Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives

Kristen Grauman[1,2], Andrew Westbury[1], Lorenzo Torresani[1], Kris Kitani[1,3], Jitendra Malik[1,4], Triantafyllos Afouras[*1], Kumar Ashutosh[*1,2], Vijay Baiyya[*5], Siddhant Bansal[*6,7], Bikram Boote[*8], Eugene Byrne[*1,9], Zach Chavis[*10], Joya Chen[*11], Feng Cheng[*1], Fu-Jen Chu[*1], Sean Crane[*9], Avijit Dasgupta[*7], Jing Dong[*5], Maria Escobar[*12], Cristhian Forigua[*12], Abrham Gebreselasie[*9], Sanjay Haresh[*13], Jing Huang[*1], Md Mohaiminul Islam[*14], Suyog Jain[*1], Rawal Khirodkar[*9], Devansh Kukreja[*1], Kevin J Liang[*1], Jia-Wei Liu[*11], Sagnik Majumder[*1,2], Yongsen Mao[*13], Miguel Martin[*1], Effrosyni Mavroudi[*1], Tushar Nagarajan[*1], Francesco Ragusa[*15], Santhosh Kumar Ramakrishnan[*2], Luigi Seminara[*15], Arjun Somayazulu[*2], Yale Song[*1], Shan Su[*16], Zihui Xue[*1,2], Edward Zhang[*16], Jinxu Zhang[*16], Angela Castillo[12], Changan Chen[2], Xinzhu Fu[11], Ryosuke Furuta[17], Cristina González[12], Prince Gupta[5], Jiabo Hu[18], Yifei Huang[17], Yiming Huang[16], Weslie Khoo[19], Anush Kumar[10], Robert Kuo[18], Sach Lakhavani[5], Miao Liu[18], Mi Luo[2], Zhengyi Luo[3], Brighid Meredith[18], Austin Miller[18], Oluwatumininu Oguntola[14], Xiaqing Pan[5], Penny Peng[18], Shraman Pramanick[20], Merey Ramazanova[21], Fiona Ryan[22], Wei Shan[14], Kiran Somasundaram[5], Chenan Song[11], Audrey Southerland[22], Masatoshi Tateno[17], Huiyu Wang[1], Yuchen Wang[19], Takuma Yagi[17], Mingfei Yan[5], Xitong Yang[1], Zecheng Yu[17], Shengxin Cindy Zha[18], Chen Zhao[21], Ziwei Zhao[19], Zhifan Zhu[6], Jeff Zhuo[14], Pablo Arbeláez[†12], Gedas Bertasius[†14], Dima Damen[†6], Jakob Engel[†5], Giovanni Maria Farinella[†15], Antonino Furnari[†15], Bernard Ghanem[†21], Judy Hoffman[†22], C. V. Jawahar[†7], Richard Newcombe[†5], Hyun Soo Park[†10], James M. Rehg[†8], Yoichi Sato[†17], Manolis Savva[†13], Jianbo Shi[†16], Mike Zheng Shou[†11], and Michael Wray[†6]

[1]*FAIR, Meta,* [2]*University of Texas at Austin,* [3]*Carnegie Mellon University,* [4]*University of California, Berkeley,* [5]*Project Aria, Meta,* [6]*University of Bristol,* [7]*International Institute of Information Technology, Hyderabad,* [8]*University of Illinois, Urbana Champaign,* [9]*Carnegie Mellon University,* [10]*University of Minnesota,* [11]*National University of Singapore,* [12]*Universidad de los Andes,* [13]*Simon Fraser University,* [14]*University of North Carolina, Chapel Hill,* [15]*University of Catania,* [16]*University of Pennsylvania,* [17]*University of Tokyo,* [18]*Meta,* [19]*Indiana University,* [20]*Johns Hopkins University,* [21]*King Abdullah University of Science and Technology,* [22]*Georgia Tech*

Abstract

We present Ego-Exo4D, a diverse, large-scale multimodal multiview video dataset and benchmark challenge. Ego-Exo4D centers around simultaneously-captured egocentric and exocentric video of skilled human activities (e.g., sports, music, dance, bike repair). 740 participants from 13 cities worldwide performed these activities in 123 different natural scene contexts, yielding long-form captures from 1 to 42 minutes each and 1,286 hours of video combined. The multimodal nature of the dataset is unprecedented: the video is accompanied by multichannel audio, eye gaze, 3D point clouds, camera poses, IMU, and multiple paired language descriptions—including a novel "expert commentary" done by coaches and teachers and tailored to the skilled-activity domain. To push the frontier of first-person video understanding of skilled human activity, we also present a suite of benchmark tasks and their annotations, including fine-grained activity understanding, proficiency estimation, cross-view translation, and 3D hand/body pose. All resources are open sourced to fuel new research in the community.

1. Introduction

A dancer leaps across a stage; Lionel Messi delivers a precise pass; your grandmother prepares her famous dumplings. We observe and seek human skills in a myriad of settings, from the practical (fixing a bike) to the aspirational (dancing beautifully). What would it mean for AI to understand human skills? And what would it take to get there?

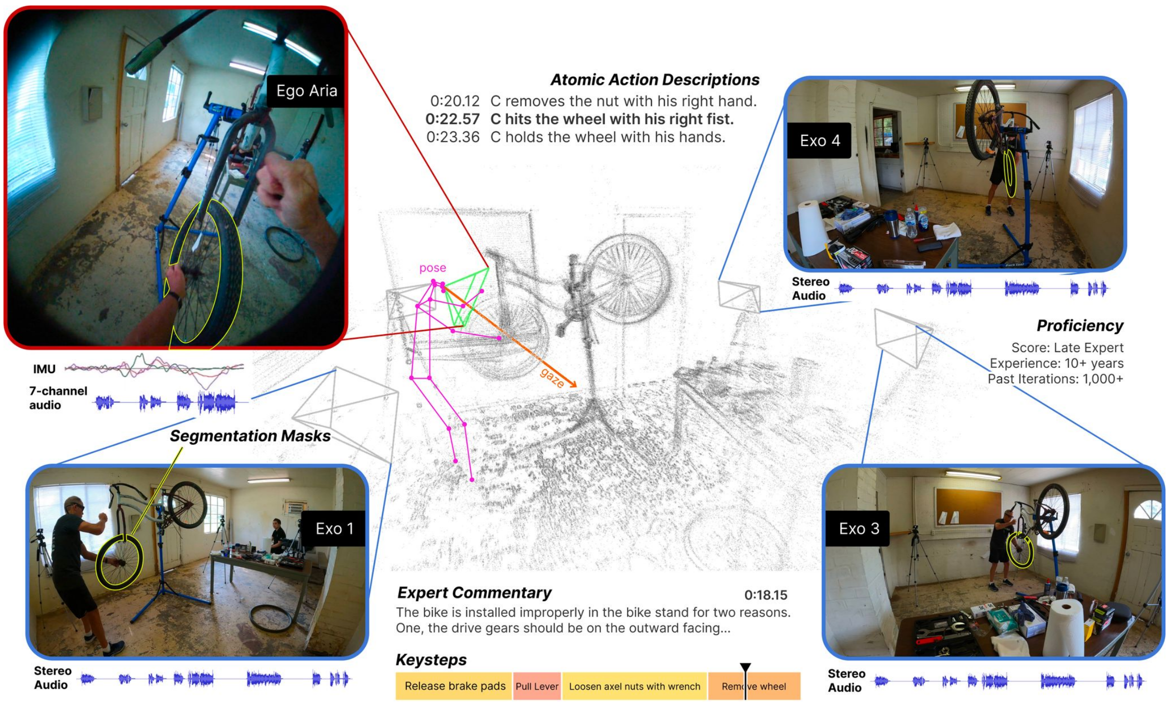

Figure 1. Ego-Exo4D offers egocentric video alongside multiple time-synchronized exocentric video streams for an array of skilled human activities—1,286 hours of ego and exo video in total. The data is both multiview and multimodal, and it is extensively annotated with language, 3D body and hand pose, keysteps, procedural dependencies, and proficiency ratings in support of our proposed benchmark tasks.

Advances in AI understanding of human skill could facilitate many applications. In augmented reality (AR), a person wearing smart glasses could quickly pick up new skills with a virtual AI coach that provides real-time guidance. In robot learning, a robot watching people in its environment could acquire new dexterous manipulation skills with less physical experience. In social networks, new communities could form based on how people share their expertise and complementary skills in video.

We contend that both the *egocentric* and *exocentric* viewpoints are critical for capturing human skill. Firstly, the two viewpoints are synergistic. The first-person (ego) perspective captures the details of close-by hand-object interactions and the camera wearer's attention, whereas the third-person (exo) perspective captures the full body pose and surrounding environment context. See Figure 1. Not coincidentally, instructional or "how-to" videos often alternate between a third-person view of the demonstrator and a close-up view of their near-field demonstration. For example, a chef may describe their approach and the equipment from an exo view, then cut to clips showing their hands manipulating the ingredients and tools from an ego-like view.

Secondly, not only are the ego and exo viewpoints synergistic, but there is a need to *translate* fluently from one to the other when acquiring skill. For example, imagine watching an expert repair a bike tire, juggle a soccer ball, or fold an origami swan—then mapping their steps to your own body. Cognitive science tells us that even from a very young age we can observe others' behavior (exo) and map it onto our own (ego) [42, 108], and this actor-observer translation remains the foundation of visual learning.

