Fine-grained Activities of People Worldwide

Jeffrey Byrne[1], Greg Castañón[2], Zhongheng Li[2], and Gil Ettinger[2]

[1]Visym Labs, Cambridge MA, USA
[2]Systems & Technology Research, Woburn MA, USA

Abstract

Every day, humans perform many closely related activities that involve subtle discriminative motions, such as putting on a shirt vs. putting on a jacket, or shaking hands vs. giving a high five. Activity recognition by ethical visual AI could provide insights into our patterns of daily life, however existing activity recognition datasets do not capture the massive diversity of these human activities around the world. To address this limitation, we introduce Collector, a free mobile app to record video while simultaneously annotating objects and activities of consented subjects. This new data collection platform was used to curate the Consented Activities of People (CAP) dataset, the first large-scale, fine-grained activity dataset of people worldwide. The CAP dataset contains 1.45M video clips of 512 fine grained activity labels of daily life, collected by 780 subjects in 33 countries. We provide activity classification and activity detection benchmarks for this dataset, and analyze baseline results to gain insight into how people around with world perform common activities. The dataset, benchmarks, evaluation tools, public leaderboards and mobile apps are available for use at https://visym.github.io/cap.

1. Introduction

Large scale activity recognition has made remarkable progress driven by the curation of large scale labeled video datasets [40, 73, 33, 1, 2, 74, 45, 9, 21, 36, 59, 35, 13, 43]. Evaluation tasks in these datasets include activity classification, activity and object detection and localization, action prediction, episodic memory for object instance retrieval, object interactions with hands/tools/people, speaker prediction and scene diarization in long duration videos.

However, performance on these important tasks remains limited by the scale, quality and applicability of data. While there are many large-scale video datasets for pretraining activity recognition such as Kinetics [41], AVA [28], Moments in Time [59][60], ActivityNet [8], YouTube-8M [1], HVU [19] and IG65M [24], these datasets are all scraped from social media platforms such as YouTube or Instagram. These datasets are easy to collect, but suffer from terms of service restrictions, non-consented subjects and link rot, making reproducible research difficult. Furthermore, the labels in these datasets sparsely sample fine-grained activities, and instead represent activities that are interesting enough for social media. Recent dataset collections efforts have transitioned to actors performing scripted [35][65] or unscripted [20] activities to introduce more diversity of the activities that we all perform every day, however these datasets have limited scale for supervised training.

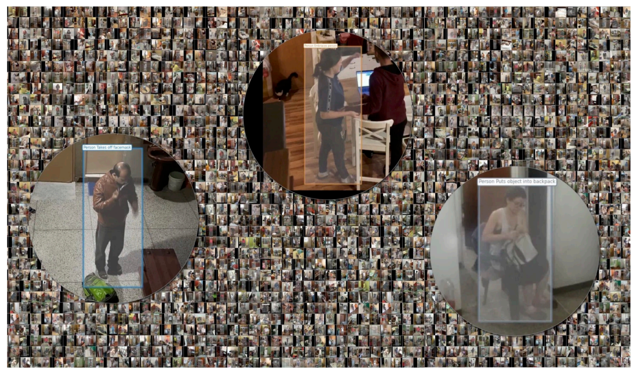

Figure 1. The Diversity of Human Activities. Humans perform a wide variety of closely related activities that involve subtle motions performed alone, while interacting with objects or with other people. The CAP dataset was designed to explore the representation of these fine grained activities of daily life around the world.

In this paper, we introduce the Consented Activities of People (CAP) dataset, a fine grained dataset of activities of daily life for visual AI research. Humans perform a wide variety of closely related activities that involve simple yet subtle motions that we perform alone, interacting with objects or interacting with other people. For example, figures 1 and 2 show examples of activities that we may perform every day: putting on a face mask, putting an object into a

Consented Activities of People (CAP)	
Clips	1454540
Fine Activity labels	512
Coarse Activity labels	144
Collectors	780
Countries	33
Cities	104
Mobile devices	355
Object labels	157

Figure 2. The Consented Activities of People Dataset, collected on-demand from consented subjects, recorded worldwide from third-person viewpoints, of fine-grained activities of daily life and submitted from handheld and rigid mobile devices. Available at visym.com/cap.

backpack or hugging another person. The CAP dataset was designed to explore the representation and recognition of fine grained activities of daily life, using open data collected on-demand from consented subjects and recorded worldwide from third-person viewpoints. Specifically, the CAP dataset contains:

- **Common activities** that we all perform each day, such as dressing or grooming that are not typically captured on video because they are rarely performed in front of a camera or are too boring to share.
- **Fine activities** that are closely related and may be easily confused, such as putting on socks vs. putting on shoes or talking on a phone vs. smoothing your hair.
- **Diverse activities** that are different ways of performing the same activity in the wild, such as activities viewed from behind or interacting with different objects.

In order to collect a large-scale visual dataset of the diversity of human activities, we introduce the *Collector platform*. Collector is a global platform for collecting consented datasets of people for visual AI applications. Collector is able to record, annotate and verify video datasets, collected with geographically diversity of people around the world.

The primary contributions of this work are:

- **Collector platform**. Section 3 describes the new platform developed to collect ethical datasets of people. This platform can be used by the research community to collect new on-demand visual datasets as easily as recording a video.
- **Consented Activities of People (CAP) dataset**. Section 4 describes the collected dataset of fine-grained activities of consented people worldwide. The dataset contains annotated videos of fine-grained activities with bounding box tracks and temporal localization.
- **Benchmark suite**. Section 5 describes the open benchmarks and baselines on this dataset, along with results and analysis in Section 6.

2. Related Work

The evolution of video datasets has progressed from a small number of classes and actors in trimmed videos [69, 6] to large-scale web video on social media [40, 73, 33, 1, 2, 74, 45, 9, 21, 36, 59]. Keyword-based search from YouTube or Instagram enabled weak labeling of videos with minimal curation, creating datasets that recorded a large set of people doing a small set of activities. The diversity and volume of video available on social media lead to massive datasets for pretraining. Recently, efforts have bootstrapped classifiers to improve the scalability of their annotation and collection efforts from noisy web video [78]. Furthermore, approaches have attempted to directly mitigate the geographic biases of web video by scraping from local versions of websites [64].

