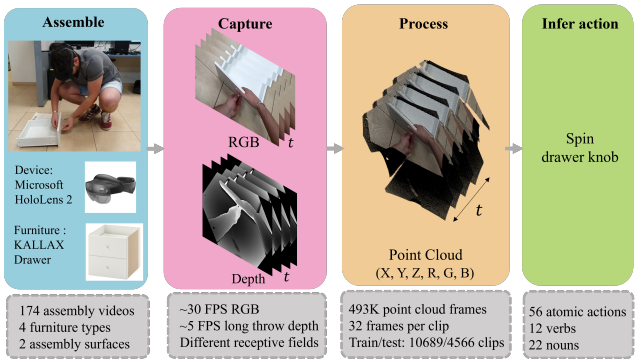


Figure 1. **Overview of the IKEA Ego 3D dataset.** The dataset includes 174 assembly sequences captured using a Microsoft HoloLens 2. The captured RGB and Depth streams are labelled with 56 action labels per frame. The data is then aligned and synced to form 3D point cloud sequences that are the input to 3D action recognition algorithms.

In recent years we have witnessed a surge in head-mounted virtual, augmented, and mixed reality devices such as the Apple Vision Pro, Microsoft HoloLens 2 and Meta Occulus Quest 3. These devices are often equipped with a suite of sensors that can produce data streams that include ego-view RGB videos and point cloud sequences (depth). Most often, the RGB video is processed to provide the user with an immersive experience. However, in many cases, a pure RGB video-based inference may not be enough and incorporating other modalities like point clouds are required. This is especially necessary for scenarios where the video is heavily degraded (e.g., due to poor lighting) or in a safety-critical application where redundancy is needed.

3D sensors offer an alternative modality through the acquisition of point clouds that sample the environment. Despite the extensive body of research on 3D vision and learning, the size of static 3D point cloud datasets remains significantly smaller compared to their RGB counterparts. This

discrepancy arises from the difficulties encountered in collecting and labelling such data. Moreover, the challenges are compounded when dealing with 3D point cloud sequence datasets, further restricting the capacity to derive meaningful representations for 3D actions. Furthermore, current 3D point cloud action recognition datasets are captured from a static camera, circumventing challenges associated with camera motion. Consequently, the exploration of effective methods for learning temporal point cloud representations remains an active and ongoing research endeavour. This pursuit is motivated by the inherent characteristics of point clouds, which lack structure and order, and present difficulties such as varying numbers of points. Particularly challenging is incorporating the temporal information into the point cloud representations, as they lack a one-to-one correspondence of points across time, unlike the well-defined pixel relationships.

We address these challenges and propose the *IKEA Ego 3D* dataset for action recognition (see Figure 1). The dataset includes 3D point cloud sequences captured from a Microsoft HoloLens 2 and consists of approximately 493k frames labelled using 56 atomic action classes composed of (verb, action)-pairs. The proposed dataset provides a unique addition to the few existing datasets that are available for this and related tasks. We introduce a benchmark and conduct extensive experiments to evaluate the performance of prominent 3D action recognition state-of-the-art methods and show that the proposed dataset is a challenging testbed for future research in the field.

The key contributions of our work are as follows:

- The collection and annotation of the IKEA Ego 3D dataset which introduces a compelling and intricate testbed for ego-view point cloud action recognition.
- A hierarchical verb-noun subdivision of actions, allowing to decouple spatio-temporal performance.
- A benchmark for ego-view 3D point cloud action recognition and evaluation of existing 3D point cloud action recognition approaches on the proposed dataset.

2. Related Work

Hereafter, we present an overview of existing research in three relevant fields: 3D action recognition datasets, learning 3D point cloud representations, and learning temporal 3D point cloud representations.

2.1. 3D action understanding datasets

A comparison between the different datasets discussed hereafter is presented in Table 1.

The availability of annotated data stands as a significant catalyst for the accomplishments observed in learning-based approaches. In the domain of 3D point cloud action recognition, standardized datasets tailored explicitly for this task are scarce and some of the existing datasets are limited in scale. This poses a constraint on the potential range of actions that can be effectively learned and recognized. Furthermore, a dataset for ego-view point cloud sequences is entirely absent in this landscape.

Small scale 3D datasets: Existing datasets include the CAD 60 and CAD 90 [21,37] datasets which contain 60 and 120 long-term activity videos of 12 and 10 classes respectively (e.g., making cereal, microwave food). This dataset provides raw RGB, skeletons, and depth data however its small scale and long-term focus limit its effectiveness.

The MSR-Action3D dataset [24] includes 10 subjects performing 20 action classes and includes a total of 567 depth map sequences composed of 23K frames, collected using a Kinect v1 device. The brevity of sequences within this dataset imposes limitations on evaluating the generalization capability of learning-based approaches. Consequently, utilizing this dataset for evaluation purposes offers only a restricted indication of overall generalizability.

The DFAUST dataset [4] offers high-resolution 4D scans capturing human subjects in motion. With 14 action categories and over 100 dynamic scans of 10 subjects, representing a balanced male-to-female ratio, and provides aligned mesh registrations. It was recently extended to the task of 3D action recognition [2] however its somewhat synthetic nature and small scale hinder the ability to evaluate methods' generalization capability.