Realizing this potential, however, is not possible using today's datasets and learning paradigms. Existing datasets comprised of both ego and exo views (i.e., ego-exo) are few [76, 77, 127, 139, 145], small in scale, lack synchronization across cameras, and/or are too staged or curated to be resilient to the diversity of the real world. Thus the current literature for activity understanding primarily attends to *either* the ego [28, 47] or exo [48, 67, 105, 149] view, leaving the ability to move fluidly between the first- and third-person perspectives out of reach. Instructional video datasets [103, 159, 204, 207] offer a compelling window into skilled human activity, but (like the above) are limited to single-viewpoint video, whether purely exocentric or mixed with "ego-like" views at certain time points.

We introduce Ego-Exo4D, a foundational dataset to support research on ego-exo video learning and multimodal perception. The result of a two-year effort by a consortium of 15 research institutions, Ego-Exo4D is a first-of-its-kind large-scale multimodal multiview dataset and benchmark suite. It constitutes the largest public dataset of time-synchronized first- and third- person video, captured by 740 diverse camera wearers in 123 distinct scenes and 13 cities worldwide. For every sequence, Ego-Exo4D provides both the camera wearer's egocentric video, as well as *multiple* (4-5) exocentric videos from tripods placed around the camera wearer. All views are time-synchronized and precisely localized in a metric, gravity-aligned frame of reference. The total collection has 1,286 hours of video and 5,035 instances, each spanning 1 to 42 min. of continuous capture.

Ego-Exo4D focuses on skilled single-person activities. The 740 participants perform skilled physical and/or procedural activities—dance, soccer, basketball, bouldering, music, cooking, bike repair, health care—in an unscripted manner and in natural settings (e.g., gym, soccer field, kitchens, bike shops, etc.), exhibiting a variety of skill levels from novice to expert. All video is recorded with rigorous privacy and ethics policies and formal consent of participants.

Ego-Exo4D is not only multiview, it is also multimodal. Captured with the unique open-source Aria glasses [38], all ego video is accompanied by 7-channel audio, IMU, eye gaze, both RGB and two grayscale SLAM cameras, and 3D environment point clouds. Additionally, Ego-Exo4D provides multiple new video-language resources, all time indexed: first-person narrations by the camera wearers describing their own actions; third-person play-by-play descriptions of every camera wearer action; and third-person spoken expert commentary critiquing their performance. The latter is particularly novel: performed by domain-specific experienced coaches and teachers, it focuses on *how* an activity is executed rather than merely *what* is being done, surfacing subtleties in skilled execution not perceivable by the untrained eye. To our knowledge, there is no prior video resource with such extensive and high quality multimodal data.

Alongside this data, we introduce benchmarks for foundational tasks for ego-exo video. We propose four families of tasks: 1) *ego-exo relation*, for relating the actions of a teacher (exo) to a learner (ego) by estimating semantic correspondences and translating viewpoints; 2) *ego(-exo) recognition*, for recognizing fine-grained keysteps and task structure; 3) *ego(-exo) proficiency estimation*, for inferring how well a person is executing a skill; and 4) *ego pose*, for recovering skilled 3D body and hand movements from ego-video. We provide annotations for each task—the result of more than 200,000 hours of annotator effort. To kickstart work in these new challenges, we also develop baseline models and report their results (Appendices). We are hosting the first public benchmark challenges in 2024.

In summary, Ego-Exo4D is the community's first diverse, large-scale multimodal multiview video resource. We have open sourced all the data, annotations, camera rig protocol, and benchmarks. With this release, we aim to fuel new research in ego-exo, multimodal activity, and beyond.

2. Related work

Next we review prior work in datasets, human skill, and cross-view analysis. Section 5 will discuss related work for each benchmark task. Table 2 in Appendix 9 summarizes Ego-Exo4D's properties vs. existing datasets.

Egocentric datasets There has been a surge of interest in egocentric video understanding, facilitated by recent ego-video datasets showing unscripted daily-life activity as in Ego4D [47], EPIC-Kitchens [27, 28, 163], UT Ego [78], ADL [119], and KrishnaCam [147], or procedural activities as in EGTea [81], AssistQ [172], Meccano [126], CMU-MMAC [77], and EgoProcel [10]. Unlike any of the above, Ego-Exo4D focuses on multimodal ego *and* exo capture, and it is focused on the domain of skilled activities.

Multiview and ego-exo datasets Most existing multi-view datasets focus on static scenes [20, 128, 151, 175, 176] and objects [133, 173], with limited (exo only) multi-view human activity [26, 169]. CMU-MMAC [77] and CharadesEgo [145] are early efforts to capture both ego and exo video. CMU-MMAC [77] features 43 participants in mocap suits who cook 5 recipes in a lab kitchen. In CharadesEgo [145], 71 Mechanical Turkers record 34 hours of scripted scenarios (e.g., "type on laptop, then pick up a pillow") from the ego and exo perspectives sequentially, yielding unsynchronized videos with non-exact activity matches. More recent ego-exo efforts focus on specific activities in one or two environments. Assembly101 [139] and H2O [76] provide time-synced ego and exo video at a lab tabletop where people assemble toy cars or manipulate handheld objects, with 53 and 4 participants, and 513 and 5 hours of footage, respectively. Homage [127] provides 30 hours of ego-exo video from 27 participants in 2 homes doing household activities like laundry.

Compared to any of the prior efforts, Ego-Exo4D offers an order of magnitude more participants, diverse locations, and hours of footage (740 participants, 123 unique scenes, 13 cities, 1,286 hours). Importantly, our focus on skilled tasks takes the participants out of the lab or home and into settings like soccer fields, dance studios, rock climbing walls, and bike repair shops. Such activities also yield a wide variety of full body poses and movements within the scene, beyond using objects at a tabletop. This variety means Ego-Exo4D augments existing 3D human body pose datasets [49, 66, 68, 80, 193]. Finally, compared to any prior ego-exo resource, Ego-Exo4D's suite of modalities and benchmark tasks are novel and will expand the research directions the community can take for egocentric and/or exocentric video understanding.

Human skill and video learning Analyzing skill and action quality has received limited attention [12, 34, 35, 113, 120, 194]. Research in instructional or "how-to" videos is

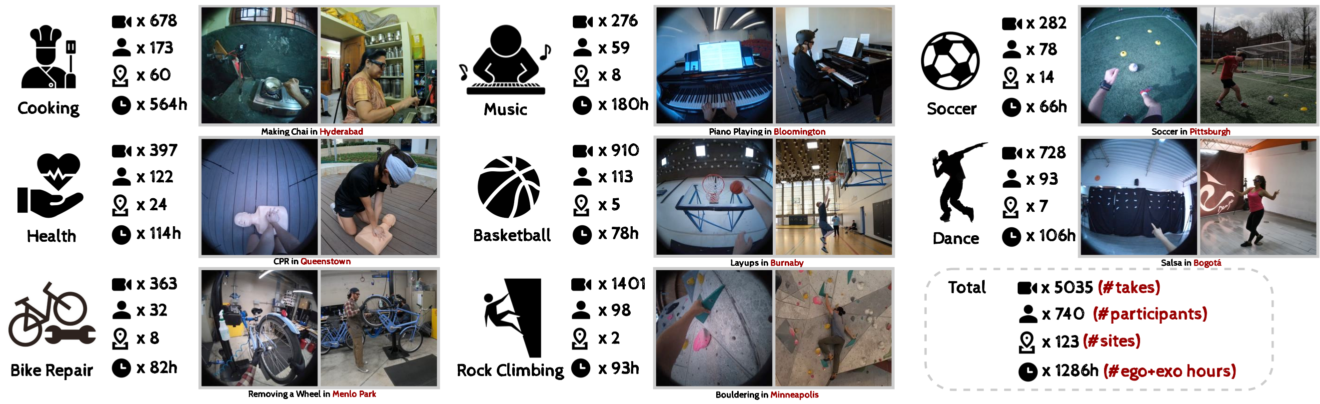

Figure 2. Ego-Exo4D captures skilled activity from 43 tasks and 689 keysteps in 8 domains, in a wide variety of 123 scenes in 13 cities in Japan, Colombia, Canada, India, Singapore, and 7 US states. Each domain is captured at multiple sites—from 2 to 64 unique locations. In total the dataset offers 1,286 hours of ego+exo video comprised of 5,035 takes from 740 camera wearers. An average take is 2.6 minutes.

facilitated by (largely exo) datasets like HowTo100M [103] and others [11, 159, 204, 207]. Challenges include grounding keysteps [10, 36, 37, 89, 103, 104, 178, 207], procedural planning [15, 17, 22, 71, 143, 167, 196, 201], learning task structure [4, 9, 37, 107, 202, 205], and leveraging noisy narrations [89, 103, 104]. A portion of Ego-Exo4D is procedural activities, but unlike the above, it offers simultaneous ego-exo capture. The scale and diversity of our data—including its three forms of language descriptions—widen the avenues for skilled activity understanding research.

Ego-exo cross-view modeling There is limited prior work on ego-exo cross-view modeling, arguably due to a lack of high-quality synchronized real-world data. Prior work explores matching people between videos [5, 6, 40, 170, 179] and learning view-invariant [7, 141, 144, 182, 184, 185] or ego features [82]. Beyond the specific case of ego-exo, cross-view methods are explored for translation [130, 131, 134, 157], novel view synthesis [19, 90, 135, 137, 164, 168, 171], and aerial to ground matching [86, 132]. Ego-Exo4D provides a testbed of unprecedented size and variety for cross-view modeling. In addition, our ego-exo relation tasks (cf. Section 5) surface new challenges in novel-view synthesis with widely varying viewpoints.

3. Ego-Exo4D dataset

Next we introduce the dataset and its scope. Notably, the video capture was a distributed but coordinated effort performed by 12 research labs. We present the common framework, and reserve site-specific details for Appendix 10.

