CDAD: A Common Daily Action Dataset with Collected Hard Negative Samples

Wangmeng Xiang[1,2], Chao Li[2], Ke Li[2], Biao Wang[2], Xian-sheng Hua[2], Lei Zhang[1*]
[1]The Hong Kong Polytechnic University [2]DAMO Academy, Alibaba Group
{cswxiang,cslzhang}@comp.polyu.edu.hk
{lllcho.lc,keli.lk,wb.wangbiao,xiansheng.hxs}@alibaba-inc.com

Abstract

The research on action understanding has achieved significant progress with the establishment of various benchmark datasets. However, the results of action understanding are far from satisfactory in practice. One reason is that the existing action datasets ignore the existence of many hard negative samples in real-world scenarios, which are usually undefined confusion actions, e.g., holding a pen near the mouth vs. smoking. In this work, we focus on the common actions in our daily life and present a novel Common Daily Action Dataset (CDAD), which consists of 57,824 video clips of 23 well-defined common daily actions with rich manual annotations. Particularly, for each daily action, we collect not only diverse positive samples but also various hard negative samples that have minor differences (share similarities) in action with the positive ones. The established CDAD dataset could not only serve as a benchmark for several important daily action understanding tasks, including multi-label action recognition, temporal action localization, and spatial-temporal action detection, but also provide a testbed for researchers to investigate the influence of highly similar negative samples in learning action understanding models. Datasets and codes are available: https://github.com/MartinXM/CDAD.

1. Introduction

Action understanding is an important field in computer vision, which aims to understand the *what*, *when* and *where* of human actions. Action recognition, temporal action localization and spatial-temporal detection are the three important subtasks of action understanding. In recent years, many methods have been proposed for action understanding. For example, TSN [45], I3D [2], P3D [35], TSM [26], Slowfast [7] are proposed for efficient spatial-temporal action modeling; BSN [28], BMN [27] and G-TAD [51] are proposed to predict the precise boundaries for temporal action localization; and YOWO [20], LFB [49], CRCNN [50], STEP [54] and T-CNN [14] are proposed for efficient detection of spatial-temporal action tubes.

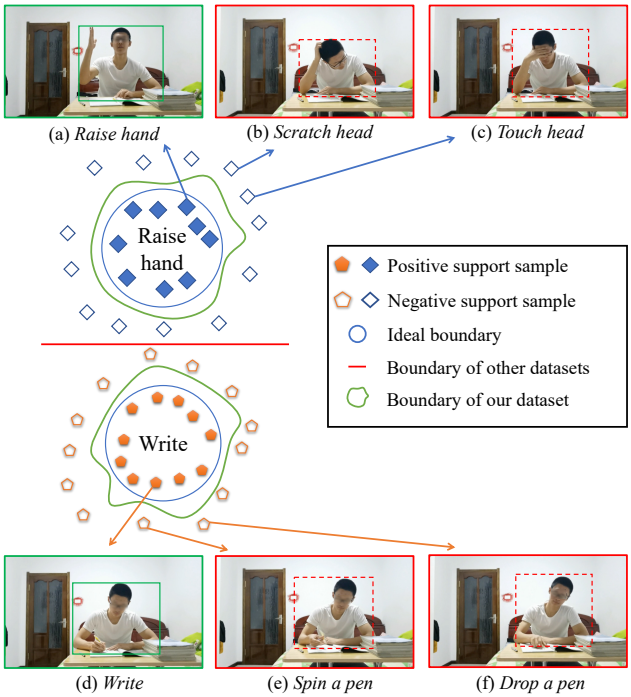


Figure 1. Crowd-worker with different actions under the same background. First row: (a) is a positive sample of the target action "raise hand", while (b) "scratch head" and (c) "touch head" are its negative samples. Second row: (d) is a positive sample of the target action "write", while (e) "spin a pen" and (f) "drop a pen" are its negative samples. We also illustrate the classification boundaries of models trained by different datasets.