Large scale 3D datasets: The NTU RGB+D 120 [25] and its predecessor the NTU RGB+D 60 [34] provide $\sim$114K and $\sim$56K clips containing 120 and 60 actions classes respectively (e.g., taking a selfie, take off a jacket). They provide 3D skeletons as well as three different simultaneous RGB views, depth and IR streams. Although these datasets can be regarded as large-scale, their focus on human-centric data introduces a challenge for prior-free approaches, as recent skeleton-based methods [11] have demonstrated exceptional performance. Consequently, justifying the adoption of a prior-free approach becomes increasingly arduous.

The Drive & Act dataset [27] focuses specifically on the domain of in-car driver activity and encompasses a rich variety of multi-view, multi-modal data, including IR streams, pose information, depth, and RGB data. While the actors in the dataset adhere to certain instructions, their actions do not follow a task-oriented paradigm in the conventional sense. It is important to note that due to the extensive effort involved in data collection, the dataset consists of a relatively low number of videos, totalling 30 in count with 9.6M frames and 83 action classes.

Most closely related to the proposed dataset is the IKEA ASM dataset [3] that provides 371 videos containing 33 action classes and clipped into $\sim$31K clips. This dataset offers a rich set of modalities, including three simultaneous RGB views, depth information, 2D and 3D skeletons, as

Dataset	Year	#Videos	#Frames	#cls	Activity type	Source	3D	Ego view
MPII Cooking [32]	2012	44	0.88M	65	cooking	collected	✗	✗
YouCook [10]	2013	88	-	-	cooking	YouTube	✗	✗
MPII Cooking 2 [33]	2016	273	2.88M	88	cooking	collected	✗	✗
IKEA-FA [39]	2017	101	0.41M	7	assembly	collected	✗	✗
YouCook2 [47]	2018	2000	-	89	cooking	YoutTube	✗	✗
EPIC-Kitchens [9]	2018	432	11.5M	149	cooking	collected	✗	✓
EPIC-Kitchens 100 [8]	2022	700	20M	90K	cooking	collected	✗	✓
COIN [38]	2019	$\sim$ 12K	-	180	12 domains	YouTube	✗	✗
IKEA in the Wild (IAW) [43]	2023	1005	-	-	assembly	YouTube	✗	✗
Drive&Act [27]	2019	30	9.6M	83	driving	collected	✓	✗
NTU RGB+D 60 [34]	2016	$\sim$ 57.6K	4M	60	Daily actions	collected	✓	✗
NTU RGB+D 120 [25]	2019	$\sim$ 114K	8M	120	Daily actions	collected	✓	✗
MSR-Action 3D [24]	2010	567	23K	20	gaming	collected	✓	✗
IKEA-ASM [3]	2020	371	3M	33	assembly	collected	✓	✗
Our IKEA Ego 3D	2023	174	493K	56	assembly	collected	✓	✓

Table 1. **Dataset comparison.** The proposed dataset is the first to have ego-view 3D point cloud data in an assembly context.

well as object segmentation. The dataset poses significant challenges due to frequent occlusion and the presence of highly unique assembly poses. Furthermore, the dataset exhibits an inherent class imbalance due to variations in assembly action duration and the potential for multiple repetitions within a single assembly instance. Note that, they used a static Kinect V2 sensor that is positioned at a distance from the assembler and therefore small components (indistinguishable due to resolution) were pre-installed and do not appear as action classes. This dataset was recently extended to point cloud action recognition [2].

In this work, we introduce the IKEA Ego 3D, a large-scale 3D point cloud action recognition dataset that includes fine-grained furniture assembly actions. Unlike all of the above 3D datasets, the proposed dataset is captured from an ego-view which introduces new challenges for this task.

The gap in 3D action recognition datasets. While existing 3D action recognition datasets differ in size, number of classes, and activity types, there remain several mutual gaps. (1) Static vs. dynamic camera — in current datasets, the camera is static and positioned at one or more fixed locations. In many real-world scenarios (e.g., AR, smartphones), cameras are dynamic and the extension of methods trained on static data to dynamic data is non-trivial. (2) Coarse vs. fine actions — in existing datasets, the camera is positioned at a distance from the action. This limits the focus to coarse actions. For example, NTU RGB+D [25, 34] focuses on daily actions that include a full body human, similarly, the IKEA ASM dataset [3] focuses on assembly actions but had to revert to pre-installing fasteners since they were indistinguishable by the camera (only 1–2 pixels). (3) Wearable device scenario — due to the fixed set up of the cameras, the collected data is not suitable for augmented/mixed-reality application use cases. Therefore, extending existing methods, developed on the existing datasets, to such scenarios requires further data collection and model development.

The proposed IKEA Ego 3D dataset uses ego-view that bridges these gaps. The head-mounted device provides dynamic point clouds that are able to capture fine actions due to the proximity to the sensor. The collected data contains very little of the surrounding environment which makes it closer to a wearable-device scenario.