These large datasets are easy to curate, but the contents have limited diversity, as the joint combination of viewpoints (e.g. exocentric, egocentric) and activity labels (e.g. dressing, eating) that are common in real scenes are not as common on social media. Centralized collection of actors [14, 65, 34], as well as crowdsourced approaches [71][10][27][65][34] have been used to generate datasets of labels and perspectives not densely sampled in social video, but are limited in the diversity and scale of training data. This style of dataset collection has specialized further into diagnostic datasets [37][76][25][3][22] that attempt to answer a specific question about performance bias, as well as fine-grained datasets which attempt to densely sample the space of actions in a specific domain [53, 61, 49, 49, 63, 68, 57, 38].

Table 1 shows a quantitative comparison of these related datasets. This comparison table focuses on egocentric, ex-

Dataset	Year	Domain	Fine Classes	Coarse Classes	Annotation	Clips	Mean Clips/Class
ActivityNet (v3) [8]	2016	Social	200	73	T	23.1K	137
Charades [71][70]	2016	Exo|Ego	-	157	T|N	68.5K	-
Something-Som... [26]	2017	Ego	-	174	T	220.8K	600
PKU-MMD [13]	2017	Exo|M	-	51	T|J	21.5K	-
Kinetics-700 [41]	2017	Social	-	700	T	650K	700
Youtube-8M (Seg) [1]	2018	Social	-	1000	T	237K	150
EPIC-Kitchens [16]	2018	Ego	97	13	T|N|C	90K	-
HACS (clips) [78]	2019	Social	-	200	T	1.55M	1100
MMAct [43]	2019	Exo|Ego|M	-	37	S|T|J	40K	-
LEMMA [35]	2020	Exo|Ego|M	863	24	S|T	11.8K	-
AVA-Kinetics [48]	2020	Social	-	60	S|T	230K	235
HVU (Actions) [19]	2020	Social	-	739	T	479.5K	2112
Moments in Time [60]	2020	Social	-	292	T	2.01M	6432
MEVA [14]	2021	Exo	-	37	S|T|E|C	35K	-
HOMAGE [65]	2021	Exo|Ego|M	453	70	S|T	24.6K	-
Ego4D (MQ) [20]	2021	Ego	-	110	S|T|N|G|C	22.2K	-
CAP	2022	Exo	512	144	S|T|N|G|C	1.45M	2880 (4501, top-250)

Table 1. CAP dataset comparison. Domains are egocentric (ego) from a first person viewpoint such as a head or body mounted camera, exocentric (exo) from a third-person viewpoint such as from a wall or building mounted camera, (social) videos scraped from online social media sources and (M)ulti-modal domains such as RGB-D, NIR, multiple viewpoints or additional non-visual sensors. Annotation ground truth considers combinations of: (T)emporal activity labels for start and end times, (S)patial object labels of bounding boxes around actors or interacted objects, (E)xtrinsic camera poses with calibrated relative position and orientation, (G)eographic locations for each video, (J)oint keypoints of human pose skeletons, (N)atural language captions or narrations and (C)onsented subjects for ethical video recording.

ocentric and social datasets for activity classification and detection tasks, comparing the number of classes, clips and mean clips per class. This shows that our Consented Activities of People (CAP) dataset is the largest consented activity dataset collected to date as measured by mean number of training clips per class.

3. Collector Platform

Collector is a new platform for visual dataset curation that was designed to address the limits of current collection strategies. The traditional approach to construction of visual dataset of people is to: (i) Set up camera networks to record videos and imagery, (ii) Gather a set of subjects who have consented to have their personally identifiable information (PII) recorded and shared for an authorized purpose and duration, (iii) Record videos of these IRB approved consented subjects only, and no one else, (iv) Send videos to an annotation team to manually search videos for ground truth labels, (v) Send the annotations to a verification team to enforce quality. This approach is slow end expensive.

There is a need for a new dataset collection approach that is *on-demand, worldwide and cost-efficient*. On-demand approaches enable an agile, adaptive collection of instances that are engineered to introduce diversity of labels or attributes such as pose, illumination or object interaction and mitigate biases. Furthermore, access to data sources from many countries and cultures avoids an imbalance of data from a specific region of the world and its implicit biases. Finally, the approach needs to be relatively cost-efficient to collect large-scale training data.

Collector is a global platform for collecting large scale consented video datasets of people for visual AI applications. Collector is able to record, annotate and verify custom video datasets of rarely occurring activities for training visual AI systems. The Collector platform provides:

- On-demand collection of rarely occurring activities from thousands of collectors worldwide.
- Simultaneous video recording, annotation and verification into a single unified platform.
- Touchscreen UI for live annotation of bounding boxes, activity clips and object categories.
- Specification of required collection attributes such as pose, illumination, location or object interactions.
- IRB approved informed consent for ethical dataset construction with in-app face anonymization.

Figure 3 shows an overview of the collector workflow. Collectors are invited onto the platform, and they download the collector mobile app to their device. Collectors are presented *collections* which are video collection tasks grouped by required objects (e.g. a car, another person) or locations (e.g. parking lot, dining room). Each collection specifies the requirements of the submitted video, which include required activities, objects, location, illumination conditions, actor pose and camera viewpoint. Once a collector chooses a collection to record, they get consent from their subject,