3.1. Ego-exo camera rig

Our goal is to capture simultaneous ego and exo video, together with multiple egocentric sensing modalities. One of our contributions is to create and share a low-cost (less than $3,000), lightweight ego-exo rig with a user-friendly calibration and time sync procedure.

Our camera configuration features Aria glasses [38] for ego capture, leveraging their rich array of sensors, including an 8 MP RGB camera, two SLAM cameras, IMU, 7 microphones, and eye tracking (see Appendix 7). The ego camera is calibrated and time-synchronized with four to five (stationary) GoPros placed on tripods as the exo capture devices, allowing 3D reconstruction of the environment point clouds and the participant's body pose. The number and placement of the exocentric cameras is determined per scenario in order to allow maximal coverage of useful viewpoints without obstructing the participants' activity.

Our time sync and calibration design relies on a QR-code procedure to auto-sync the cameras and auto-separate the individual "takes", meaning instances of an activity. We can do continuous recordings of up to $\sim$60 minutes, based on the Aria battery life. See Appendix 8 for more details.

3.2. Domains and environments

Ego-Exo4D focuses on *skilled human activity*. This is in contrast to existing ego-only efforts like Ego4D [47], which has a broad span of daily-life activities. We intentionally select the domains based on a few criteria: Will it illustrate skill and a variety of expertise? Is there visual variety to be expected across different instances? Will the ego and exo views offer complementary information? Will it present new challenges unaddressed by current datasets?

Intersecting these criteria, we arrived at two broad categories[1] of skilled activity: *physical* and *procedural*, together comprising eight total domains. The physical domains are soccer, basketball, dance, bouldering, and music. They emphasize body pose and movements as well as

[1] Note that in general physical and procedural are not mutually exclusive labels. An activity can both require physical skill and procedural steps.

interaction with objects (e.g., a ball, musical instrument). The procedural domains are cooking, bike repair, and health care. They require performing a sequence of steps to reach a goal state (e.g., a completed recipe, a repaired bike) and generally entail intricate hand-object manipulations with a variety of objects (e.g., bike repair tools; cooking utensils, appliances, and ingredients).

In total, we have 43 activities derived from the eight domains (see Appendix 9). For example, cooking is comprised of 14 recipes; soccer is comprised of 3 drills. The length of a take ranges from 8 sec to 42 min, with procedural activities like cooking having the longest sustained captures.

To achieve visual diversity in the data, multiple labs across our team (typically 3-5) captured each Ego-Exo4D domain. The data is collected in authentic settings—such as real-world bike shops, soccer pitches, or bouldering gyms––as opposed to lab environments. For example, we have videos of chefs in New York City, Vancouver, Philadelphia, Bogota, and others; soccer players in Tokyo, Chapel Hill, Hyderabad, Singapore, and Pittsburgh. See Figure 2.

3.3. Participants: expertise and diversity

We recruited 740 total participants from the local communities of 12 labs. All scenarios feature real-world experts, where the camera-wearer participant has specific credentials, training, or expertise in the skill being demonstrated. For example, among the Ego-Exo4D camera wearers are professional and college athletes; jazz, salsa, and Chinese folk dancers and instructors; competitive boulderers; professional chefs who work in industrial-scale kitchens; bike technicians who service dozens of bikes per day. Many of them have (individually) over 10 years of experience.

Experts are prioritized given they are likely to conduct activities without mistakes or distractions, providing a strong ground truth for how to approach a given task. However, we also include capture from people with varying skill levels, as well—essential for our proposed skill proficiency estimation task (Section 5). Notably, Ego-Exo4D represents human intelligence in a new way by capturing domain-specific expertise—both in the video as well as the accompanying expert commentary (see Section 4)—portraying the evolution of a skill from beginners to experts.

According to the participant surveys (Appendix 11), the camera wearers range in age from 18 to 74 years old, with 37% self-identifying as female 60% male and 3% as non-binary or preferring not to say. In total, the participants self report more than 24 different ethnicities.

3.4. Privacy and ethics

Ego-Exo4D was collected following rigorous privacy and ethics standards. This included undergoing formal independent review processes at each institution to establish the standards for collection, management, and informed con-

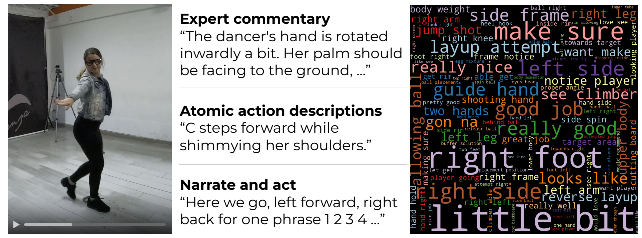

Figure 3. Ego-Exo4D offers 3 paired language corpora. Word cloud is from expert commentary which critiques the performance.

sent. Similarly, all Ego-Exo4D data collection adhered to the Project Aria Research Community Guidelines for responsible research. Since the scenarios allow for closed environments (e.g., no passerbys) nearly all video is available without de-identification. For information about each individual partners' protocols and restrictions, please see Appendix 10. Ego-Exo4D data is gated behind a license system, which defines permitted uses, restrictions, and consequences for non-compliance.

4. Natural language descriptions

Ego-Exo4D also offers three kinds of paired natural language datasets, each time-indexed alongside the video. See Figure 3. These language annotations are not steered towards any single benchmark, but rather are a general resource that will support browsing and mining the dataset—as well as challenges in video-language learning like grounding actions and objects, self-supervised representation learning, video-conditioned language models, and skill assessment. See Appendix 12.

The first language dataset is spoken *expert commentary*. The goal is to reveal nuances of the skill that are not always visible to non-experts. We recruited 52 experts (distinct from the participants) to critique the recorded videos, call out strengths and weaknesses, explain how the specific behavior of the participant (e.g., hand/body pose, use of objects) affects the performance, and provide spatial markings to support their commentary. The experts are not only well-credentialed in their areas of expertise, but also have coaching or teaching experience, which facilitates clear communication. They watch the video and pause every time they have a comment, typically 7 times per minute of video. Each piece of commentary is unbounded in length, and averages 4 sentences. We provide both the transcribed speech and the raw audio (interesting for its inflection and non-word utterances), as well as the experts' spatial drawings and numeric ratings of each participant's skill. Videos have expert commentary by 2-5 distinct experts, offering a variety of perspectives for the same content. In total, we have 117,812 pieces of time-stamped, video-aligned commentary. These commentaries are quite novel: they focus on *how* the activity is executed rather than *what* it entails,

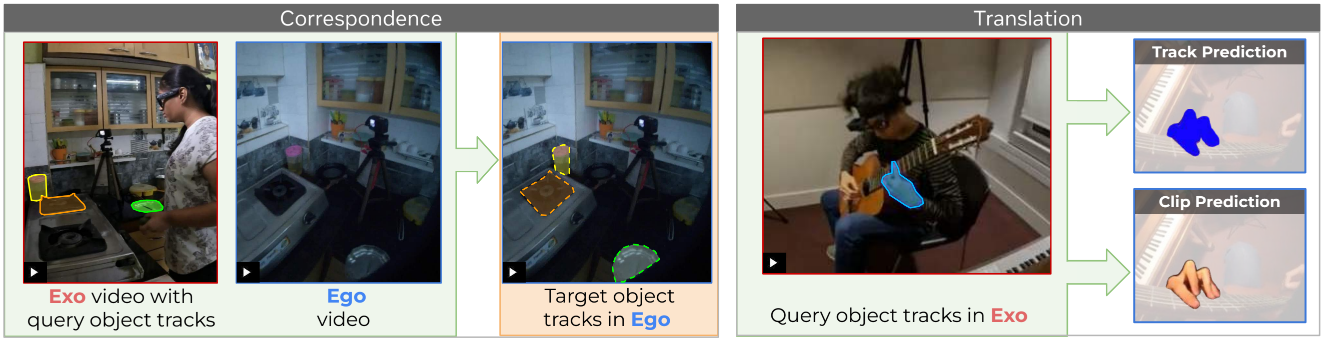

Figure 4. The ego-exo relation family consists of the tasks of correspondence (left) and translation (right).

capturing subtle differences in skilled execution. We believe this can unlock new fundamental problems (e.g., proficiency estimation below) and disruptive future applications (e.g., AI coaching).

The second language dataset consists of *narrate-and-act descriptions* provided by the participants themselves. They are in the style of a tutorial or how-to video, where the participant explains what they are doing and why. Unlike the third-party expert commentary above, these are first-person reflections on the activity given by the people doing them. These narrations are available for about 10% of all takes in the dataset, since we wanted participants to execute the tasks without pausing for the bulk of the recordings.

The third language dataset consists of *atomic action descriptions*. Whereas the commentary and narrate-and-act language reveals spoken opinions and reasons for the actions (the "why and how"), this stream of text is specifically about the "what". Inspired by Ego4D's narrations [47], these are short statements written by third-party (non-domain expert) annotators, timestamped for every atomic action performed by the participant for all videos in the dataset, for a total of 432K sentences. This data is valuable for mining for taxonomies of objects and actions in the data, indexing the videos with keywords for exploring the dataset, and for future research in video-language learning, as has been quite successful for the Ego4D narrations [8, 85, 123].

5. Ego-Exo4D benchmark tasks

Our second major contribution is to define the core research challenges in the domain of egocentric perception of skilled activity, particularly when ego-exo data is available for training (if not testing). To that end, we devise a suite of foundational benchmark tasks organized into four task families: relation (Sec. 5.1), recognition (Sec. 5.2), proficiency (Sec. 5.3), and ego-pose (Sec. 5.4). For each task, we provide high quality annotations and baselines that provide a starting point from which the research community can build. We will run the first formal Ego-Exo4D challenges in 2024. Due to space limits, we briefly overview each task; see the referenced Appendices for all details including baseline models and results. There are two publicly released versions of Ego-Exo4D annotations: v1 is used to train/test baselines in this paper; the larger v2 will be used for future challenge leaderboards (see Table 7 in Appendix).