2D action recognition datasets. For the past decade, multiple 2D action recognition datasets have been either manually collected or scraped from YouTube and annotated. These datasets have laid the groundwork for significant advances in the development of action recognition algorithms from RGB videos. Several datasets focus on cooking-related actions. These include the MPII Cooking [32] and MPII Cooking 2 [33] datasets that consist of 44 and 273 videos respectively with a frame count of 0.88M and 2.88M and include 65 and 88 action classes respectively. Also in this category are YouCook2 [47] and YouCook [10].

Other datasets focus on furniture assembly-related actions. These include the IKEA-FA [39] that includes 101 videos consisting of 0.41M frames and 7 action classes. Also in this category is the recent IKEA in the wild (IAW) dataset that proposes a new task of image-to-video alignment between assembly videos and manual images [43].

Ego-view datasets. In recent years, a tremendous effort has been made in order to collect ego-view RGB videos for action understanding. Notable are the Epic Kitchens [9] and the Epic Kitchens 100 [8] datasets that include 432 and 700 videos consisting of 11.5M and 20M frames with 149 and 90K action classes respectively. Also notable is the massive Ego4D [18] dataset that contains 3,670 hours of everyday activities in hundreds of scenarios. Although previous

datasets have made attempts to incorporate various modalities, none of them match the nature of the proposed dataset, which specifically focuses on 3D point cloud sequences.

2.2. Learning 3D point cloud representations

Point clouds pose a challenge for neural networks due to their unstructured and point-wise unordered nature. To address these challenges, several approaches have been proposed. PointNet^{++} and its predecessor PointNet [29, 30] uses permutation-invariant operators, such as pointwise MLPs and pooling layers, to aggregate features across a point set. Some approaches construct graphs or trees from the point set to impose structure [20, 35, 40]. Alternatively, the structure can be imposed using a grid of Gaussians [1] or voxels [28, 42]. Another alternative avoids the structure by using Transformers [22, 23, 45] and their attention mechanism. For a comprehensive survey of point cloud architectures please refer to Guo *et al*. [19].

Recently, various factors that can impact the training of different architectures have been investigated [17, 31]. This includes exploring data augmentation strategies and loss functions that are not specific to a particular architecture. The results of this study showed that older PointNet-based architectures [29, 30] can perform comparably to newer architectures with minor changes.

All of the above methods deal with static, single-frame, or single-shape point clouds. In this work, the input is a temporal point cloud where a representation for a short sequence is required and point correspondence between frames is unknown. We propose a benchmark for 3D action recognition and report the results of PoineNet [29], PointNet^{++} [30] and Set Transformer [22] as baselines for per-frame inference in addition to a temporally smoothed version of these methods.

2.3. Learning temporal point cloud representations

Temporal point clouds, particularly in the context of action recognition, have not received the same level of extensive investigation as their static counterparts. In Meteornet [26], a PointNet^{++} architecture is utilized by incorporating a temporal dimension appended to the spatial coordinates, enabling the processing of point cloud sequences. A spatio temporal convolution was proposed in PSTNet [14, 15] that leverages temporal consistency for action recognition. Similarly, P4Transformer [12] uses a transformer architecture and adopts 4D convolutions to capture appearance and motion using self-attention. In a subsequent work, PST-Transformer [13] explore similarities across entire videos, by introducing video-level self-attention that encodes spatio-temporal structures. MinkowskiNet [7] first converts the point cloud into an occupancy grid and then applies a 4D spatio-temporal CNN. 3DV [41] encodes 3D motion information from depth videos into a condensed voxel set. Kinet [46] implicitly encodes feature-level dynamics in feature space by unrolling the normal solver of ST-surfaces. Most recently, 3DinAction [2] proposed to extract temporal patches (t-patches) from the temporal point cloud and employ a hierarchical architecture based on MLPs to obtain a spatiotemporal representation. In this paper, we propose a benchmark for 3D action recognition and report the results for PSTNet [14], P4Transformer [12] and 3DinAction [2].

3. The IKEA Ego3D dataset

The IKEA Ego 3D dataset and code are publicly available on the project website for research purposes under the Creative Commons Attribution-NonCommercial 4.0 International License. It includes 174 sequences ($\sim$493K frames) with 56 ground truth action annotations (per frame labels). It provides point clouds with RGB color and normal vectors per point from an ego view. Additionally, we provide code for loading, processing, training, testing, evaluating, and visualizing the data.

3.1. Data collection

Our data collection hardware is a Microsoft HoloLens 2, a head-mounted mixed-reality device equipped with an array of sensors. We used the depth long-throw sensor for acquiring the depth sequences that will be later processed into 3D point cloud sequences. It provides a smoothed point cloud and a wide receptive field at a frame rate of $\sim$5 frames per second. The sensor itself is positioned a few centimetres above the eyes next to an RGB camera. This relative eye-sensor location is not fixed since the device head-mount is flexible for ergonomic reasons. This configuration poses a challenge for capturing since the assemblers have no indication of the difference between what they see and what the device is capturing. To mitigate this we projected a hologram of a thin opaque rectangular outline, guiding the person to keep the assembly within the sensors frustum. We also capture the RGB in a frame rate of $\sim$ 30 frames per second. In a post-processing stage we first sync the RGB frames to the depth frames by matching the nearest time stamps. Then, we project the RGB onto the depth map to get an RGB-D frame. Finally, in a post-processing stage we extract orineted 3D point clouds $(x, y, z, R, G, B, N_x, N_y, N_z)$ using the camera parameters. Note that the RGB receptive field is smaller, therefore not all points have a corresponding color value.