Action datasets play an important role for the development of action understanding. Kinetics [19] and something-something [11] boost significantly the study on action recognition. ActivityNet [1], HACS [56] and FineGym [39] enable the study on temporal action localization. AVA [12] provides bounding box annotations for spatial-temporal action detection. The statistics of existing popular datasets are summarized in Table 1. With these datasets, researchers have developed many action understanding methods for ac-

*Corresponding author

tion recognition [2, 4, 7, 8, 24, 26, 31, 33, 35, 38, 41, 44, 44–46, 53, 58, 60], temporal action localization [3, 25, 27–30, 47, 52, 53, 55, 57, 59], and spatial-temporal action detection [14, 20, 49, 50, 54]; however, there remains a gap between these public action understanding datasets and the real-world application scenarios.

Taking daily action understanding as an example, besides the target actions that we are interested in, there are numerous undefined "negative" actions that are very similar to the target actions. For example, in Figure 1, only "raise hand" in (a) and "write" in (d) are target actions, while others are similar-looking negative actions of them. Unfortunately, these negative actions have a huge impact on the action recognition performance during the testing stage. The reason behind this is that using the positive samples alone is not sufficient to learn the robust boundary of actions. Figure 1 shows the ideal boundary for action "raise hand" in the blue circle. If we only collect data that belong to target actions for model training, the learned classification boundary would be the "red line", and the learned model is unable to differentiate similar-looking negative samples, *e.g*. "scratch head" and "raise hand".

To our best knowledge, most of the existing action understanding datasets only have positive samples for the target actions, ignoring the collection of negative samples. Something-something [11] collected samples of "pretend to do something" for several actions but not in a systematic way. In addition, there lacks a dataset especially collected for common daily action understanding, which is very important for understanding our daily life. To address these issues, we establish a large-scale common daily action dataset (CDAD), which has several distinguishing features.

- *Informative support samples.* The action samples are collected in groups, and each positive action sample is coupled with similar-looking negative action samples. These support samples define the fine-grained class boundaries close to ideal class boundaries, as illustrated in Figure 1.
- *Fine-grained spatial-temporal definition and annotation.* We provide a precise spatial-temporal definition for each action and rich annotations. Therefore, multiple tasks can be learned on the dataset.
- *Decoupling of actions, person identities and scenes.* In each group, the target action and its associated negative actions are collected with the same person and background scene. In addition, various target actions are collected at the same scene, which enables the learning process to focus on actions rather than person identities or background.

Overall, the established CDAD dataset consists of 57,824 videos of 23 types of common daily actions. Several negative support samples are collected for each positive support sample to learn more accurate class boundaries. We collect rich annotations for CDAD, including multi-class action labels, start-end time, and spatial bounding box of the actions. Extensive experiments are conducted to investigate the influence of negative samples, and the results demonstrate that introducing negative samples during testing will significantly influence the performance of models on all tasks, while utilizing negative samples during training could benefit both action classification and localization.

2. Related Work

Existing action understanding datasets can be roughly divided into two categories depending on the richness of annotations: action recognition datasets and action detection datasets. A detailed comparison can be found in Table 1.

2.1. Action recognition dataset

Coarse-grained action recognition dataset. For coarse-grained action recognition, the richness, quality and diversity of the datasets have been greatly improved in recent years. HMDB51 [22] and UCF101 [42] are early attempts to enrich the action classes and video volumes to support training deep models. In recent years, large-scale datasets such as THUMOS [15], Sports-1M [18], Kinetics [19] and Moments in Time [32] have been proposed. These datasets greatly promote the development of action recognition methods. However, all these datasets focus on coarse-grained action recognition. The inter-classes boundaries learned on these datasets are only useful for distinguishing the positive samples, but less useful for recognizing the similar-looking negative actions, which are commonly seen in real-world applications. What's more, the background context often tangles with actions in model learning.