5.1. Ego-exo relation

Our ego-exo *relation* tasks deal with relating the video content across the extreme ego-exo viewpoint changes. They take the form of object-level matching (correspondence) and synthesis of one view from the other (translation).

5.1.1 Ego-exo correspondence

Motivation. Establishing object-level correspondences between ego and exo viewpoints would allow AI assistants to provide visual instructions by matching third-person observations of objects from instructional videos to those in the user's first-person view. Compared to the general correspondence problem, our setting requires tackling a number of challenges: extreme viewpoint differences, high degrees of object occlusion, and many small objects (e.g., cooking utensils and bike repair tools).

Task definition. Given a pair of synchronized ego-exo videos and a sequence of query masks of an object of interest in one of the videos, the task is to predict the corresponding mask for the same object in each synchronized frame of the other view, if it is visible. See Figure 4, left. The task can be posed with query objects in either the ego or exo video, with both directions presenting interesting challenges (e.g., high degree of occlusion in ego views, and small object size in exo views). See Appendix 13.A.1.

Related work. Related tasks are image-level sparse correspondence given query points (instead of object masks) [65] and image-level object co-segmentation [166] for jointly segmenting semantically similar objects. Our task goes beyond static object correspondence, since the interplay between human pose and object state changes during manipulation necessitate using temporal context and tracking as the query object can be highly occluded or blurry [158].

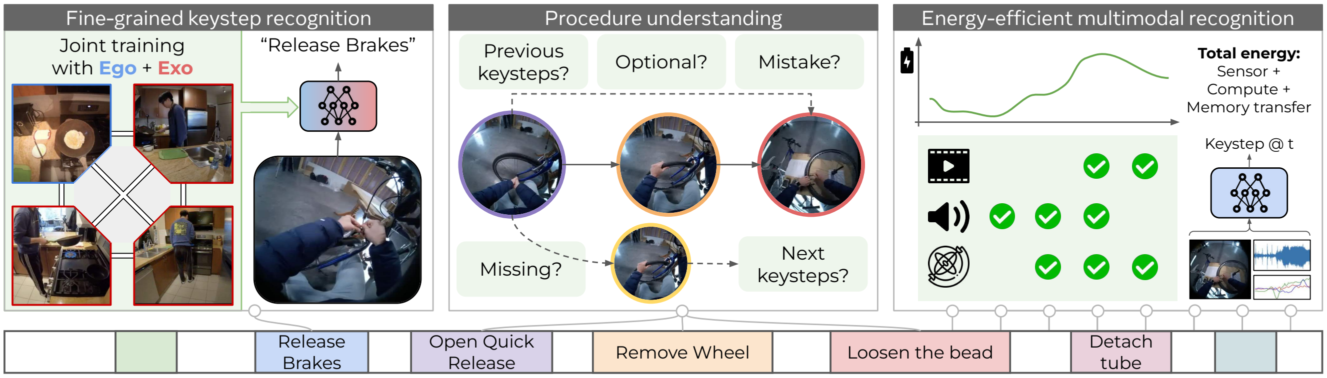

Figure 5. **Ego-exo keystep recognition.** This family of tasks consists of fine-grained recognition (left, Section 5.2.1), procedure understanding (center, Section 5.2.3), and energy-efficient multimodal recognition (right, Section 5.2.2).

5.1.2 Ego-exo translation

Motivation. Our translation task entails synthesizing a target ego clip from a given exo clip. We believe this problem will drive novel research for combining recognition and object synthesis. For example, in Figure 4 (right), the approach must make effective use of the hand's object-specific shape and appearance priors in order to synthesize the ego view of the fingertips—which are not visible in the exo clip. Furthermore, this task will stimulate advances in visual odometry, as the method must be able to infer the ego camera pose from the third-person clip. Ego-exo translation also holds strong application potential, as it may unlock the ability to generate first-person renderings of videos that were originally captured from a third-person perspective, e.g., benefitting robot perception or AR coaching.

Task definition. We decompose ego-exo translation into two separate tasks: *ego track prediction* and *ego clip generation* (Figure 4, right). Ego track prediction estimates the segmentation mask of an object in the *unobserved* ego frames given the object masks in the observed exo clip. Ego clip generation must generate the image values (i.e., RGB) within the given ground-truth ego mask by making use of the exo clip and the object masks in those frames. This decomposition effectively splits the problem into two tasks: 1) predicting the location and shape of the object in the ego clip, and 2) synthesizing its appearance given the ground-truth position. For each, we consider a variant where the pose of the ego camera with respect to the exo camera is available to use at inference time. This simplifies the problem but reduces the applicability of the method, since this information is typically not available for arbitrary third-person videos. See Appendix 13.A.2.

Related work. Ego-exo translation relates to cross-view image synthesis [96, 130, 157]. Within this genre, the problem of exo-to-ego generation was recently introduced for both images [93] and video [94, 97], and approached using GANs or diffusion conditioned on the input view. Our work not only formalizes this task with ample data, but its formulation also draws attention to the need for a *semantic* basis to new view synthesis across extreme view changes.

5.2. Ego-exo keystep recognition

This family of tasks centers around recognizing the keysteps of a procedural activity and modeling their dependencies.

5.2.1 Fine-grained keystep recognition

Motivation. Recognizing the step a camera wearer is performing is non-trivial: keysteps in the same activity may look similar (folding vs. smoothing the bedsheet) and may involve hand-object interactions with heavy occlusions and head motion. Models with access to multiple views during training can leverage their complementarity to account for the deficiencies of each one, by learning viewpoint invariant representations or distilling multi-view signals into a single model (e.g., human hands from ego; body pose from exo).

Task definition. We study ego-exo for video recognition. During training, models have access to paired ego-exo data—time-synchronized captures of the same activity from multiple known viewpoints. Each training instance has one ego view, N exo views, and a corresponding keystep label (e.g., "flip the omelette"). At test time, given only a trimmed *egocentric* video clip, the model must identify the keystep performed from a taxonomy of 689 keysteps across 17 procedural activities. See Figure 5, left. Importantly, all extra supervision (time-alignment, camera poses etc.) is only available at training time; inference is standard keystep recognition, but with models that benefit from cross-viewpoint training. See Appendix 13.B.1.

Related work. Keystep recognition has been studied in first-person [10, 126, 145, 148] or third-person [9, 100, 159, 205, 207] videos; however, limited work considers both views together. Prior work considers cross-view learning with unpaired videos [7, 82, 182] and view-invariant feature learning on paired videos [144]. In contrast, we explore keystep recognition in large-scale, procedural activities with fully synchronized training videos.

5.2.2 Energy-efficient multimodal keystep recognition

Motivation. Current activity detection models assume access to densely sampled clips from the full video and ample computational resources to process them. These assumptions are incompatible with real-world devices (e.g., mobile phones, AR glasses) where the camera is not always on and the compute budget is limited by battery life. This task focuses on building energy-efficient video models to pave the way for feasibility on real-world hardware.

Task definition. We formulate the problem as an online action detection task, with a given energy budget. See Figure 5, right. Given a stream of audio, IMU, and RGB video data, a model must identify the keystep being performed at each frame, as well as decide which sensor(s) to use for subsequent time-steps. This task will inspire models that are strategic about which modality to deploy when. Energy consumption is the sum of sensor energy (operating the camera/audio/IMU sensors), model inference costs, and memory transfer costs, and must be within 20mW to reflect real-world device power constraints. See Appendix 13.B.2.

Related work. Prior work on efficient models considers light-weight architectures [41, 56, 101, 155, 165, 195], efficient input processing [43, 44, 73, 102, 156], or inference optimizations [39, 58, 121, 174, 206]. In all cases, they optimize computation (FLOPs), parameter count, or prediction throughput (FPS), which in isolation are insufficient to characterize running on real-world devices. To address this, we propose the first benchmark for *energy-efficient* video recognition that is tied to real-world, on-device constraints, and measures total power consumed.

5.2.3 Procedure understanding

Motivation. Automatically understanding the *structure* of a procedure from video (inferring keystep ordering, preconditions, etc.) would allow assisting AR users in a task or informing robots that learn from human demonstrations.

Task definition. In our procedure understanding task, given a video segment s_t and its previous video segment history, models have to 1) determine *previous keysteps* (to be performed before s_t); infer if s_t is 2) *optional* or 3) a *procedural mistake*; 4) predict *missing keysteps* (should have been performed before s_t but were not); and 5) *next keysteps* (for which dependencies are satisfied). The task offers two versions of weak supervision: instance-level: segments and their keystep labels are available for train/test; and procedure-level: only unlabeled segments and procedure-specific keystep names are given for train/test. See Figure 5 (center) and Appendix 13.B.3.

Related work. Prior work focusing on procedural understanding learns an explicit graph [62, 150, 177] as ground truth or uses a task graph for representation learning [9, 107, 202] and short-term step understanding [9, 36, 202]. Other work [32, 139] studies mistake detection in a supervised setting. We are the first to propose procedural understanding to evaluate the long-term structure of the task in a weakly-supervised setting.

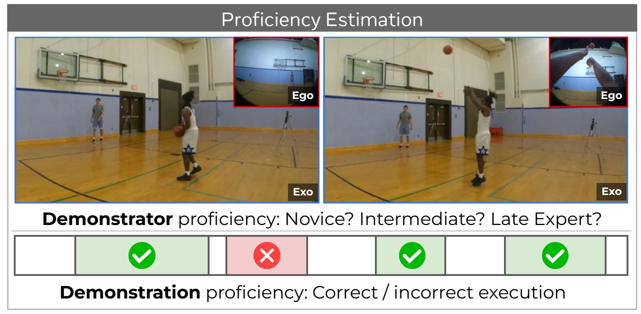

Figure 6. Demonstrator and demonstration proficiency estimation.