The IKEA Ego 3D dataset consists of 174 unique assemblies of four furniture types (LACK side table, LACK TV bench, KALLAX drawer, and BEKVAM stool step) see Figure 2. These furniture types were selected to partly overlap with the existing IKEA ASM dataset [3] but also to introduce a new type with a higher level of assembly complexity and more actions from the IKEA Object State dataset [36]. Note that in the IKEA ASM dataset, most fas-

(a) (b) (c) (d)

Figure 2. **IKEA Ego3D Furniture.** The assemblies in the IKEA Ego3D dataset include the (a) LACK Side table, (b) LACK TV bench, (c) KALLAX Drawer, and BEKVAM (d) Stepping stool.

teners were pre-installed, rendering the assembly process limited to coarser motions. In this dataset, however, fasteners installation is captured.

To collect the dataset we had two experts wear the HoloLens 2 on their heads while they assemble furniture in an office environment. The assembly was done on two different surfaces, a desk, and the floor. The tools and parts were scattered randomly for each assembly. The assembly was not done in a definitive sequential order, however, the assembly process is simple and therefore there is some repetitiveness in the assembly order. In Figure 3 we provide visualizations of the RGB frame, its corresponding point cloud, and the performed action. It shows the difference in lighting, assembly surfaces, tools, and action diversity.

3.2. Data annotations

We manually annotate our dataset using an open-source video annotation tool (Vidat) [44]. Annotators were asked to specify the temporal boundaries, *i.e.*, start and end frame, of all actions in the video from a pre-defined list of atomic actions. Each atomic action in the action list is composed of (verb, noun)-pair, where the noun is either an assembly component or a tool and the verb specifies the manner of using it. For visualization simplicity, the labelling is performed on the RGB video and then synchronized with the corresponding point clouds frame.

3.3. Data statistics

Overall, the dataset contains 493568 frames of footage with an average of 2836.6 frames per video (frames captured at $\sim$ 30 FPS for RGB and $\sim$ 5 for depth) and an average action duration of 46.2 (with a standard deviation of 64.4 frames). We captured 52 sequences for the Drawer assembly, 50 for the Side Table , 43 for the Stool, and 29 for the Coffee Table. The assemblies were done on two different surfaces - floor and table that include 52 and 122 sequences respectively. Figure 5 shows the distribution of sequence lengths (a) and action lengths (b) in terms of the number of frames. The statistics in the context of learning are presented in Figure 4 which shows the action distribution in the train and test sets. Each action class contains at least 100 frames. Due to the nature of the assemblies, there is a high action imbalance since some actions are longer and are repeated multiple times in an assembly and others are short and occur once e.g., *spin leg* (class 53) repeats four times when assembling a table and takes much longer than *lay down screwdriver* (class 26). Additionally, there are many unlabeled frames (labelled *NA*, class 0) since in an assembly task there are multiple transitions between different actions. For a full list of actions see the supplementary.

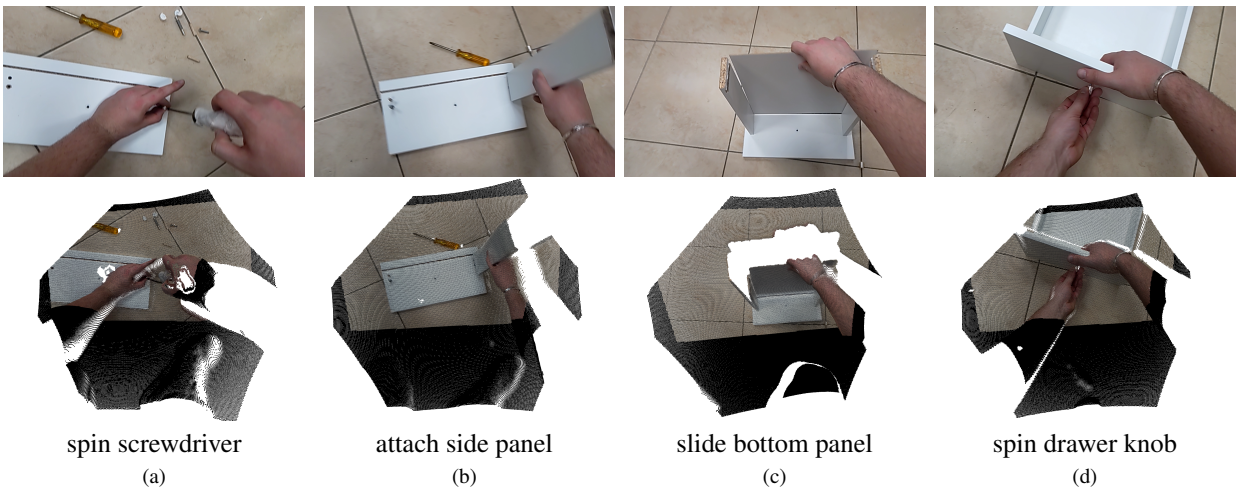
In addition to the atomic actions, we introduce another level of hierarchy that subdivides the actions into nouns and verbs. The verb set captures temporal semantics since different objects are used in a similar manner. These classes are characterized by geometric differences but temporal similarities. The verb label set includes 12 classes: *align, attach, flip, insert, lay down, spin, move, pick up, slide, move, interface, NA*. On the other hand, the noun set captures objects (assembly components or tools). These classes are characterized by geometric similarities but temporal differences. Here, we group similar components under the same class, e.g., different screw objects are now a single screw class. The noun set includes 22 classes: *coffee table shelf, screw, connector, leg, beam, step, back panel, side panel, stool side, coffee table top, coffee table, stool, table top, drawer, cam lock, dowel, drill, screwdriver, bottom panel, knob, front panel, NA*.