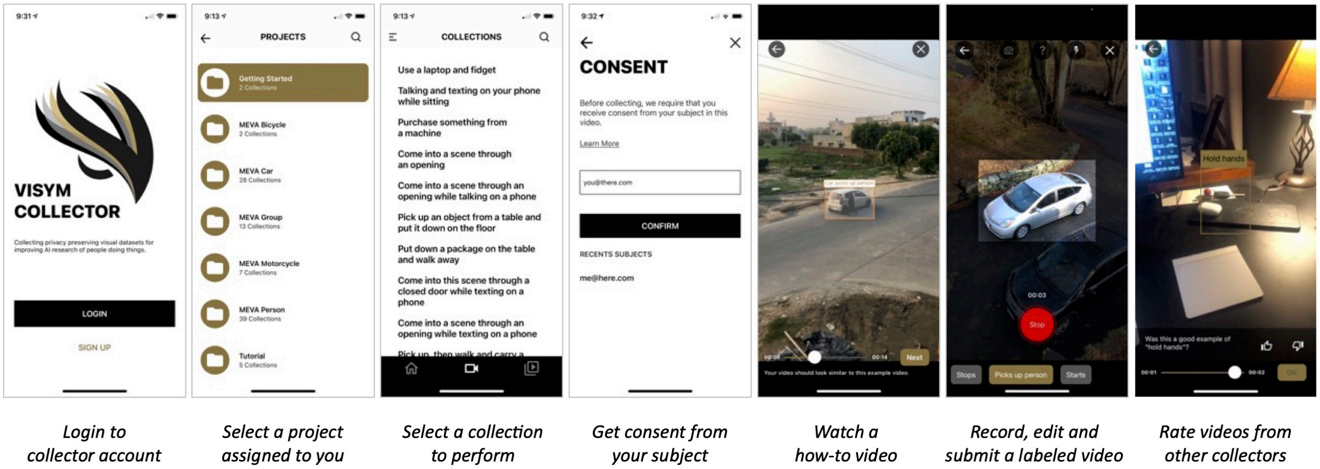

Figure 3. The Collector platform curates visual datasets of people by enabling thousands of collectors worldwide to record and submit videos using a mobile app. This workflow shows the mobile interface for collecting on-demand video datasets of people.

including a video recording to ensure that the person consenting is the person being recorded. Next, the collector watches an example video which shows a gold standard exemplar of the collection. We use visual exemplars to bypass language issues and communicate an idea of what the collection should look like. Finally, the collector records and annotates the video live using touch gestures on their device, optionally corrects errors using an in-app annotation editor and submits the annotated collection for review.

The Collector mobile app has been downloaded by thousands of freelance collectors worldwide, and is freely available in the iOS and Android app stores. Appendix A provides more information on mobile app for recording and annotation (§A.1), campaign dashboard for global coordination (§A.2) and human review for annotation quality (§A.3).

4. Consented Activities of People Dataset

The Consented Activities of People (CAP) dataset is a fine grained visual dataset of the activities of daily life, curated using the Collector platform. Humans perform a wide variety of closely related activities every day that are subtle, localized and socially informative. The CAP dataset was designed to explore the problem of representation of simple, fine-grained activities and provide a benchmark to characterize performance for classification and detection of these closely related activities.

How do we define the set of labels in a dataset of fine-grained activities? What exactly is a fine-grained activity? The discussion of this question in appendix B.3 suggests that a fine-grained activity is defined relative to other activities and should specify the following:

- **Who?** Fine-grained activity labels should be performed by the same noun (e.g. Person).
- **What?** Fine-grained activity labels should include simple verbs that can be performed in a few seconds along with other "closely related" verbs.
- **With?** Fine-grained activity labels should include object interactions that induce a visually distinct motion.
- **How?** Fine-grained activities should include visually grounded styles as within class variation.

Label expansion. In order to downselect labels that satisfy these criteria, we perform a new strategy called *label expansion*. Label expansion starts from the source labels in AVA [28], Charades [71], Moments in Time [59][60], Kinetics-700 [41], Something-Something [26] and MEVA [14]. We augment this set with the Activities of Daily Living [62][47]. Next, we remove activities that are complex, non-visually grounded, non-person centered, not commonly performed around the house, or require skilled execution. The remaining verbs are label expansion candidates.

We perform label expansion by selecting one or more closely related verbs and nouns for each label candidate that satisfy the CAP design goals. We are all experts when it comes to understanding the subtle discrimination between gestures, social interactions or simple activities that we perform every day. Therefore, the collector team leveraged their social expertise to manually perform label expansion for each candidate label. For example, closely related verbs *person puts on socks* to *person puts on shoes* or closely related object interactions with different appearances *person puts on shoes* to *person puts on hat*.

The result of the label expansion is is shown in figures 4 and appendix B.22. Appendix figure B.22 shows a circular tree plot of the hierarchical organization of the fine-grained labels grouped by "Noun Verb" structure, such as *person*

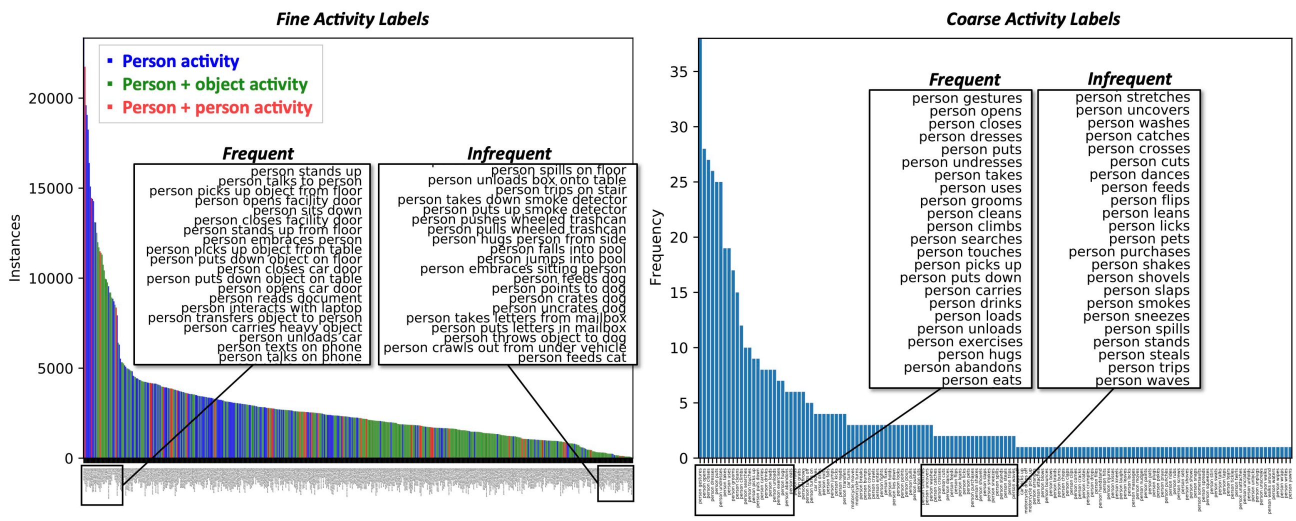

Figure 4. CAP label distribution. (left) Instance histogram for fine-grained categories, colored by person-only, person-person or person-object interactions showing the most and least common labels by frequency, (right) Fine-grained histogram for each coarse-grained category to show the number of fine-grained categories in each hierarchical grouping. Figure B.22 shows the full hierarchical label set.

dresses or *person gestures* into a two level, tree structured hierarchy.