Fine-grained action recognition dataset. Fine-grained action datasets [6, 9, 11, 21, 38] focus on subtle details of actions. This is reflected by the fine-grained action labels [11, 39] and sub-actions [6, 9, 21, 38, 39]. Especially, something-something [11] and Charades [40] focus on daily human-object interaction actions. Breakfast [21], MPII-Cooking [37], and Epic-kitchens-100 [5] provide annotations for individual steps of various cooking activities. These fine-grained datasets provide more detailed annotations, which greatly promote the study of action recognition at a finer level. The similar-looking sub-actions in these datasets help define fine-grained class boundaries. However, these datasets have two problems: 1) Not every action has similar-looking sub-actions and many classes are vastly different in semantic and motion; 2) In the real-world case, similar-looking actions are often undefined due to their diversity. Therefore, we tag datasets with sub-actions or sim-

Anno. Type	Dataset	Videos	Class	Total duration(min)	IPC	Negative sample	Datasource	Year
Category	Kinetics400 [19]	306k	400	50k	765	No	Internet	2017
	SthSthv1 [11]	108k	174	7.3k	620	Partial	Crowdsourcing	2017
Category Temporal	MPII-Cooking2 [37]	273	88	1.6k	158	Partial	Self-recorded	2012
	Breakfast [21]	433	50	180	61	Partial	Self-recorded	2014
	THUMOS14 [17]	413	20	1.7k	21	No	Internet	2014
	ActivityNet-1.3 [1]	19,994	200	40k	150	No	Internet	2015
	Charades [40]	9,848	157	4.9k	63	Partial	Crowdsourcing	2016
	HACS-1.1 [56]	49,581	200	124k	694	No	Internet	2019
	FineGym-1.1 [39]	12,818	530	9.8k	97	Partial	Internet	2020
Category Temporal Spatial	UCF-sports [36]	150	10	5.3	15	No	Internet	2008
	ADL [34]	20	18	600	22	Partial	Crowdsourcing	2012
	J-HMDB [22]	928	21	22	44	No	Internet	2013
	UCF101-24 [42]	3207	24	385	134	No	Internet	2017
	AVA [12]	430	80	6.5k	1013	No	Internet	2018
	Epic-kitchens-100 [5]	700	4,053	6k	22	Partial	Self-recorded	2020
	CDAD (ours)	**57,824**	23	**12.9k**	**1577**	**Collected**	Crowdsourcing	2021

Table 1. Comparison of different video datasets, including annotation type, video clip and class number, total video duration, instance per class (IPC), negative sample, datasource and publication year.

ilar action categories as "partial" in "negative sample" column in Table 1.

Our dataset addresses the above issues by relieving the background influences and collecting balanced positive and negative video samples. The details of our dataset construction process can be found in Sec. 3.

2.2. Action detection dataset

Temporal action detection dataset. THUMOS14 [17] is an early attempt of constructing dataset for temporal action detection. ActivityNet [1] and HACS [56] are recently developed large-scale temporal action detection datasets, which contain activity classes that belong to 7 different top-level categories: personal care, eating and Drinking, household, caring and Helping, working, socializing and leisure and sports and exercises. FineGym [39] is a fine-grained dataset that provides three semantic level action classes and sub-action annotations on gymnastic videos. Unlike previous datasets, our CDAD provides 30,000 well-defined temporal annotations for commonly seen daily actions. In addition, the collected negative samples help define clear boundaries for target action instances, which are valuable for both research and industrial applications.

Spatial-temporal action detection dataset. In recent years, some datasets [12,16,42] have been developed to provide both action bounding boxes and temporal annotations for spatial-temporal action detection. J-HMDB [16] gives comprehensive annotations per frame and boosts the early researches on spatial-temporal action detection. UCF101-24 [42] provides spatial-temporal annotation for a subset of UCF101. AVA [12] conducts a semi-auto annotation process. AVA generates a large amount of person bounding boxes by faster-RCNN, and tracklets by linking bounding boxes using Hungarian algorithm [23]. The generated bounding boxes and tracklets are then validated by annotators. ADL [34] contains spatial temporal annotations for egocentric actions. Our dataset differs from previous ones in annotation scale, granularity, and precision. CDAD provides over 200,000 manually annotated, well-defined bounding boxes for actions in five categories.

3. The CDAD dataset

In this section, we introduce in detail the dataset construction process of CDAD, including data collection, data cleaning, preprocessing and annotation. The illustration of annotation tools, video samples, details of experiment settings and analysis will be provided in the supplementary materials. The full dataset and codes will be released for research purpose.