3.4. Data split

We aim to enable model training that will generalize to previously unseen environments (tools and part locations). We split the data into 121 and 53 sequences in the train and test sets respectively. We then further subdivide the sequences into 32 frame clips yielding 10689 and 4566 clips in the train and test sets respectively ($\sim$ 340K and $\sim$144K frames respectively). The subdivision was made in a way that maintains diversity so that all action classes are available in both train and test splits.

3.5. Dataset unique challanges

The proposed dataset provides unique challenges compared to existing counterparts at the data level and on the class level. First, the ego-view includes camera motion that, for point clouds, is very challenging since points are appearing and disappearing within each frame. Additionally, this setup introduces several types of point motion: (a) action-related motion, (b) noise-related motion, and (c) camera-related motion. All are difficult to distinguish between one another. In existing datasets (e.g., IKEA ASM, NTU), only (a) and (b) point motions are available. Second, some classes are very similar to each other. Since our actions are composed of a verb + noun pair, some nouns are contextually different but visually similar e.g., screws of different lengths for different furniture types. The ego-view plays a key role in enabling the discrimination of such small components. Conversely, in the IKEA ASM dataset,



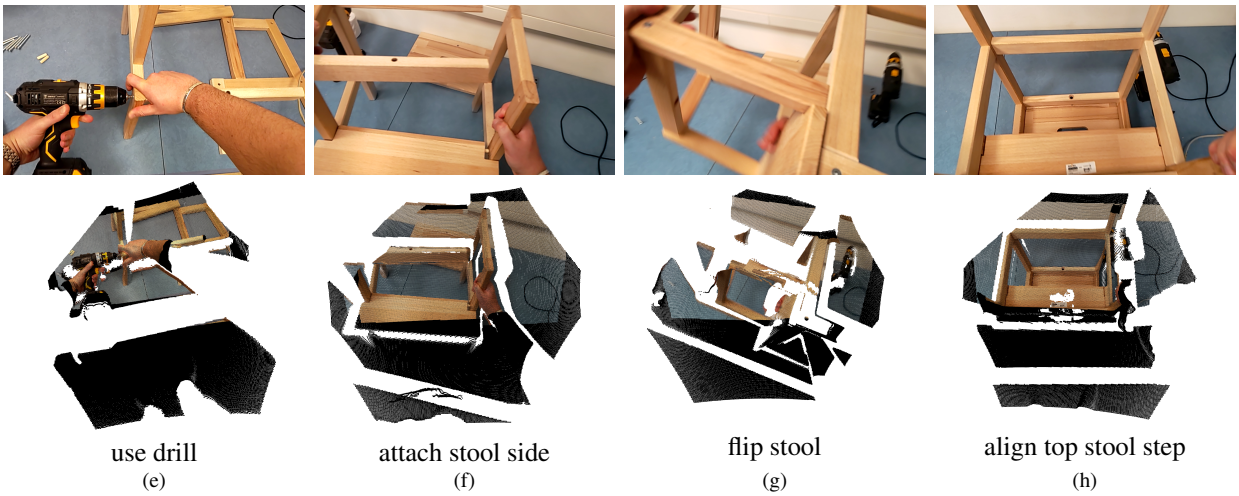


Figure 3. **IKEA Ego 3D Dataset actions.** Visualizing the RGB image (top), 3D point cloud (middle), and action label (bottom) for the Drawer (a-d) and Stool (e-h) assemblies.

these fasteners were pre-installed since the static camera was too distant to distinguish such small components. Finally, the dataset provides many class-level challenges that include ambiguity in the start and end time of an action (when does *align leg* to the table end and *spin leg* starts?), and class imbalance due to multi-occurrence actions and action duration e.g., *spin leg* action is very long and repeats four times in each table assembly while *spin drawer knob* is very short and occurs once. All of the aforementioned challenges make this dataset important for the development of future 3D point cloud-based action recognition methods.

4. Benchmark and Experiments

We report the performance of prominent 3D action understanding methods on the proposed IKEA Ego 3D datasets. The results show that IKEA Ego 3D is a challenging dataset that sheds new light on the strengths and weaknesses of existing 3D action recognition methods.