Collection Campaign. The CAP campaign was set up to run on the Collector platform during the period of Apr 2020 to Dec 2021. The CAP dataset was collected in two stages, Apr 2020 - Mar 2021 which focused on collection of MEVA activity classes [14] and July - Dec 2021 which focused on the remaining CAP activity classes. The campaign specification includes 842 unique collection types, each specifies one of 512 activity labels and 157 object types. In total, 288/842 collection types were specified so that the subject is facing away from the camera to increase diversity, 87/842 collections were specified to be collected to support temporal activity detection and 38/842 collection types were physically stabilized. The overall collection statistics are shown in figure 2, such that 905,369 clips are for activity classification (AC) train/val, 132,271 clips for AC sequestered test and 416,900 clips for activity detection (AD). Figure 4 shows the overall label frequency. Note that this histogram is unbalanced due to frequent organic activities, such as *person sits down* which often precedes object interactions.

The appendix discusses the key challenges (§B.3), dataset design goals (§B.2), collection methodology (§B.4), distribution format (§B.5) and visualizations (Figure B.19, B.21) for curating a large scale dataset of daily activities.

5. Benchmark Suite

Performance benchmarking is the specification of an evaluation methodology, task, dataset and a baseline system design to evaluate system performance. Typical benchmarking considers test data that is in-domain, meaning it is collected and annotated exactly as it will be used in practice. However, consider the challenge of benchmarking fine-grained activity recognition in third-person security video. We may collect many hours of video from many security cameras, without ever collecting an organic instance of a fine-grained target label like *person puts down backpack*. If our goal is to benchmark performance for rarely occurring activities, then how do we benchmark in practice when the labels to evaluate almost never occur?

We address this key challenge by introducing *domain adjacent benchmarking*. In this strategy, we collect test sets that are from the required viewpoint, but with actors performing the test activities in short bursts. This provides performance evaluation of a target domain (e.g. third person, long duration videos, organic activities) in a closely related adjacent domain (e.g. third person, short duration videos, actors). The test data in the adjacent domain can be collected and distributed ethically, and performance evaluation on the domain adjacent data is used as a surrogate for the target domain. Further discussion of the implicit biases in this strategy is provided in appendix B.7.

5.1. Evaluation Tasks

Activity Classification (AC). The Activity Classification (AC) task is to assign one or more activity class labels and confidence scores to each video clip from a set of predefined classes. The metric for AC performance is Mean Average Precision (mAP), top-1 and top-5 classification performance averaged over all classes.

The AC task is separated into two domains, AC (Handheld) and AC (Stabilized). AC (Handheld) is constructed using videos collected from handheld cameras, and AC

	Activity Classification (Handheld)			Activity Classification (Stabilized)			Activity Detection (Handheld)			Activity Detection (Rigid)		
Experiment	**mAP**	**Top-1**	**Top-5**	**mAP**	**Top-1**	**Top-5**	**mAP$_{(.2)}$**	**mAP$_{(.5)}$**	**mAP$_{(.8)}$**	**mAP$_{(.2)}$**	**mAP$_{(.5)}$**	**mAP$_{(.8)}$**
Handheld (Fine)	0.453	0.435	0.690	0.421	0.395	0.638	0.171	0.064	0.003	0.182	0.073	0.007
Stabilized (Fine)	0.341	0.302	0.555	0.448	0.423	0.674	0.113	0.044	0.002	0.193	0.079	0.005
Handheld (Coarse)	0.483	0.534	0.783	0.421	0.491	0.731	0.182	0.075	0.004	0.200	0.081	0.003
Stabilized (Coarse)	0.362	0.387	0.683	0.465	0.515	0.754	0.136	0.054	0.003	0.225	0.090	0.004
Handheld (Coarsened)	0.470	0.518	0.781	0.413	0.474	0.724	0.177	0.069	0.003	0.184	0.071	0.003
Stabilized (Coarsened)	0.345	0.370	0.662	0.451	0.499	0.755	0.112	0.043	0.002	0.195	0.076	0.003

Figure 5. CAP Benchmark Evaluation. This result shows the performance of six experimental systems (rows) on four evaluation tasks (columns). The experimental systems differ in the training set, such that Handheld|Stabilized refers to the handheld or background stabilized video data and Fine|Coarse|Coarsened refers the training set labels (e.g. Fine labels, Coarse labels, or Coarsened labels trained on fine labels, then transformed to coarse labels at test time). The evaluation tasks are Activity Classification|Activity Detection (§5.1) evaluated on Handheld|Stabilized|Rigid video subsets (e.g. handheld, software background stabilized or rigidly mounted video).

(Stabilized) is constructed by performing software background stabilization on AC (Handheld) videos. Appendix B.5 discusses this background stabilization algorithm with examples shown in figure B.18. The stabilization is used as a post-processing step to evaluate the domain mismatch of stabilized videos to rigidly mounted cameras.

Figures B.17 and B.18 show examples from the training set for the activity classification task. The videos show untrimmed clips which include repetitions of an activity performed multiple times in a row by a subject. The objective of the activity classification task is to specify a label for a three second trimmed clip containing one activity.

Temporal Activity Detection (AD). The Temporal Activity Detection (AD) task is to detect and temporally localize all activity instances in untrimmed video. The metric for AD performance is Mean Average Precision (mAP) at a fixed temporal intersection over union (IoU) of 0.2, 0.5 and 0.8.

The AD Task is separated into two collection domains, AD (Handheld) and AD (Rigid). AD (Handheld) is constructed from handheld cameras, and AD (Rigid) is constructed from rigidly mounted, unmoving cameras. This separation is designed to evaluate a system trained with software stabilization, and tested on rigid cameras.

Appendix figure B.21 shows eight sample videos in the activity detection task. This visualization shows seven frames extracted from a video on each row. Each video is from a specific collection scenario, as described in section B.4. Each scenario has a subject performing between 7 and 11 activities in a sequence that is chosen by the subject.