3.1. Data collection, cleaning and preprocessing

We choose 23 indoor daily actions as our target actions. These actions are selected for two reasons. First, they are common and meaningful daily actions for both research and practical applications. Second, there exists a large amount of similar-looking undefined negative actions for these target actions, which will degrade the performance of models in real-world applications.

Many existing video datasets collect videos from online video websites using keyword searching. However, we found that such a data collection method has two drawbacks in our case. 1) It is hard to describe possible similar-looking negative actions using keywords. 2) It is hard to collect actions with the same background scene and by the same person to disentangle target action from background and per-

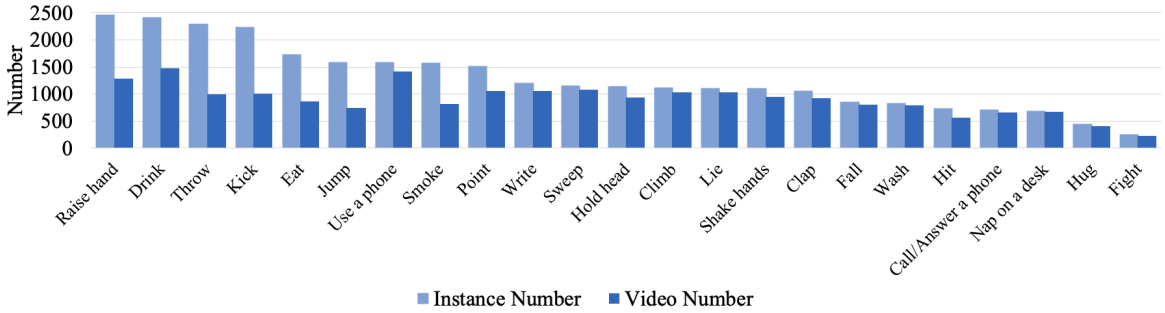


Figure 2. Distribution of the number of action instances and videos per class.

son identities. Therefore, we choose using crowd-sourcing website and ask users to upload videos.

In order to collect desired high-quality data, we use several data collection strategies. *First, disentangle actions from person identities.* Crowd-workers are asked to upload positive action samples as well as similar-looking negative action samples by the same person. As shown in Figure 1, only (a) and (d) are positive samples for target classes, while others are negative samples. *Second, disentangle actions from scenes.* Crowd-workers are asked to record action samples with the same scene for each action group, emphasizing the focused actions rather than the background. *Third, diverse negative actions.* We provide precise definitions for the positive actions, while collecting diverse negative samples for the target actions defined by individual crowd workers. Using all the strategies above, we collected a high-quality and informative dataset for human daily action understanding.

In the data cleaning stage, annotators are asked to strictly follow our data collecting policies and carefully check videos to remove those with the following problems: unrelated to our topic, containing watermarks, redundant, low-resolution, poor image quality, recorded at abnormal or poor conditions, with copy-right issues. After data cleaning, we resize the videos' spatial resolution to the long side of 1280 regarding the original width/height ratio. The FPS of videos are kept the same as the original ones. For videos exceeding 30 FPS, we re-encode the videos to 30 FPS.

3.2. Action instance definition

Generally speaking, an action instance can be defined from two different perspectives: spatial and temporal.

Action instance definition. Different from previous datasets, which provide only vague full-body annotation, we collect fine-grained spatial annotations for the action. We divide daily actions into four categories: half-body action, full-body action, human-object interaction action and multi-person action. For half-body action, such as "raise hand", annotators are asked to annotate the upper body bounding box. For full-body action such as "lie", annotators are asked to annotate the full-body bounding box. For human-object interaction action such as "call/answer a phone", upper body and object should be included in the bounding box. For actions such as "sweep", the bounding box should include full body and object. For multi-person action, such as "shake hands", the bounding box needs to contain both persons. We provide a detailed illustration of action type, bounding box type, and bounding box number for all actions in the supplementary materials.

We use clear signals to determine the start and end frames of the action. If the actor do the same action several times during the video, we annotate them as different action instances with the same label. If the interval of two consecutive action instances is less than one second, we label them as one action instance.