Evaluation metrics. For evaluation, we report several standard [5] metrics: the top1 and top3 frame-wise accuracy that are the de facto standard for action classification. We compute it by summing the number of correctly classified frames and dividing by the total number of frames in each video and then averaging over all videos in the test set. Additionally, since the dataset is imbalanced, we also report the macro-recall by separately computing recall for each category and then performing averaging (macro). Finally, we report the mean average precision (mAP) since all untrimmed videos contain multiple action labels.

Baselines. As a first sanity check baseline we report PointNet [29], PointNet^{++} [30], and Set Transformer [22] on each point cloud frame. These methods were not designed for temporal understanding but provided a per-point cloud frame global representation. To incorporate the temporal information in the most naive way, we implemented a temporal smoothing version of each

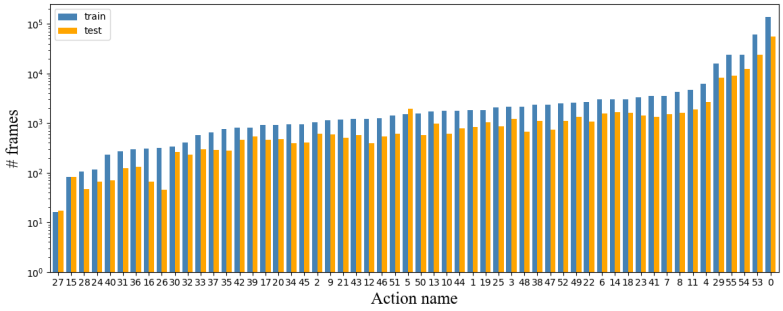
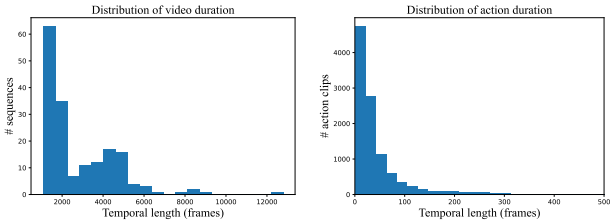


Figure 4. **IKEA Ego 3D dataset action occurrence**. A highly imbalanced dataset provides a challenge for learning-based algorithms. Note the y-axis is log scaled, therefore a small gap in this axis reflects a significant gap in proportions.



(a) (b)

Figure 5. **Duration statistics.** Sequence (a) and action (b) duration distributions in the IKEA Ego 3D dataset.

(PoinNet+TS, Pointnet^{++}+TS, and Set Transformer+TS respectively) by learning the weights of a convolutional layer over the temporal dimension. Temporal smoothing aims to provide a simple baseline for utilizing temporal information in addition to spatial information. As the second and important baselines, we report current SoTA methods for 3D action recognition that fuse between the geometric and temporal information. These include PSTNet [14], P4Transformer [12], and 3DinAction [2].

4.1. Experiment setup

We specify a unified protocol for all baselines.

Preprocessing. Every point cloud frame contains a varying number of points. Additionally, the number of points in each frame can be very large, ranging from tens to hundreds of thousands of points (scene dependent). Therefore, loading such large sequences is very computationally demanding. To alleviate these issues, we first sample 4096 points using farthest point sampling (FPS) for each frame. We then subdivide the dataset into short clips of 32 frames and save the data in a compact format (pickle) that can load a full clip instead of individual point clouds.

Training. We train each baseline for 100 epochs. To mitigate the class imbalance, we follow [3] and use a weighted random sampler where each class is weighted inversely proportional to its abundance in the dataset. We use an effective batch size of 160 (accumulating 20 batches of size 8) We utilize an Adam optimizer with an initial learning rate of 10^{-3} and a learning rate scheduler that reduces the learning rate by 50% every 25 epochs.

4.2. Benchmark results discussion

The benchmark's results are reported in Table 2. The results clearly show that there is a significant boost in performance when incorporating the naive temporal smoothing baselines into all of the reported per-frame approaches. This strengthens the notion that temporal information is crucial for inferring actions. Furthermore, among the designated action recognition methods, PSTNet [14] performance stands out despite 3DinAction [2] reporting outperforming it on other datasets (IKEA ASM and DFAUST). We attribute this to the limitations of the t-patch construction method. In our dataset, there are large motions that do not originate from the action but rather from the head (sensor) motion. This will cause many t-patches to collapse despite the bi-directional solution proposed in [2]. Surprisingly, the best-performing method is PointNet^{++} with temporal smoothing. This result suggests that the global representation in the designated 3D action recognition methods could be improved further.

4.3. Noun and Verb clustering experiment

The goal of this experiment is to decouple the temporal and spatial representation power of each method. In this experiment, we extend and decompose the above benchmark by clustering the atomic actions into 12 verb classes and 22 noun classes. We report the top1 accuracy, macro and mAP metrics. Note that the top3 metric is not valid here because all of three top predictions may belong to the same cluster. We emphasize that the baselines are not retrained (since this will likely result in an improvement independent of the desired decomposition) but rather their results are clustered according to the hierarchical noun and verb classes to provide a lower bound.