5.2. Baseline system

The baseline system for activity detection is based on activity classification of tracked cuboids [39]. The system operates by performing SORT tracking [5] of people and vehicles, using a framewise YOLO-v5 [67] object detector on 5Hz videos followed by spatiotemporal IoU track association. For each track above a minimum length ($> 1s$) and minimum confidence (> 0.2), define an activity cuboid proposal as the spatiotemporal sequence of bounding boxes for the object instance. The cuboid is split into three second proposals, with overlap (4 frames), with replicated boundary conditions for short tracks, dilated by a constant factor (1.2), cropped to maximum square shape preserving the centroid and resized to 16x4x224x224 (frames, channels,height, width). The cuboid is converted into a RGBA representation, with an alpha channel (A) encoding a binary mask for the tracked bounding box within the cuboid. Finally, we classify each cuboid proposal using a 3D-Resnet-50 [32] with softmax classification followed by a non-maximum suppression at temporal IoU ≥ 0.5.

	AC (Handheld)			AC (Δ Baseline)		
Ablation Experiment	**mAP**	**Top-1**	**Top-5**	**mAP**	**Top-1**	**Top-5**
No video augmentation	0.450	0.424	0.676	-0.004	-0.011	-0.014
Low collection diversity	0.451	0.416	0.668	-0.002	-0.019	-0.022
Top-100 collectors only	0.471	0.445	0.688	0.018	0.010	-0.002

Figure 6. CAP Ablation Study. We retrained the baseline system removing video augmentation, removing the "from behind" collection diversity or removing all but the top-100 collectors, then compared the relative performance to the baseline.

Baseline training is performed using uniform random weight initialization, cross-entropy focal loss [51], on 8 GPUs with minibatch size 256, ADAM optimization [42] and inverse class frequency instance weighting on CAP dataset until validation loss saturates. Data augmentation includes spatial mirroring and random clips by shifting ± 3 frames. The baseline system is GPU optimized, real-time, python only and available at github.com/visym/heyvi.

6. Performance Evaluation

In this section, we describe the benchmark results on the CAP dataset, and results on the Activities in Extended Video (ActEV) Sequestered Data Leaderboard (SDL) [14].

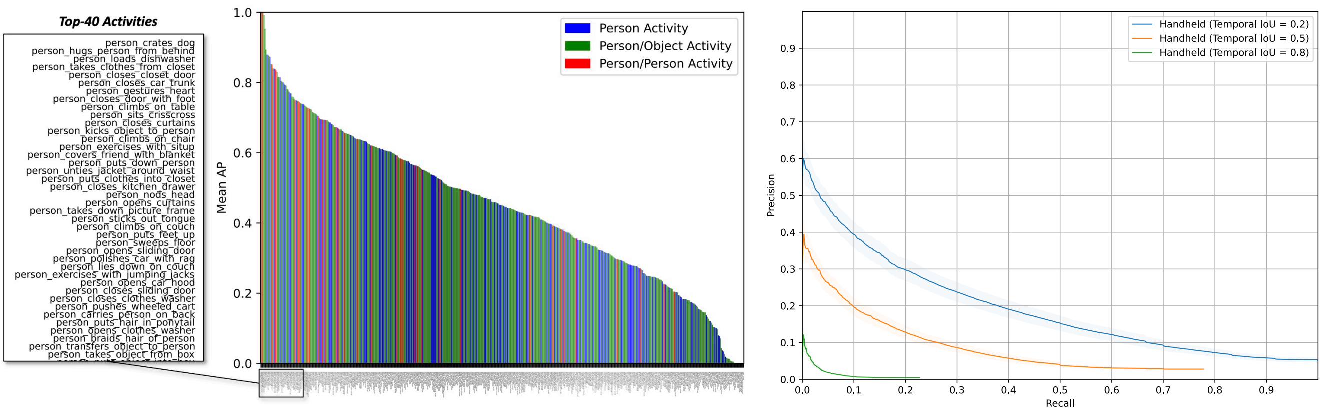

Figure 7. CAP Benchmark Evaluation Plots. (left) Activity Classification (Stabilized) performance per class, sorted in decreasing order by mean AP, colored by person, person/object or person/person interactions, showing top-40 classes (zoom into PDF for rest), (right) Activity Detection (Handheld) performance showing the mean precision recall at ground truth assignment IoU=0.2, 0.5 or 0.8, with 1σ error bars.

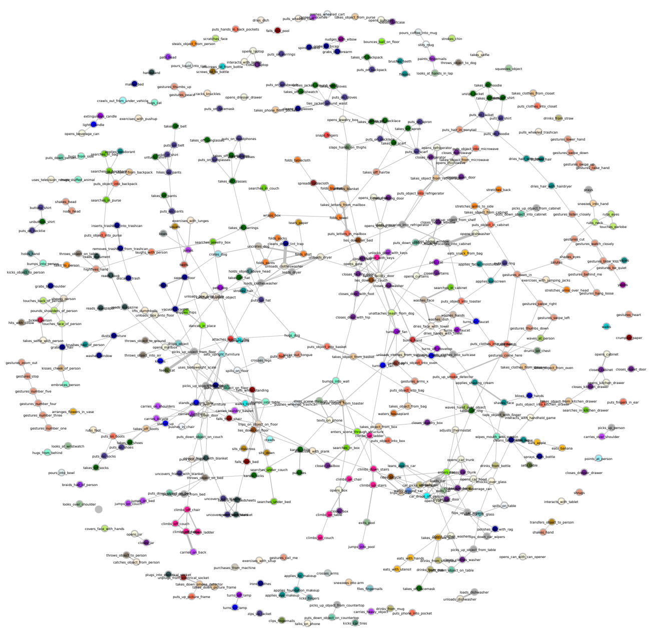

Figure 8. Confusion graph for activity classification showing edges connecting commonly confused fine-grained activity labels.

6.1. Benchmark Results

The experimental system runs the baseline with the following six combinations: Handheld, Stabilized or Rigid videos with Fine, Coarse or Coarsened label sets. *Handheld* refers to videos recorded directly from handheld mobile devices, *Stabilized* are the handheld videos with software background stabilization, and *Rigid* are test subset collected using rigidly mounted cameras. The baseline system is trained using either handheld or background stabilized videos, with Fine labels (e.g. figure B.22 outer), Coarse labels (e.g. figure B.22 inner), or *Coarsened* labels which is trained using fine labels, then mapped via lookup to the coarse label at test time. The benchmark datasets are the AC or AD sequestered test sets, subdivided into AC (Handheld), AC (Stabilized) and AD (Rigid) video subsets.