Action instance annotation. Every positive video sample contains several action instances. The annotation for every action instance $\mathbf{A}$ contains action label cls, action start time t_s, end time t_e and a series of bounding boxes of it:

$$\mathbf{A} = (cls, t_s, t_e, \{b_1, b_2, ..., b_N\}), \mathbf{b}_i = [f_i, x_{i1}, y_{i1}, x_{i2}, y_{i2}].$$

where f_i are frame-id and (x_{i1}, y_{i1}), (x_{i2}, y_{i2}) are corner coordinates. We provide one spatial annotation per second, and for actions less than one second, the start, end and the middle frame of the action instance are annotated for the completeness of fast-tempo actions (*e.g.* jump).

3.3. Data annotation

As fine-grained daily action understanding aims to differentiate between subtle daily actions, the quality of annotations is extremely important. In order to provide high-quality annotations, we hire a professional annotation team and conduct rigorous annotation. The whole annotation process can be divided into five steps:

1) *Annotation standard training.* We provide annotators with detailed annotation standard documents and train them with the background knowledge and the spatial-temporal definition of every action.

2) *On-site demo.* After annotators understand the annotation standard, we give them on-site annotation demos to clarify our annotation requirements.

3) *Trial annotation.* In this stage, annotators are asked to annotate a sample set of videos. The annotation results

Figure 3. Frame and annotation samples of CDAD, including multi-person action (e.g. shake hands, hug), single person action (e.g. raise hand), human-object interaction action (e.g. smoke, sweep). Spatial-temporal annotations are provided for all actions. "Green" indicates temporal annotation of action instances, "blue" represents background (no action), and red rectangles are spatial annotations.

are carefully checked to make sure their understanding of annotation is correct.

4) *Formal annotation.* The formal annotation follows the trial annotation. During the annotation, we create an online discussion group for clarifying confusion cases in time.

5) *Cross-validation.* We conduct cross-validation when the formal annotation ends. Every action instance is checked by at least two annotators to ensure its correctness.

3.4. Dataset statistics and property

Finally, our CDAD dataset contains 57,824 videos of 23 action types, including 24,174 positive video samples and 33,650 negative video samples. The duration of video ranges from 8 to 21 seconds. We provide 36,271 action instances and 212,151 action bounding boxes in total. The dataset is split into train, validation and test in the ratio of 8:1:1. We ensured that all videos provided by the same crowd worker occur only in one split (train, validation, test). In its current version, the dataset was generated by 3,866 crowd workers with an average of 621 workers per class. Figure 2 shows the distribution of the number of videos and action instances per class. The average video per class is 903. The action "fight" has the minimum number of 228 videos, and the action "drink" has the maximum number of 1,479 videos. The average number of action instance is 1,577 per class. The action "fight" also has the minimum number of 255 action instances, and the action "raise hand" has the maximum number of 2,471 action instances. Frame samples from the collected videos are shown in Figure 3.

The videos of CDAD are recorded and uploaded by crowd-workers. All the videos are of good resolution (720P). The annotators are well trained and the annotation results are double-checked to ensure consistency and quality. For every target action, crowd-workers are asked to upload one positive sample and two negative samples. These negative samples are similar to the positive sample, which provide rich information to decide the boundary of the target actions. For every positive video, we carefully annotate the start-end time of actions as well as the spatial bounding box for the action. The rich annotations enable many video understanding tasks on our dataset.

4. Experiments

With the constructed CDAD dataset, we are able to conduct experiments on different tasks of human daily action understanding, including multi-label classification, temporal action localization and spatial-temporal action detection.

4.1. Multi-label classification

Experiment setup. We employ the popular and representative temporal segment network (TSN) [45] and temporal shift module (TSM) [26] for temporal modeling, and use ResNet [13] as backbone to extract features of individual

Method	Backbone	Frame	N-test	mAP
TSN	resnet18	8	×	76.8
			✓	57.8
		16	×	78.9
			✓	59.8
TSM	resnet18	8	×	82.1
			✓	62.7
		16	×	85.4
			✓	66.8
	resnet50	8	×	88.1
			✓	68.1
		16	×	90.4
			✓	71.1

Table 2. Multi-label classification experiments on CDAD by models trained without negative samples. "N-test" represents whether the negative samples are used in testing.