The results for the noun classes are reported in Table 3 and show that, as expected, all methods benefit from

Method	Frame acc.			
	top 1	top 3	macro	mAP
PN [29]	42.97	71.86	21.63	0.2988
PN^{++} [30]	52.08	78.65	22.83	0.3795
Set Trans. [22]	42.04	70.57	16.13	0.2141
PN [29]+ TS	50.31	76.97	21.17	0.2932
PN^{++} [30] + TS	**52.98**	**81.22**	**26.72**	0.3984
Set Trans. [22] + TS	44.80	73.30	21.39	0.2944
PSTNet [14]	50.91	76.54	22.87	**0.4024**
P4Transformer [12]	42.21	66.71	12.33	0.2025
3DinAction [2]	48.71	75.98	19.89	0.3591

Table 2. **Action recognition results on IKEA Ego 3D.** Comparing between recent state-of-the-art approaches using frame accuracy (top1 and top3), macro recall and mAP metrics.

Method	Acc.	macro	mAP
PN [29]	49.08	35.39	0.3343
PN^{++} [30]	57.33	38.21	**0.4811**
Set Transformer [22]	45.17	28.12	0.2681
PN [29]+TS	55.34	35.88	0.3510
PN^{++} [30]+TS	**59.08**	**44.58**	0.4707
Set Transformer [22]+ TS	50.06	35.48	0.3442
PSTNet [14]	56.00	37.47	0.4725
P4Transformer [12]	44.34	22.95	0.3333
3DinAction [2]	54.21	35.15	0.4375

Table 3. **Noun recognition results.** We cluster the action classes by nouns and report the performance of state-of-the-art approaches. This experiment demonstrates the spatial quality of the methods (how well they capture objects).

noun clustering. Additionally, the per-frame methods have demonstrated a larger boost in performance. This can be attributed to their ability to obtain a better global geometric representation since there are no temporal parameters to optimize. This supports the notion that arose in the benchmark that the global geometric representation of designated 3D action recognition methods could be improved further.

The results for the verb classes are reported in Table 4. The results show that all methods benefit from verb clustering in most metrics. Despite the clustering, these results are consistent with the atomic action (non-clustered) benchmark and demonstrate that, as expected, the temporal information is beneficial for recognizing the verbs. Note that verbs are inherently more difficult to recognize than nouns because verbs must accumulate sequence information, whereas nouns can be recognized by static frame. The key takeaway from this experiment is that the temporal representation power across all method can be further improved compared to its spatial counterpart.

Method	Acc.	macro	mAP
PN [29]	44.67	28.30	0.2874
PN^{++} [30]	52.71	29.44	0.3777
Set Transformer [22]	42.92	22.43	0.2245
PN [29]+TS	51.64	27.44	0.3106
PN^{++} [30]+TS	**53.66**	**33.04**	0.3683
Set Transformer [22]+ TS	46.81	28.27	0.3263
PSTNet [14]	51.77	29.27	**0.4212**
P4Transformer [12]	43.08	17.85	0.2414
3DinAction [2]	50.49	26.41	0.3671

Table 4. **Verb recognition results.** We cluster the action classes by verbs and report the performance of state-of-the-art approaches. This experiment demonstrates the temporal quality of the methods (how well they distinguish motions).

5. Future work and applications

In this paper, we focused on using the IKEA Ego 3D dataset for the task of 3D action recognition. This dataset, however, enables many future research directions and real-world applications. Future research may focus on comparing and fusing between modalities provided in the dataset (point clouds and RGB). Other directions include, for example, focusing on action anticipation and forecasting. For real-world applications, our dataset can be used in developing AR human assistive systems for assembly, and instructional tasks more generally, that provides online visual feedback on different assembly tasks in a factory or home environment, aiding in safe execution of the task and the prevention of errors.

6. Conclusions

We have introduced a large-scale annotated dataset for understanding fine-grained human actions from 3D point clouds captured from an ego viewpoint. Our dataset provides a challenging testbed for 3D computer vision algorithms, focusing on action recognition when there is sensor motion in addition to the action movements. Furthermore, we have reported benchmark results of prominent baseline methods on the task of 3D action recognition with an insightful verb and noun action decomposition. Through recognizing human actions, we believe that our dataset will facilitate an understanding of action temporal and geometrical consistency of human-object interactions and lay the groundwork for the perceptual understanding required for lengthy activities in real-world environments.

Acknowledgement. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 893465. We also thank the NVIDIA Academic Hardware Grant Program for providing high-speed A5000 GPU.