Figure 5 provides a benchmark evaluation on the activity classification and activity detection tasks. Results in figure 5 show: (i) background stabilization training helps for AD (Rigid), (ii) stabilized vs. handheld training exhibits a domain bias, (iii) AC (Handheld) performance is slightly better than AC (Stabilized) due to minor stabilization artifacts, (iv) AD is significantly more difficult than AC and (v) coarse labels are better than coarsened.

Figure 7 (left) shows activities on the AC (Stabilized) task ranked by mean AP per class and colored by object interaction type. This provides a deeper insight into the classes that are the highest and lowest performing. Results show that the best performing classes still leverage scene context (e.g. *crates dog, loads dishwasher*) and worst performing classes (mAP=0) and are poorly represented using the baseline system (e.g. *puts up smoke detector*). Figure 7 (right) shows an aggregate result on the AD (Handheld) task, which demonstrates that fine-grained activity detection at precise temporal localization (IoU=0.8) is challenging.

6.2. Benchmark Analysis

Figure 6 shows the results of an ablation study to understand the effect of three training set configurations on baseline performance. In all experiments, we renormalized the inverse class frequency weighting for the revised trainset, retrained the baseline system, then used the revised valset for model selection. First, we removed only the video augmentation (e.g. collectors performing activities multiple times), preserving all other data augmentation. Results show that relative baseline performance is lower, which demonstrates that video augmentation helps. Next, we removed only the "from behind" collections introduced for diversity. Relative performance for this trainset is lower, which shows that collection diversity helps. Finally, we kept only the videos from the top-100 collectors, comprising 65% of training set. Relative performance for top-100 trainset is higher, which suggests that for fine-grained activities (and our baseline system), it is better to have each collector perform many fine-grained activities.

Figure 8 shows a confusion graph of the AC task. This visualization shows a 2-d graph embedding constructed by transforming a confusion matrix to a graph adjacency matrix such that nodes are fine grained activity labels, node colors are coarse grained labels, and edge thickness corresponds to commonly confused fine-grained activities. A larger version is shown in appendix figure B.20.

Analysis of the confusion graph provides four insights. First, casual pairs (e.g. open and close) are commonly confused, since causal pairs often co-occur in a short temporal sequence. Second, we observe approximately one fourth of labels are not significantly confused, as shown by disconnected nodes. Third, there are small connected components with long range connections for common activities performed in sequence, such as interacting with drawers and cabinets in a kitchen. Finally, the three nodes that are most confused are *person trips on object on floor*, *person enters car* and *person opens facility door*, which suggests that improvement on these high degree labels should be prioritized.

6.3. ActEV SDL

The ActEV SDL is a sequestered data leaderboard for activity detection in long duration security videos. The ActEV SDL is labeled using the MEVA label set [14], which include 37 simple activities in security video. The MEVA labels are a subset of the CAP labels with five additional labels for vehicles turning, stopping and starting. We split the CAP dataset into a CAP-MEVA subset containing only the MEVA labels, which contains 405,781 background stabilized clips, split into 370K/35K train/val set. CAP-MEVA was used to re-train the baseline system, and compared results to training using MEVA only, which contained 35,022 training clips, as of when this analysis was performed.

Figure 9 shows an evaluation result on this dataset. The performance metric is mean probability of missed detection over activity classes vs. time based false alarm rate (TFA). We trained the baseline system using the MEVA dataset only or the union of CAP-MEVA and MEVA. We made four submissions to the ActEV SDL that differed only by the training set and validation set assumptions. Results show a 32% improvement at a fixed TFA=0.2 due only to training with the CAP-MEVA data, when controlling for the training hyperparameters and system configuration. Both green (MEVA + CAP-MEVA training) and red (MEVA only training) were trained from scratch rather than fine-tuned

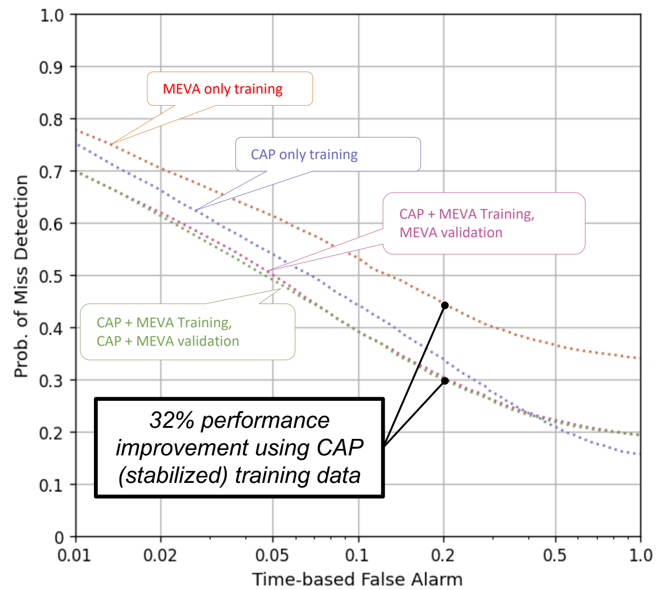

Figure 9. ActEV SDL Evaluation. (red) Trained with MEVA data only, (purple) trained with the union of MEVA and a CAP subset containing MEVA labels. Comparing the red/purple curves shows a 32% improvement using CAP data for identical systems.

starting from a pretrained model. All training data is background stabilized. This result shows that when controlling all other hyperparameters, the CAP dataset improves sequestered temporal AD performance in long duration video. This provides an independent validation of the CAP data for activity detection on static, long duration security video.

7. Conclusions

In this paper, we introduced the Consented Activities of People dataset, the largest fine grained activity dataset of people ever collected. Our benchmark provides a standardized evaluation of this new problem, with analysis to highlight the unique challenges of representing fine-grained activities. Finally, we believe that the Collector platform may be useful for the research community to address the never-ending demand for more ethical visual data.

Acknowledgement. Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/ Interior Business Center (DOI/IBC) contract number D17PC00344. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/IBC, or the U.S. Government. Dataset distribution supported by the AWS Open Data Sponsorship Program.