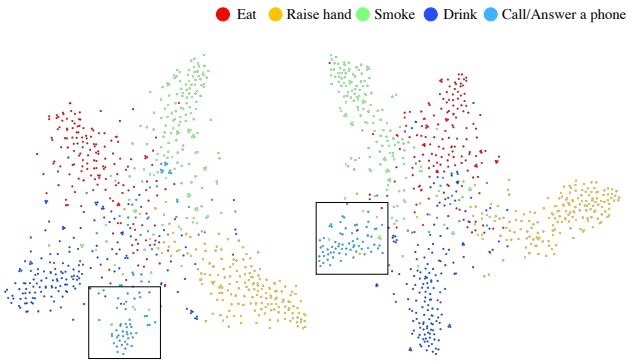
Method	Backbone	Frame	N-train	mAP	Top1-err	HL
TSN	resnet18	8	×	57.8	37.4	0.023
			✓	58.4	20.2	0.017
		16	×	59.8	36.5	0.023
			✓	60.4	19.3	0.017
TSM	resnet18	8	×	62.7	42.9	0.026
			✓	68.9	16.8	0.015
		16	×	66.8	38.2	0.022
			✓	72.8	17.2	0.014
	resnet50	8	×	68.1	42.5	0.024
			✓	72.3	16.4	0.013
		16	×	71.1	43.6	0.025
			✓	76.5	16.8	0.013

Table 3. Multi-label classification experiments on CDAD by models tested with negative samples. "N-train" represents whether negative samples are used in training.

frames. We average the obtained features for each frame in the video to form the final encoding. We use uniform sampling in TSN [45] and take 8 or 16 frames as inputs. The networks were pre-trained on ImageNet. For multi-label classification, we use sigmoid activation and binary cross-entropy loss as the final loss function. The negative samples are used as samples with label to be full zeros. Based on [43], we use example-based evaluation metric hamming loss and ranking-based metrics mAP and Top1-error as our evaluation metrics.

Results and analysis. We conduct experiments to validate the roles of collected negative samples in our dataset. There are two experimental settings. First, we train models without the collected negative samples, and test the trained models with and without the negative samples. Second, we train models with the collected negative samples, and test them with the negative samples.

The results of the first experiment are shown in Table 2. We see that without negative samples in training, the testing stage is largely influenced by negative samples. By introducing negative samples in the query samples, the mAP drops around 20% for all models. This indicates that if a model is trained only by the positive samples, which is a common practice in previous action datasets, it would perform poorly at the testing in real scenarios, where there are a lot of undefined negative samples. As many negative actions are similar to target actions, many false positives (FP) are expected to be reported by the trained model.



a. Trained without negative samples b. Trained with negative samples

Figure 4. Feature visualization of models trained w/w.o. negative samples.

We further conduct experiments by introducing negative samples to train the models. From Table 3, we can see all the models achieve improvements on mAP and Top1-error. Hamming loss (the lower the better) is significantly decreased. This performance bump demonstrates the great benefits of utilizing negative samples during training. The robustness of the trained model is largely improved and FPs are significantly reduced.

It is found that models with better temporal modeling modules would achieve greater performance bump when using negative samples for training. TSM-R18-8frame achieves 6.2% improvements on mAP, while TSN-R18-8frame only improves by 0.6%. Using more frames also helps the learning of subtle differences between positive and negative samples. For example, TSM-R18-16frame achieves 3.9% performance improvements than TSM-R18-8frame. The Top-error and HL begin to saturate when use TSM-R50-16frame compared to 8frame models. In general, models with better temporal modeling capabilities can better differentiate positive and negative samples.

In Fig. 4, we show feature visualization results by t-SNE of five similar-looking actions on validation set. TSM-R18-8frame is used for both training settings. We can see that features in b are more tightly aggregated than features in a. Especially, in the rectangle areas, more samples are aggregated in b than a, for action "call/answer a phone". This indicates negative samples "push" the positive actions further from other confusing actions and closer to actions from the same categories.