5.3. Ego-exo proficiency estimation

Motivation. Going beyond recognizing what a person is doing, this task aims to infer the user's skill level. Such an ability could lead to novel coaching tools that let people learn new skills more effectively, or new ways to *evaluate* human performance in domains like sports or music.

Task definition. We consider two variants: (1) *demonstrator* and (2) *demonstration* proficiency estimation. Both tasks consider one egocentric and (optionally) M exocentric videos synchronized in time as their inputs. Demonstrator proficiency is formulated as a video classification task, where the model has to output one of four labels (novice, early, intermediate, or late expert). Demonstration proficiency is formulated as a temporal action localization task where given an untrimmed video, the model must output a list of tuples, each containing a timestamp, a proficiency category (i.e., good execution or needs improvement), and its probability. Note that parts of the video that do not reveal the participant's skill are left unlabeled. See Figure 6 and Appendix 13.C.

Related work. Prior work uses egocentric [12, 35] or exocentric [60, 114, 115] views for proficiency estimation in sports [12, 115, 120], health [60, 92, 191, 208], and others [35, 186]. We propose the first multi-view egocentric and exocentric proficiency estimation benchmark. Unlike prior work, our benchmark spans diverse, day-to-day physical and procedural scenarios and includes temporally localized annotations of (in)correct executions.

5.4. Ego pose

This family of tasks is motivated by recovering the skilled body movements of participants, even in the extreme setting of monocular ego-video input in dynamic environments.

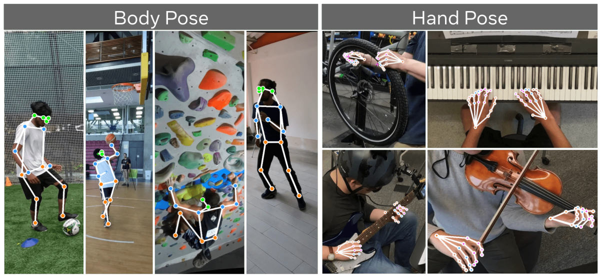

Figure 7. Hand and body keypoints for ego-pose estimation

Motivation. Estimating the physical state of a person's body—the 3D positions of the arms, legs, hands—from the ego view is essential for wearable AI systems that can support human activity. Challenges include subtle and flexible movements, frequent occlusion, and body parts out of view.
Task definition. For both the body and hand pose ("ego pose") estimation tasks, the input is an ego video. The output is a set of 3D joint positions of the camera wearer's body and hands for each time step, parameterized as 17 3D body joint positions and 21 3D joint positions per hand, following the MS COCO convention [87]. To our knowledge, Ego-Exo4D offers the largest manual ground truth (GT) egocentric body and hand pose annotations to date. And, in total, it offers ∼14M frames of 3D GT and pseudo-GT combined. See Figure 7 and Appendix 13.D.
Related work. Limited prior work explores 3D body pose from a wearable camera. Some methods assume no body visibility [63, 80, 98, 187, 188], while others assume partial observability by modifying cameras to capture the body [3, 57, 136, 161, 181]. Our dataset can be used for both paradigms. Existing hand pose datasets use constrained environments [106, 146] with simple hand motion [50, 76, 109], whereas we include diverse real-world scenarios, e.g., with expert musicians and bike mechanics.

6. Conclusions

Ego-Exo4D provides a dataset of unprecedented scale and realism for ego-exo video learning. It offers a unique window into skilled human activity from 8 compelling domains by hundreds of real-world experts around the globe. Together with the proposed benchmarks, we hope that this new open source resource will set the stage for substantial new research for the years to come.

Though we are motivated by skill learning, Ego-Exo4D is poised for even broader influence, beyond the proposed benchmarks. Whereas existing datasets lack activity modeling in real-world 3D contexts (e.g., restricted to mocap suits and/or lab settings). Ego-Exo4D is a resource for **general 3D vision**—such as environment reconstruction, camera relocalization, audio-visual mapping, and many others. Similarly, our novel **video-language** resources will offer many opportunities for grounding of actions and objects, multimodal representation learning, and language generation. Finally, though our tasks prioritize perception from the "ego-only" perspective, the exo component of our data ensures its utility for the more **traditional exo viewpoint** too, e.g., for activity recognition and body pose estimation.

Contribution statement

This project is the result of a large collaboration between many institutions over the last two years. Initial authors represent the leadership team of the project. Kristen Grauman initiated the project, served as the technical lead, initiated the recognition and proficiency benchmarks and expert commentary, and coordinated their working groups. Andrew Westbury served as the program manager and operations lead for all aspects of the project. Lorenzo Torresani led development of the capture domains, initiated the relation and ego-pose benchmarks, and coordinated their working groups. Kris Kitani led development of the multi-camera rig and supported the Ego-Exo4D engineering team on all aspects of the data annotation and organization. Jitendra Malik served as a scientific advisor. Authors with stars (*) were key drivers of implementation, collection, and/or annotation development throughout the project. Authors with daggers (†) are faculty and senior researcher PIs for the project. The Appendices detail the contributions of individual authors for the various benchmarks, data collection, and annotation pipelines.

Acknowledgements

We gratefully acknowledge the following colleagues for valuable discussions and support of our project: Vittorio Caggiano, Ilé Danza, Ahmad Darkhalil, Zona de Bloque, Alex Dinh, Rene Martinez Doehner, Ivan Cruz, Matt Feiszli, Vance Feutz, Kelly Forbes, Rohit Girdhar, Pierre Gleize, Andrés Hernández, Shun Iwase, Bolin Lai, Vivian Lee, Brighid Meredith, Ashley Massie, Natalia Neverova, Manohar Paluri, Joelle Pineau, Artsiom Sanakoyeu, Paresh Shenoy, Jiaray Shi, Jiasheng Shi, Gaurav Shrivastava, Mitesh Singh, Manasi Swaminathan, Arjang Talattof, Ali Thabet, Laurens van der Maaten, Andrea Vedaldi, and Tobby Zhu. We also sincerely thank the 52 experts who contributed to the expert commentary for their expertise and support; they are listed individually in Appendix 12. Thank you to the Common Visual Data Foundation (CVDF) for hosting the Ego-Exo4D dataset. Finally, thank you to the 740 participants who contributed to this dataset and shared their skills in video.

Univ. of Bristol is supported in part by EPSRC UMPIRE (EP/T004991/1) and EPSRC PG Visual AI (EP/T028572/1). Z. Zhu is supported by UoB-CSC Scholarship. Univ. of Catania is supported in part by the project Future Artificial Intelligence Research (FAIR) – PNRR MUR Cod. PE0000013 - CUP: E63C22001940006. Simon Fraser Univ. is supported in part by the Canada Research Chairs Program (CRC-2019-00298) and NSERC Discovery (2019-06489). Georgia Tech is supported in part by NSF award CNS-2308994. IU is supported in part by NSF DRL-2112635 (AI Institute for Engaged Learning). UT Austin is supported in part by the IFML NSF AI Institute.

References

[1] Ahmed Tawfik Aboukhadra, Jameel Malik, Ahmed Elhayek, Nadia Robertini, and Didier Stricker. Thor-net: End-to-end graformer-based realistic two hands and object reconstruction with self-supervision. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1001–1010, 2023. 58, 59

[2] Michael Abrash. Creating the future: Augmented reality, the next human-machine interface. In *2021 IEEE International Electron Devices Meeting (IEDM)*, 2021. 46

[3] Karan Ahuja, Chris Harrison, Mayank Goel, and Robert Xiao. Mecap: Whole-body digitization for low-cost vr/ar headsets. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, pages 453–462, 2019. 9

[4] Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. Unsupervised learning from narrated instruction videos. In *CVPR*, 2016. 4

[5] Shervin Ardeshir and Ali Borji. Ego2top: Matching viewers in egocentric and top-view videos. In *ECCV*, 2016. 4

[6] Shervin Ardeshir and Ali Borji. Egocentric meets top-view. *IEEE transactions on pattern analysis and machine intelligence*, 41(6), 2018. 4

[7] Shervin Ardeshir and Ali Borji. An exocentric look at egocentric actions and vice versa. *Computer Vision and Image Understanding*, 171, 2018. 4, 7

[8] Kumar Ashutosh, Rohit Girdhar, Lorenzo Torresani, and Kristen Grauman. Hiervl: Learning hierarchical video-language embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 6

[9] Kumar Ashutosh, Santhosh Kumar Ramakrishnan, Triantafyllos Afouras, and Kristen Grauman. Video-mined task graphs for keystep recognition in instructional videos. In *NeurIPS*, 2023. 4, 7, 8, 29, 49

[10] Siddhant Bansal, Chetan Arora, and C.V. Jawahar. My view is the best view: Procedure learning from egocentric videos. In *European Conference on Computer Vision (ECCV)*, 2022. 3, 4, 7, 49

[11] Yizhak Ben-Shabat, Xin Yu, Fatemehsadat Saleh, Dylan Campbell, Cristian Rodriguez-Opazo, Hongdong Li, and Stephen Gould. The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose. 2020. 4

[12] Gedas Bertasius, Hyun Soo Park, Stella Yu, and Jianbo Shi. Am i a baller? basketball performance assessment from first-person videos. In *ICCV*, 2017. 3, 8

[13] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, 2021. 39, 43, 44, 47, 54, 55, 56

[14] Jing Bi, Jiebo Luo, and Chenliang Xu. Procedure planning in instructional videos via contextual modeling and model-based policy learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15611–15620, 2021. 49