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016. 1, 2, 3

[2] Jean Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. Unsupervised learning from narrated instruction videos. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-Decem:4575–4583, 2016. 1, 2

[3] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems*, page 9448–9458, 2019. 2

[4] C. Berridge and T. Wetle. Why Older Adults and Their Children Disagree About In-Home Surveillance Technology, Sensors, and Tracking. *The Gerontologist*, 60(5), 2020. 19

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

[6] M Blank and L Gorelick. Actions as space-time shapes. *International Conference on Computer Vision*, 29(12):2247–2253, 2005. 2

[7] J. Byrne, B. Decann, and S. Bloom. Key-nets: Optical transformation convolutional networks for privacy preserving vision sensors. In *British Machine Vision Conference (BMVC)*, 2020. 19

[8] 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. 1, 3, 21

[9] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. *arXiv 1705.07750*, pages 6299–6308, 2017. 1, 2

[10] G. Castanon, N. Shnidman, T. Anderson, and J. Byrne. Out-the-window dataset for activity detection in security video. In *arXiv:1908.10899*, 2019. 2

[11] Chun-Fu Chen, Rameswar Panda, Kandan Ramakrishnan, Rogério Schmidt Feris, J. M. Cohn, Aude Oliva, and Quanfu Fan. Deep analysis of cnn-based spatio-temporal representations for action recognition. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6161–6171, 2021. 17

[12] Yue Chen, Yalong Bai, Wei Zhang, and Tao Mei. Destruction and construction learning for fine-grained image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:5152–5161, 2019. 15

[13] Liu Chunhui, Hu Yueyu, Li Yanghao, Song Sijie, and Liu Jiaying. Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding. *arXiv preprint arXiv:1703.07475*, 2017. 1, 3

[14] Kellie Corona, Katie Osterdahl, Roderic Collins, and Anthony Hoogs. Meva: A large-scale multiview, multimodal video dataset for activity detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1060–1068, January 2021. 2, 3, 4, 5, 6, 8, 19, 21

[15] Yin Cui, Feng Zhou, Jiang Wang, Xiao Liu, Yuanqing Lin, and Serge Belongie. Kernel pooling for convolutional neural networks. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-January:3049–3058, 2017. 15

[16] 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

[17] Adrien Deliege, Anthony Cioppa, Silvio Giancola, Meisam J. Seikavandi, Jacob V. Dueholm, Kamal Nasrollahi, Bernard Ghanem, Thomas B. Moeslund, and Marc Van Droogenbroeck. Soccernet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2021. 15

[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 21

[19] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Manohar Paluri, Jürgen Gall, Rainer Stiefelhagen, and Luc Van Gool. Large scale holistic video understanding. In *European Conference on Computer Vision*, pages 593–610. Springer, 2020. 1, 3

[20] K. Grauman et. al. Ego4d: Around the World in 3,000 Hours of Egocentric Video. *CoRR*, abs/2110.07058, 2021. 1, 3, 21

[21] David F. Fouhey, Wei Cheng Kuo, Alexei A. Efros, and Jitendra Malik. From Lifestyle Vlogs to Everyday Interactions. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 4991–5000, 2018. 1, 2

[22] Chuang Gan, Jeremy Schwartz, Seth Alter, Damian Mrowca, Martin Schrimpf, James Traer, Julian De Freitas, Jonas Kubilius, Abhishek Bhandwaldar, Nick Haber, Megumi Sano, Kuno Kim, Elias Wang, Michael Lingelbach, Aidan Curtis, Kevin Feigelis, Daniel M. Bear, Dan Gutfreund, David Cox, Antonio Torralba, James J. DiCarlo, Joshua B. Tenenbaum, Josh H. McDermott, and Daniel L. K. Yamins. Threedworld: A platform for interactive multi-modal physical simulation, 2020. 2

[23] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-December(2):317–326, 2016. 15

[24] Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. Large-scale weakly-supervised pre-training for video action recognition. *CVPR*, 2019. 1

[25] Rohit Girdhar and Deva Ramanan. Cater: A diagnostic dataset for compositional actions and temporal reasoning, 2019. 2

[26] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzýnska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. *arXiv*, pages 5842–5850, 2017. 3, 4

[27] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzyńska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The ”something something” video database for learning and evaluating visual common sense, 2017. 2, 21

[28] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6047–6056, 2018. 1, 4, 21

[29] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 21

[30] Albert Haque, Michelle Guo, Alexandre Alahi, Serena Yeung, Zelun Luo, Alisha Rege, Jeffrey Jopling, Lance Downing, William Beninati, Amit Singh, Terry Platchek, Arnold Milstein, and Li Fei-Fei. Towards vision-based smart hospitals: A system for tracking and monitoring hand hygiene compliance. In *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pages 75–87. PMLR, 18–19 Aug 2017. 19

[31] Albert Haque, Arnold Milstein, and Li Fei-Fei. Illuminating the dark spaces of healthcare with ambient intelligence. In *Nature (585)*, page 193–202, 2020. 19

[32] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6546–6555, 2018. 6

[33] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June:961–970, 2015. 1, 2

[34] Jingwei Ji, Ranjay Krishna Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10236–10247, 2020. 2, 19

[35] Baoxiong Jia, Yixin Chen, Siyuan Huang, Yixin Zhu, and Song-Chun Zhu. Lemma: A multiview dataset for learning multi-agent multi-view activities. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1, 3

[36] Yu Gang Jiang, Zuxuan Wu, Jun Wang, Xiangyang Xue, and Shih Fu Chang. Exploiting Feature and Class Relationships in Video Categorization with Regularized Deep Neural Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(2):352–364, 2018. 1, 2

[37] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning, 2016. 2

[38] Jonathan D. Jones, Cathryn Cortesa, Amy Shelton, Barbara Landau, Sanjeev Khudanpur, and Gregory D. Hager. Fine-grained activity recognition for assembly videos. *arXiv:2012.01392*, 2020. 2

[39] Vicky S. Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. Tubelet detector for spatio-temporal action localization. In *ICCV*, 2017. 6

[40] Andrej Karpathy and Thomas Leung. Large-Scale Video Classification with Convolutional Neural Networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2014. 1, 2, 15

[41] 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. 1, 3, 4, 21

[42] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015. 6

[43] Quan Kong, Ziming Wu, Ziwei Deng, Martin Klinkigt, Bin Tong, and Tomokazu Murakami. Mmact: A large-scale dataset for cross modal human action understanding. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 1, 3