References

[1] Yizhak Ben-Shabat, Michael Lindenbaum, and Anath Fischer. 3DMFV: Three-dimensional point cloud classification in real-time using convolutional neural networks. *RAL*, 3:3145–3152, 2018. 4

[2] Yizhak Ben-Shabat, Oren Shrout, and Stephen Gould. 3dinaction: Understanding human actions in 3d point clouds. *arXiv preprint arXiv:2303.06346*, 2023. 2, 3, 4, 7, 8

[3] Yizhak Ben-Shabat, Xin Yu, Fatemeh Saleh, Dylan Campbell, Cristian Rodriguez-Opazo, Hongdong Li, and Stephen Gould. The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 847–859, 2021. 2, 3, 4, 7

[4] Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Dynamic FAUST: Registering human bodies in motion. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[5] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, 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. 6

[6] 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

[7] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3075–3084, 2019. 4

[8] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pages 1–23, 2022. 3

[9] 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. 3

[10] 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. 3

[11] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2969–2978, 2022. 2

[12] Hehe Fan, Yi Yang, and Mohan Kankanhalli. Point 4d transformer networks for spatio-temporal modeling in point cloud videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14204–14213, 2021. 4, 7, 8

[13] Hehe Fan, Yi Yang, and Mohan Kankanhalli. Point spatio-temporal transformer networks for point cloud video modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2181–2192, 2022. 4

[14] Hehe Fan, Xin Yu, Yuhang Ding, Yi Yang, and Mohan Kankanhalli. Pstnet: Point spatio-temporal convolution on point cloud sequences. In *International Conference on Learning Representations*, 2021. 4, 7, 8

[15] Hehe Fan, Xin Yu, Yi Yang, and Mohan Kankanhalli. Deep hierarchical representation of point cloud videos via spatio-temporal decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9918–9930, 2021. 4

[16] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 1

[17] Ankit Goyal, Hei Law, Bowei Liu, Alejandro Newell, and Jia Deng. Revisiting point cloud shape classification with a simple and effective baseline. In *International Conference on Machine Learning*, pages 3809–3820. PMLR, 2021. 4

[18] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 3

[19] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3d point clouds: A survey. *PAMI*, 2020. 4

[20] Roman Klokov and Victor Lempitsky. Escape from cells: Deep kd-networks for the recognition of 3d point cloud models. In *Proceedings of the IEEE international conference on computer vision*, pages 863–872, 2017. 4

[21] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. Learning human activities and object affordances from rgb-d videos. *The International Journal of Robotics Research*, 32(8):951–970, 2013. 2

[22] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pages 3744–3753. PMLR, 2019. 4, 6, 8

[23] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pages 3744–3753. PMLR, 2019. 4

[24] Wanqing Li, Zhengyou Zhang, and Zicheng Liu. Action recognition based on a bag of 3d points. In *2010 IEEE computer society conference on computer vision and pattern recognition-workshops*, pages 9–14. IEEE, 2010. 2, 3

[25] 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, 42(10):2684–2701, 2019. 2, 3

[26] Xingyu Liu, Mengyuan Yan, and Jeannette Bohg. Meteornet: Deep learning on dynamic 3d point cloud sequences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9246–9255, 2019. 4

[27] Manuel Martin, Alina Roitberg, Monica Haurilet, Matthias Horne, Simon Reiß, 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

[28] Daniel Maturana and Sebastian Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 922–928. IEEE, 2015. 4

[29] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, pages 652–660, 2017. 4, 6, 8

[30] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NeurIPS*, volume 30, 2017. 4, 6, 8

[31] Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Abed Al Kader Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. *arXiv preprint arXiv:2206.04670*, 2022. 4

[32] 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. 3

[33] 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

[34] 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, 3

[35] Yiru Shen, Chen Feng, Yaoqing Yang, and Dong Tian. Mining point cloud local structures by kernel correlation and graph pooling. In *CVPR*, pages 4548–4557, 2018. 4

[36] Yongzhi Su, Mingxin Liu, Jason Rambach, Antonia Pehrson, Anton Berg, and Didier Stricker. Ikea object state dataset: A 6dof object pose estimation dataset and benchmark for multi-state assembly objects, 2021. 4

[37] Jaeyong Sung, Colin Ponce, Bart Selman, and Ashutosh Saxena. Unstructured human activity detection from rgbd images. In *2012 IEEE international conference on robotics and automation*, pages 842–849. IEEE, 2012. 2

[38] 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. 3

[39] 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. 3

[40] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)*, 38:1–12, 2019. 4

[41] Yancheng Wang, Yang Xiao, Fu Xiong, Wenxiang Jiang, Zhiguo Cao, Joey Tianyi Zhou, and Junsong Yuan. 3dv: 3d dynamic voxel for action recognition in depth video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 511–520, 2020. 4

[42] Cheng Zhang, Haocheng Wan, Shengqiang Liu, Xinyi Shen, and Zizhao Wu. Pvt: Point-voxel transformer for 3d deep learning. *arXiv preprint arXiv:2108.06076*, 2021. 4

[43] Jiahao Zhang, Anoop Cherian, Yanbin Liu, Yizhak Ben-Shabat, Cristian Rodriguez, and Stephen Gould. Aligning step-by-step instructional diagrams to video demonstrations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2483–2492, 2023. 3

[44] Jiahao Zhang, Stephen Gould, and Itzik Ben-Shabat. Vidat—ANU CVML video annotation tool. https://github.com/anucvml/vidat, 2020. 5

[45] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip H.S. Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16259–16268, October 2021. 4

[46] Jia-Xing Zhong, Kaichen Zhou, Qingyong Hu, Bing Wang, Niki Trigoni, and Andrew Markham. No pain, big gain: classify dynamic point cloud sequences with static models by fitting feature-level space-time surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8510–8520, 2022. 4

[47] 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. 3