[15] Jing Bi, Jiebo Luo, and Chenliang Xu. Procedure planning in instructional videos via contextual modeling and model-based policy learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15611–15620, 2021. 4

[16] Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann. The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition*, pages 3121–3124. IEEE, 2010. 35

[17] Meng Cao, Tianyu Yang, Junwu Weng, Can Zhang, Jue Wang, and Yuexian Zou. Locvtp: Video-text pre-training for temporal localization. In *European Conference on Computer Vision*, 2022. 4

[18] Angela Castillo, Maria Escobar, Guillaume Jeanneret, Albert Pumarola, Pablo Arbeláez, Ali Thabet, and Artsiom Sanakoyeu. Bodiffusion: Diffusing sparse observations for full-body human motion synthesis. *CV4Metaverse workshop, International Conference on Computer Vision*, 2023. 57, 58

[19] Eric R Chan, Koki Nagano, Matthew A Chan, Alexander W Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3d-aware diffusion models. *arXiv preprint arXiv:2304.02602*, 2023. 4

[20] Angel Chang, Angela Dai, Tom Funkhouser, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2017. MatterPort3D dataset license available at: `http://kaldir.vc.in.tum.de/matterport/MP_TOS.pdf`. 3

[21] Chien-Yi Chang, De-An Huang, Danfei Xu, Ehsan Adeli, Li Fei-Fei, and Juan Carlos Niebles. Procedure planning in instructional videos. In *European Conference on Computer Vision*, pages 334–350. Springer, 2020. 49

[22] Chien-Yi Chang, De-An Huang, Danfei Xu, Ehsan Adeli, Li Fei-Fei, and Juan Carlos Niebles. Procedure planning in instructional videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI*, pages 334–350. Springer, 2020. 4

[23] Yu-Hsin Chen, Tien-Ju Yang, Joel Emer, and Vivienne Sze. Eyeriss v2: A flexible accelerator for emerging deep neural networks on mobile devices. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 2019. 46

[24] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European Conference on Computer Vision*, pages 640–658. Springer, 2022. 36

[25] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. `https://github.com/open-mmlab/mmpose`, 2020. 56

[26] Kellie Corona, Katie Osterdahl, Roderic Collins, and Anthony Hoogs. Meva: A large-scale multiview, multimodal video dataset for activity detection. In *WACV*, 2021. 3, 7

[27] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018. 3, 27

[28] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, , Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision. *IJCV*, 2021. 2, 3

[29] 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*, 130(1):33–55, 2022. 7

[30] Roeland De Geest, Efstratios Gavves, Amir Ghodrati, Zhenyang Li, Cees Snoek, and Tinne Tuytelaars. Online action detection. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 269–284. Springer, 2016. 46, 51

[31] Radosvet Desislavov, Fernando Martínez-Plumed, and José Hernández-Orallo. Trends in ai inference energy consumption: Beyond the performance-vs-parameter laws of deep learning. *Sustainable Computing: Informatics and Systems*, 38:100857, 2023. 46

[32] Guodong Ding, Fadime Sener, Shugao Ma, and Angela Yao. Every mistake counts in assembly. *arXiv preprint arXiv:2307.16453*, 2023. 8

[33] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020. 39

[34] H Doughty, D Damen, and W Mayol-Cuevas. Who's better? who's best? pairwise deep ranking for skill determination. In *CVPR*, 2018. 3

[35] Hazel Doughty, Walterio Mayol-Cuevas, and Dima Damen. The Pros and Cons: Rank-aware Temporal Attention for Skill Determination in Long Videos. 2019. 3, 8

[36] Nikita Dvornik, Isma Hadji, Hai Pham, Dhaivat Bhatt, Brais Martinez, Afsaneh Fazly, and Allan D Jepson. Flow graph to video grounding for weakly-supervised multi-step localization. In *ECCV*, pages 319–335. Springer, 2022. 4, 8

[37] Ehsan Elhamifar and Dat Huynh. Self-supervised multi-task procedure learning from instructional videos. In *European Conference on Computer Vision*, pages 557–573. Springer, 2020. 4

[38] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, Cheng Peng, Chris Sweeney, Cole Wilson, Dan Barnes, Daniel DeTone, David Caruso, Derek Valleroy, Dinesh Ginjupalli, Duncan Frost, Edward Miller, Elias Mueggler, Evgeniy Oleinik, Fan Zhang, Guruprasad Somasundaram, Gustavo Solaira, Harry Lanaras, Henry Howard-Jenkins, Huixuan Tang, Hyo Jin Kim, Jaime Rivera, Ji Luo, Jing Dong, Julian Straub, Kevin Bailey, Kevin Eckenhoff, Lingni Ma, Luis Pesqueira, Mark Schwesinger, Maurizio Monge, Nan Yang, Nick Charron, Nikhil Raina, Omkar Parkhi, Peter Borschowa, Pierre Moulon, Prince Gupta, Raul Mur-Artal, Robbie Pennington, Sachin Kulkarni, Sagar Miglani, Santosh Gondi, Saransh Solanki, Sean Diener, Shangyi Cheng, Simon Green, Steve Saarinen, Suvam Patra, Tassos Mourikis, Thomas Whelan, Tripti Singh, Vasileios Balntas, Vijay Baiyya, Wilson Dreewes, Xiaqing Pan, Yang Lou, Yipu Zhao, Yusuf Mansour, Yuyang Zou, Zhaoyang Lv, Zijian Wang, Mingfei Yan, Carl Ren, Renzo De Nardi, and Richard Newcombe. Project Aria: A new tool for egocentric multi-modal AI research, 2023. 3, 4, 2

[39] Steven K Esser, Jeffrey L McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S Modha. Learned step size quantization. *arXiv preprint arXiv:1902.08153*, 2019. 8

[40] Chenyou Fan, Jangwon Lee, Mingze Xu, Krishna Kumar Singh, Yong Jae Lee, David J Crandall, and Michael S Ryoo. Identifying first-person camera wearers in third-person videos. In *CVPR*, 2017. 4

[41] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 203–213, 2020. 8, 45, 46, 47, 48

[42] John H. Flavell, Eleanor R. Flavell, Frances L. Green, and Sharon A. Wilcox. The development of three spatial perspective-taking rules. *Child Development*, 1981. 2

[43] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *CVPR*, 2020. 8

[44] Amir Ghodrati, Babak Ehteshami Bejnordi, and Amirhossein Habibian. Frameexit: Conditional early exiting for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 8

[45] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16102–16112, 2022. 54

[46] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 42

[47] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph

Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolář, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4D: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022. 2, 3, 4, 6, 7, 29, 34, 43

[48] Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*, 2018. 2

[49] Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In *CVPR*, 2021. 3

[50] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3196–3206, 2020. 9

[51] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 36, 58

[52] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 43, 44

[53] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 57

[54] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010. 39

[55] Mark Horowitz. 1.1 computing's energy problem (and what we can do about it). In *2014 IEEE international solid-state circuits conference digest of technical papers (ISSCC)*, 2014. 46

[56] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 8

[57] Dong-Hyun Hwang, Kohei Aso, Ye Yuan, Kris Kitani, and Hideki Koike. Monoeye: Multimodal human motion capture system using a single ultra-wide fisheye camera. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, pages 98–111, 2020. 9

[58] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016. 8

[59] Vladimir Iashin and Esa Rahtu. Multi-modal dense video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 958–959, 2020. 26

[60] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Evaluating surgical skills from kinematic data using convolutional neural networks. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 214–221, 2018. 8

[61] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017. 39

[62] Yunseok Jang, Sungryull Sohn, Lajanugen Logeswaran, Tiange Luo, Moontae Lee, and Honglak Lee. Multimodal subtask graph generation from instructional videos. *arXiv preprint arXiv:2302.08672*, 2023. 8

[63] Hao Jiang and Kristen Grauman. Seeing invisible poses: Estimating 3d body pose from egocentric video. In *CVPR*, 2017. 9

[64] Jiaxi Jiang, Paul Streli, Huajian Qiu, Andreas Fender, Larissa Laich, Patrick Snape, and Christian Holz. Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In *European Conference on Computer Vision*, pages 443–460. Springer, 2022. 57, 58

[65] Wei Jiang, Eduard Trulls, Jan Hosang, Andrea Tagliasacchi, and Kwang Moo Yi. Cotr: Correspondence transformer for matching across images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6207–6217, 2021. 6, 35

[66] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 3

[67] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 2, 42, 43

[68] Rawal Khirodkar, Aayush Bansal, Lingni Ma, Richard Newcombe, Minh Vo, and Kris Kitani. Egohumans: An egocentric 3d multi-human benchmark. In *ICCV*, 2023. 3

[69] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 36

[70] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar,

and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 33

[71] Dohwan Ko, Joonmyung Choi, Juyeon Ko, Shinyeong Noh, Kyoung-Woon On, Eun-Sol Kim, and Hyunwoo J Kim. Video-text representation learning via differentiable weak temporal alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5016–5025, 2022. 4

[72] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2252–2261, 2019. 59

[73] Bruno Korbar, Du Tran, and Lorenzo Torresani. Scsampler: Sampling salient clips from video for efficient action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 8

[74] Zuzana Kukelova, Jan Heller, and Andrew Fitzgibbon. Efficient intersection of three quadrics and applications in computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1799–1808, 2016. 3

[75] Iljung Kwak, Jian-Zhong Guo, Adam Hantman, David Kriegman, and Kristin Branson. Detecting the starting frame of actions in video. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 489–497, 2020. 54

[76] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10138–10148, 2021. 2, 3, 9, 7