[44] Shu Kong and Charless Fowlkes. Low-rank bilinear pooling for fine-grained classification. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-January:7025–7034, 2017. 15

[45] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. *arXiv*, pages 1–10, 2017. 1, 2

[46] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: A large video database for human motion recognition. *Proceedings of the IEEE International Conference on Computer Vision*, pages 2556–2563, 2011. 21

[47] Lawton, M. Powell, Brody, and Elaine M. Assessment of Older People: Self-Maintaining and Instrumental Activities of Daily Living. *The Gerontologist*, 9(3.1):179–186, 10 1969. 4

[48] Ang Li, Meghana Thotakuri, David A. Ross, João Carreira, Alexander Vostrikov, and Andrew Zisserman. The ava-kinetics localized human actions video dataset. *arXiv:2005.00214*, 2020. 3

[49] Yixuan Li, Lei Chen, Runyu He, Zhenzhi Wang, Gangshan Wu, and Limin Wang. Multisports: A multi-person video dataset of spatio-temporally localized sports actions, 2021. 2, 15

[50] Yixuan Li, Lei Chen, Runyu He, Zhenzhi Wang, Gangshan Wu, and Limin Wang. Multisports: A multi-person video dataset of spatio-temporally localized sports actions. *arXiv preprint arXiv:2105.07404*, 2021. 15

[51] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 2017. 6

[52] Tsung Yu Lin and Subhransu Maji. Improved bilinear pooling with CNNs. *British Machine Vision Conference 2017, BMVC 2017*, 2017. 15

[53] Yi Liu, Limin Wang, Xiao Ma, Yali Wang, and Yu Qiao. Fineaction: A fine-grained video dataset for temporal action localization, 2021. 2, 15

[54] Yi Liu, Limin Wang, Xiao Ma, Yali Wang, and Yu Qiao. Fineaction: A fine-grained video dataset for temporal action localization. *ICCV 2021 workshops*, 2021. 15

[55] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. *ECCV*, 2018. 17

[56] Nicole Martinez-Martin, Zelun Luo, Amit Kaushal, Ehsan Adeli Albert Haque, Sara S Kelly, Sarah Wieten, Mildred K Cho, David Magnus, Fei-Fei Li, Kevin Schulman, and Arnold Milstein. Ethical issues in using ambient intelligence in health-care settings. In *The Lancet*, December 21 2020. 19

[57] Mohammad Mahdi Kazemi Moghaddam, Ehsan Abbasnejad, and Javen Shi. Follow the attention: Combining partial pose and object motion for fine-grained action detection, 2019. 2

[58] Mohammad Moghimi, Mohammad Saberian, Jian Yang, Li Jia Li, Nuno Vasconcelos, and Serge Belongie. Boosted convolutional neural networks. *British Machine Vision Conference 2016, BMVC 2016*, 2016-September:24.1–24.13, 2016. 15

[59] 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. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):502–508, 2020. 1, 2, 4, 21

[60] Mathew Monfort, Bowen Pan, Kandan Ramakrishnan, Alex Andonian, Barry A Mcnamara, Alex Lascelles, Quanfu Fan, Dan Gutfreund, Rogerio Feris, and Aude Oliva. Multi-moments in time: Learning and interpreting models for multi-action video understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. 1, 3, 4

[61] E. Nicora, G. Goyal, N. Noceti, A. Vignolo, A. Sciutti, and F. Odone. The moca dataset, kinematic and multi-view visual streams of fine-grained cooking actions. *Scientific Data 7 (1), 1-15*, 2020. 2

[62] Linda S. Noelker and Richard Browdie. Sidney Katz, MD: A New Paradigm for Chronic Illness and Long-Term Care. *The Gerontologist*, 54(1):13–20, 08 2013. 4

[63] AJ Piergiovanni and Michael S. Ryoo. Fine-grained activity recognition in baseball videos. *CVPR Workshop on Computer Vision in Sports*, 2018. 2, 15

[64] AJ Piergiovanni and Michael S. Ryoo. Avid dataset: Anonymized videos from diverse countries. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[65] N. Rai, H. Chen, J. Ji, R. Desai, K. Kozuka, S. Ishizaka, Ehsan Adeli, and J.C. Niebles. Home action genome: Contrastive compositional action understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021)*, 2021. 1, 2, 3, 19

[66] Yongming Rao, Guangyi Chen, Jiwen Lu, and Jie Zhou. Counterfactual Attention Learning for Fine-Grained Visual Categorization and Re-identification. *arXiv:2108.08728*, pages 1025–1034, 2021. 15

[67] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016. 6

[68] 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, 2012. 2, 15

[69] C Schuldt, I Laptev, and B Caputo. Recognizing human actions: a local SVM approach. *Proceedings of the International Conference on Pattern Recognition*, pages 3–7, 2004. 2

[70] Gunnar A. Sigurdsson, Abhinav Kumar Gupta, Cordelia Schmid, Ali Farhadi, and Alahari Karteek. Actor and observer: Joint modeling of first and third-person videos. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7396–7404, 2018. 3

[71] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9905 LNCS:510–526, 2016. 2, 3, 4

[72] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer, 2016. 21

[73] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *arXiv preprint arXiv:1212.0402*, (November), 2012. 1, 2

[74] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-Decem:5288–5296, 2016. 1, 2

[75] Artem Babenko Yandex and Victor Lempitsky. Aggregating local deep features for image retrieval. *Proceedings of the IEEE International Conference on Computer Vision*,

2015 International Conference on Computer Vision, ICCV 2015:1269–1277, 2015. 15

[76] Kexin Yi*, Chuang Gan*, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. Clevrer: Collision events for video representation and reasoning. *ICLR*, 2020. 2

[77] Chaojian Yu, Xinyi Zhao, Qi Zheng, Peng Zhang, and Xinge You. Hierarchical Bilinear Pooling for Fine-Grained Visual Recognition. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11220 LNCS:595–610, 2018. 15

[78] Hang Zhao, Zhicheng Yan, Lorenzo Torresani, and Antonio Torralba. Hacs: Human action clips and segments dataset for recognition and temporal localization. *arXiv preprint arXiv:1712.09374*, 2019. 2, 3