[77] F. De la Torre, J. Hodgins, J. Montano, S. Valcarcel, R. Forcada, and J. Macey. Guide to the carnegie mellon university multimodal activity (cmu-mmac) database. In *Tech. report CMU-RI-TR-08-22, Robotics Institute, Carnegie Mellon University*, 2009. 2, 3, 7

[78] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR*, 2012. 3

[79] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3383–3393, 2021. 59

[80] Jiaman Li, Karen Liu, and Jiajun Wu. Ego-body pose estimation via ego-head pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17142–17151, 2023. 3, 9, 56, 57, 58

[81] Yin Li, Miao Liu, and James M. Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *ECCV*, 2018. 3, 7

[82] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6943–6953, 2021. 4, 7, 21, 43, 44

[83] Junhua Liao, Haihan Duan, Kanghui Feng, Wanbing Zhao, Yanbing Yang, and Liangyin Chen. A light weight model for active speaker detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22932–22941, 2023. 47, 48

[84] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1954–1963, 2021. 59

[85] Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *NeurIPS*, 2022. 6, 26, 50

[86] TY Lin, Y Cui, S Belongie, and J Hays. Learning deep representations for ground-to-aerial geolocalization. In *CVPR*, 2015. 4

[87] 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 *ECCV*, 2014. 9

[88] 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. 58

[89] Xudong Lin, Fabio Petroni, Gedas Bertasius, Marcus Rohrbach, Shih-Fu Chang, and Lorenzo Torresani. Learning to recognize procedural activities with distant supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13853–13863, 2022. 4, 29

[90] Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14458–14467, 2021. 4

[91] Chiao Liu, Lyle Bainbridge, Andrew Berkovich, Song Chen, Wei Gao, Tsung-Hsun Tsai, Kazuya Mori, Rimon Ikeno, Masayuki Uno, Toshiyuki Isozaki, et al. A 4.6 μm, 512$\times$ 512, ultra-low power stacked digital pixel sensor with triple quantization and 127db dynamic range. In *2020 IEEE International Electron Devices Meeting (IEDM)*, 2020. 46

[92] Daochang Liu, Qiyue Li, Tingting Jiang, Yizhou Wang, Rulin Miao, Fei Shan, and Ziyu Li. Towards unified surgical skill assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9522–9531, 2021. 8

[93] Gaowen Liu, Hao Tang, Hugo Latapie, and Yan Yan. Exocentric to egocentric image generation via parallel generative adversarial network. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1843–1847. IEEE, 2020. 7

[94] Gaowen Liu, Hao Tang, Hugo M Latapie, Jason J Corso, and Yan Yan. Cross-view exocentric to egocentric video

synthesis. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 974–982, 2021. 7

[95] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 47

[96] Xiaohu Lu, Zuoyue Li, Zhaopeng Cui, Martin R. Oswald, Marc Pollefeys, and Rongjun Qin. Geometry-aware satellite-to-ground image synthesis for urban areas. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 7

[97] Mi Luo, Zihui Xue, Alex Dimakis, and Kristen Grauman. Put myself in your shoes: Lifting the egocentric perspective from exocentric videos. *arXiv:2403.06351*, 2024. 7

[98] Zhengyi Luo, Ryo Hachiuma, Ye Yuan, and Kris Kitani. Dynamics-regulated kinematic policy for egocentric pose estimation. In *Advances in Neural Information Processing Systems*, 2021. 9, 56, 57, 58

[99] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 57

[100] Effrosyni Mavroudi, Triantafyllos Afouras, and Lorenzo Torresani. Learning to ground instructional articles in videos through narrations. 2022. 7

[101] Sachin Mehta and Mohammad Rastegari. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178*, 2021. 8

[102] Yue Meng, Chung-Ching Lin, Rameswar Panda, Prasanna Sattigeri, Leonid Karlinsky, Aude Oliva, Kate Saenko, and Rogerio Feris. Ar-net: Adaptive frame resolution for efficient action recognition. In *ECCV 2020*, 2020. 8

[103] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*, 2019. 2, 4, 29

[104] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020. 4, 29

[105] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, and Aude Oliva. Moments in time dataset: one million videos for event understanding. *PAMI*, 2019. 2

[106] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 548–564. Springer, 2020. 9

[107] Medhini Narasimhan, Licheng Yu, Sean Bell, Ning Zhang, and Trevor Darrell. Learning and verification of task structure in instructional videos. *arXiv preprint arXiv:2303.13519*, 2023. 4, 8

[108] Nora Newcombe. The development of spatial perspective taking. *Advances in child development and behavior*, 1989. 2

[109] Takehiko Ohkawa, Kun He, Fadime Sener, Tomas Hodan, Luan Tran, and Cem Keskin. Assemblyhands: Towards egocentric activity understanding via 3d hand pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12999–13008, 2023. 9, 7

[110] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 43, 44

[111] Boxiao Pan, Haoye Cai, De-An Huang, Kuan-Hui Lee, Adrien Gaidon, Ehsan Adeli, and Juan Carlos Niebles. Spatio-temporal graph for video captioning with knowledge distillation. In *CVPR*, 2020. 26

[112] JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Handoccnet: Occlusion-robust 3d hand mesh estimation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1496–1505, 2022. 58, 59

[113] Paritosh Parmar and Brendan Morris. Action quality assessment across multiple actions. In *WACV*, 2019. 3

[114] Paritosh Parmar and Brendan Tran Morris. Learning to score olympic events, 2017. 8

[115] Paritosh Parmar and Brendan Tran Morris. What and how well you performed? a multitask learning approach to action quality assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 304–313, 2019. 8

[116] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022. 39

[117] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 724–732, 2016. 35, 39

[118] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 39

[119] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *CVPR*, 2012. 3

[120] H. Pirsiavash, C. Vondrick, and A. Torralba. Assessing the quality of actions. In *ECCV*, 2014. 3, 8

[121] Antonio Polino, Razvan Pascanu, and Dan Alistarh. Model compression via distillation and quantization. *arXiv preprint arXiv:1802.05668*, 2018. 8

[122] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Luk*á*vs. Burget, Ondrej Glembek, Nagendra Kumar Goel, Mirko Hannemann, Petr Motlícek, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer, and Karel Veselý. The kaldi speech recognition toolkit. 2011. 47

[123] Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou,

Rama Chellappa, and Pengchuan Zhang. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5285–5297, 2023. 6, 26, 43, 44

[124] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 39

[125] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023. 27

[126] Francesco Ragusa, Antonino Furnari, Salvatore Livatino, and Giovanni Maria Farinella. The meccano dataset: Understanding human-object interactions from egocentric videos in an industrial-like domain. In *WACV*, 2021. 3, 7

[127] N. Rai, H. Chen, J. Ji, R. Desai, K. Kozuka, S. Ishizaka, E. Adeli, and J.C. Niebles. Home action genome: Contrastive compositional action understanding. In *CVPR*, 2021. 2, 3

[128] Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3d dataset (HM3d): 1000 large-scale 3d environments for embodied AI. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. 3

[129] Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014. 39

[130] Krishna Regmi and Ali Borji. Cross-view image synthesis using conditional gans. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 4, 7

[131] Krishna Regmi and Ali Borji. Cross-view image synthesis using geometry-guided conditional gans. *Computer Vision and Image Understanding*, 2019. 4

[132] Krishna Regmi and Mubarak Shah. Bridging the domain gap for ground-to-aerial image matching. In *ICCV*, 2019. 4

[133] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *ICCV*, 2021. 3

[134] Bin Ren, Hao Tang, and Nicu Sebe. Cascaded cross mlp-mixer gans for cross-view image translation. *arXiv preprint arXiv:2110.10183*, 2021. 4

[135] Xuanchi Ren and Xiaolong Wang. Look outside the room: Synthesizing a consistent long-term 3d scene video from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3563–3573, 2022. 4

[136] Helge Rhodin, Christian Richardt, Dan Casas, Eldar Insafutdinov, Mohammad Shafiei, Hans-Peter Seidel, Bernt Schiele, and Christian Theobalt. Egocap: egocentric marker-less motion capture with two fisheye cameras. *ACM Transactions on Graphics (TOG)*, 35(6):1–11, 2016. 9

[137] Robin Rombach, Patrick Esser, and Björn Ommer. Geometry-free view synthesis: Transformers and no 3d priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14356–14366, 2021. 4

[138] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 59

[139] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21096–21106, 2022. 2, 3, 8, 7

[140] Pierre Sermanet, Corey Lynch, Jasmine Hsu, and Sergey Levine. Time-contrastive networks: Self-supervised learning from multi-view observation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 486–487. IEEE, 2017. 43

[141] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, and Sergey Levine. Time-contrastive networks: Self-supervised learning from video. *Proceedings of International Conference in Robotics and Automation (ICRA)*, 2018. 4

[142] Xi Shen, Alexei A Efros, Armand Joulin, and Mathieu Aubry. Learning co-segmentation by segment swapping for retrieval and discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5082–5092, 2022. 35, 36

[143] Nina Shvetsova, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogerio S Feris, David Harwath, James Glass, and Hilde Kuehne. Everything at once-multi-modal fusion transformer for video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20020–20029, 2022. 4

[144] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *CVPR*, 2018. 4, 7, 43

[145] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*, 2018. 2, 3, 7

[146] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1145–1153, 2017. 9

[147] Krishna Kumar Singh, Kayvon Fatahalian, and Alexei A. Efros. Krishnacam: Using a longitudinal, single-person, egocentric dataset for scene understanding tasks. In *WACV*, 2016. 3

[148] Yale Song, Eugene Byrne, Tushar Nagarajan, Huiyu Wang, Miguel Martin, and Lorenzo Torresani. Ego4d goal-step:

Toward hierarchical understanding of procedural activities. In *NeurIPS*, 2023. 7

[149] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human action classes from videos in the wild. In *CRCV-TR-12-01*, 2012. 2, 42

[150] Bilge Soran, Ali Farhadi, and Linda Shapiro. Generating notifications for missing actions: Don't forget to turn the lights off! In *ICCV*, pages 4669–4677, 2015. 8

[151] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 3

[152] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, pages 240–248. Springer, 2017. 39

[153] Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, and Joel S Emer. How to evaluate deep neural network processors: Tops/w (alone) considered harmful. *IEEE Solid-State Circuits Magazine*, 2020. 46

[154] Mukund Varma T, Peihao Wang, Xuxi Chen, Tianlong Chen, Subhashini Venugopalan, and Zhangyang Wang. Is attention all that neRF needs? In *The Eleventh International Conference on Learning Representations*, 2023. 39

[155] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. Mnasnet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2820–2828, 2019. 8

[156] Shuhan Tan, Tushar Nagarajan, and Kristen Grauman. Egodistill: Egocentric head motion distillation for efficient video understanding. *NeurIPS*, 2023. 8

[157] Hao Tang, Dan Xu, Nicu Sebe, Yanzhi Wang, Jason J Corso, and Yan Yan. Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2417–2426, 2019. 4, 7

[158] Hao Tang, Kevin Liang, Kristen Grauman, Matt Feiszli, and Weiyao Wang. Egotracks: A long-term egocentric visual object tracking dataset. *Advances in Neural Information Processing Systems*, 2023. 6

[159] Yansong Tang, Jiwen Lu, and Jie Zhou. Comprehensive instructional video analysis: The coin dataset and performance evaluation. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 2, 4, 7

[160] Zachary Teed and Jia Deng. DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras. *Advances in neural information processing systems*, 2021. 57

[161] Denis Tome, Patrick Peluse, Lourdes Agapito, and Hernan Badino. xr-egopose: Egocentric 3d human pose from an hmd camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 9

[162] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. 2022. 43

[163] Vadim Tschernezki, Ahmad Darkhalil, Zhifan Zhu, David Fouhey, Iro Larina, Diane Larlus, Dima Damen, and Andrea Vedaldi. EPIC Fields: Marrying 3D Geometry and Video Understanding. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2023. 3

[164] Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhib Alsisan, Jia-Bin Huang, and Johannes Kopf. Consistent view synthesis with pose-guided diffusion models. *arXiv preprint arXiv:2303.17598*, 2023. 4

[165] Pavan Kumar Anasosalu Vasu, James Gabriel, Jeff Zhu, Oncel Tuzel, and Anurag Ranjan. Mobileone: An improved one millisecond mobile backbone. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7907–7917, 2023. 8

[166] Sara Vicente, Carsten Rother, and Vladimir Kolmogorov. Object cosegmentation. In *CVPR 2011*, pages 2217–2224. IEEE, 2011. 6

[167] Hanlin Wang, Yilu Wu, Sheng Guo, and Limin Wang. Pdpp: Projected diffusion for procedure planning in instructional videos. *arXiv preprint arXiv:2303.14676*, 2023. 4

[168] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022. 4

[169] Daniel Weinland, Remi Ronfard, and Edmond Boyer. Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding (CVIU)*, 2006. 3, 7

[170] Yangming Wen, Krishna Kumar Singh, Markham Anderson, Wei-Pang Jan, and Yong Jae Lee. Seeing the unseen: Predicting the first-person camera wearer's location and pose in third-person scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 3446–3455, 2021. 4

[171] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7467–7477, 2020. 4

[172] Benita Wong, Joya Chen, You Wu, Stan Weixian Lei, Dongxing Mao, Difei Gao, and Mike Zheng Shou. Assistq: Affordance-centric question-driven task completion for egocentric assistant. In *European Conference on Computer Vision*, 2022. 3

[173] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In

Computer Vision and Pattern Recognition, IEEE Conference on, 2015. 3

[174] Zuxuan Wu, Tushar Nagarajan, Abhishek Kumar, Steven Rennie, Larry S Davis, Kristen Grauman, and Rogerio Feris. Blockdrop: Dynamic inference paths in residual networks. In *CVPR*, 2018. 8

[175] Fei Xia, Amir R. Zamir, Zhi-Yang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson Env: real-world perception for embodied agents. In *CVPR*. IEEE, 2018. Gibson license is available at `http://svl.stanford.edu/gibson2/assets/GDS_agreement.pdf`. 3

[176] J. Xiao, A. Owens, and A. Torralba. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *ICCV*, 2013. 3

[177] Frank F Xu, Lei Ji, Botian Shi, Junyi Du, Graham Neubig, Yonatan Bisk, and Nan Duan. A benchmark for structured procedural knowledge extraction from cooking videos. *arXiv preprint arXiv:2005.00706*, 2020. 8

[178] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pretraining for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021. 4

[179] Mingze Xu, Chenyou Fan, Yuchen Wang, Michael S Ryoo, and David J Crandall. Joint person segmentation and identification in synchronized first-and third-person videos. In *ECCV*, 2018. 4

[180] Mingze Xu, Yuanjun Xiong, Hao Chen, Xinyu Li, Wei Xia, Zhuowen Tu, and Stefano Soatto. Long short-term transformer for online action detection. *Advances in Neural Information Processing Systems*, 34:1086–1099, 2021. 49

[181] Weipeng Xu, Avishek Chatterjee, Michael Zollhoefer, Helge Rhodin, Pascal Fua, Hans-Peter Seidel, and Christian Theobalt. Mo 2 cap 2: Real-time mobile 3d motion capture with a cap-mounted fisheye camera. *IEEE transactions on visualization and computer graphics*, 25(5):2093–2101, 2019. 9

[182] Zihui Xue and Kristen Grauman. Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment. In *NeurIPS*, 2023. 4, 7

[183] Lita Yang, Robert M Radway, Yu-Hsin Chen, Tony F Wu, Huichu Liu, Elnaz Ansari, Vikas Chandra, Subhasish Mitra, and Edith Beigné. Three-dimensional stacked neural network accelerator architectures for ar/vr applications. *IEEE Micro*, 2022. 46

[184] Huangyue Yu, Minjie Cai, Yunfei Liu, and Feng Lu. What i see is what you see: Joint attention learning for first and third person video co-analysis. In *ACM MM*, 2019. 4

[185] Huangyue Yu, Minjie Cai, Yunfei Liu, and Feng Lu. First- and third-person video co-analysis by learning spatial-temporal joint attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 4

[186] Xumin Yu, Yongming Rao, Wenliang Zhao, Jiwen Lu, and Jie Zhou. Group-aware contrastive regression for action quality assessment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7919–7928, 2021. 8

[187] Ye Yuan and Kris Kitani. 3d ego-pose estimation via imitation learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 9

[188] Ye Yuan and Kris Kitani. Ego-pose estimation and forecasting as real-time pd control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 9

[189] Chenlin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. *arXiv preprint arXiv:2202.07925*, 2022. 21

[190] Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, pages 492–510, 2022. 54, 55, 57

[191] Qiang Zhang and Baoxin Li. Relative hidden markov models for evaluating motion skill. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 8

[192] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 39

[193] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *ECCV*, 2022. 3

[194] Shiyi Zhang, Wenxun Dai, Sujia Wang, Xiangwei Shen, Jiwen Lu, Jie Zhou, and Yansong Tang. Logo: A long-form video dataset for group action quality assessment. In *CVPR*, 2023. 3

[195] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856, 2018. 8

[196] He Zhao, Isma Hadji, Nikita Dvornik, Konstantinos G Derpanis, Richard P Wildes, and Allan D Jepson. P3iv: Probabilistic procedure planning from instructional videos with weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2938–2948, 2022. 4

[197] Weixi Zhao, Weiqiang Wang, and Yunjie Tian. Graformer: Graph-oriented transformer for 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20438–20447, 2022. 58

[198] Yue Zhao and Philipp Krähenbühl. Real-time online video detection with temporal smoothing transformers. In *European Conference on Computer Vision*, 2022. 49

[199] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6586–6597, 2023. 26, 46, 47, 48

[200] Ce Zheng, Xianpeng Liu, Guo-Jun Qi, and Chen Chen. Potter: Pooling attention transformer for efficient human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1611–1620, 2023. 59

[201] Yiwu Zhong, Licheng Yu, Yang Bai, Shangwen Li, Xueting Yan, and Yin Li. Learning procedure-aware video representation from instructional videos and their narrations. *arXiv preprint arXiv:2303.17839*, 2023. 4

[202] Honglu Zhou, Roberto Martin-Martin, Mubbasir Kapadia, Silvio Savarese, and Juan Carlos Niebles. Procedure-aware pretraining for instructional video understanding. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 4, 8

[203] Honglu Zhou, Roberto Martín-Martín, Mubbasir Kapadia, Silvio Savarese, and Juan Carlos Niebles. Procedure-aware pretraining for instructional video understanding. In *CVPR*, pages 10727–10738, 2023. 49

[204] L. Zhou, N. Louis, and J. Corso. Weakly-supervised video object grounding from text by loss weighting and object interaction. In *BMVC*, 2018. 2, 4

[205] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2018. 4, 7

[206] Michael Zhu and Suyog Gupta. To prune, or not to prune: exploring the efficacy of pruning for model compression. *arXiv preprint arXiv:1710.01878*, 2017. 8

[207] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 2, 4, 7

[208] Aneeq Zia, Yachna Sharma, Vinay Bettadapura, Eric L. Sarin, and Irfan A. Essa. Video and accelerometer-based motion analysis for automated surgical skills assessment. *CoRR*, abs/1702.07772, 2017. 8