The APs of different classes by using negative samples during the training of model TSM-R18-8frame are shown in Figure 5. It can be seen that "throw", "eat", "use a phone"

Method	N-train	N-test	mAP@tIOU			avg-mAP
			0.5	0.75	0.95	
G-TAD	×	×	44.96	37.44	16.66	35.72
	×	✓	29.58	25.52	12.60	24.33
	✓	✓	**32.23**	**27.39**	**13.50**	**26.18**
BMN	×	×	44.50	37.25	12.77	35.03
	×	✓	17.43	12.65	5.65	12.58
	✓	✓	**32.29**	**27.06**	**10.65**	**25.85**

Table 4. Temporal action localization experiments on CDAD.

are the most difficult ones, due to the fast-tempo of actions or small-size of objects. "nap on a desk" and "sweep" are the easiest among these actions. Negative samples improve the AP of all categories. The AP improvements of "point", "kick", "smoke", "throw" are very significant.

Feature analysis. To further illustrate the properties of CDAD compared to former datasets such as Kinetics, we conduct feature analysis on Charades with models pre-trained on Kinetics and CDAD. We use several actions as target actions (e.g. "eat","drink"), and cut video clips and re-organize labels in Charades for our study as these actions also appear in Charades, Kinetics, and CDAD. For Charades, we select actions such as "Drinking from a cup/glass/bottle" as positive category for "drink", and "Pouring something into a cup/glass/bottle","Putting a cup/glass/bottle somewhere", *etc*. as negative categories. We apply the same strategy for other actions and create a subset of Charades (more details can be found in supplementary materials). TSM-R18-8frame pre-trained on Kinetics and CDAD is used to generate features separately. We then use SVM for classification and plot the confusion matrix of four classes (0:eat, 1:neg eat, 2:drink, 3:neg drink) for visualization and analysis. The model pre-trained on CDAD achieves 43.9 % accuracy, while models pre-trained on Kinetics achieve 41.7%. Fig. 6 shows the confusion matrix. On target labels 0 (eat) and 2 (drink), the model trained on CDAD achieves 48.5% accuracy, with a 6.7% improvement over the model trained on Kinetics. As pointed out in [48], actions in Kinetics are highly correlated with scenes. The experiments above indicate CDAD's model can better handle similar actions with the same scene. In Fig. 7, we show some action samples from charades: a woman (a) drinks water, (b) pours water, (c) eats a sandwich, and (d) rubs mouth with a towel in the kitchen. The model pre-trained on CDAD performs better under all these scenarios. It is worth noting that Kinetics has much more videos than CDAD, which indicates CDAD's advantages in distinguishing similar-looking confusion actions in daily life.

4.2. Temporal action localization

We then perform experiments to test the influence of collected negative samples on action localization capability. We follow the pipeline in previous works [27, 51] and conduct feature extraction, temporal proposal generation and classification, respectively.

Feature extraction. We first train a video classification model for frame-level feature extraction. We choose the TSM model pre-trained on ImageNet with resnet50 backbone as our feature extractor. During training, we sample 8 frames in a randomly selected one-second time window in a video clip, and use multi-label classification as the target task to train the feature extractor. After training the model, we extract frame-level features from a video clip by 0.2s time step, and keep the output feature dimension to 256.

Temporal proposal generation. We apply BMN [27] and G-TAD [51] for temporal action localization. BMN integrates the 2D confidence map and start/end score prediction. G-TAD utilizes the temporal and semantic graph neural network to improve the capacity of proposal generation. We train these models under two settings: with or without negative samples. As we focus on evaluating the localization performance of the model, during testing we generate video-level class labels using TSM-R50 model and combine labels with the proposals generated above.

Evaluation metrics. We take mean Average Precision (mAP) at 0.5, 0.75, 0.95 IoU thresholds as the main evaluation metric. We also report average mAP over 10 different IoU thresholds [0.5 : 0.05 : 0.95].

Results and analysis. The temporal localization results are reported in Table 4. We can see a clear performance drop on average-mAP (11.4% on G-TAD and 22.5% on BMN) when models are tested with negative samples. This indicates that the proposals generated by negative samples will also have a high score, which degrades the mAP performance of the model. It is also worth noticing that negative samples have larger influences on BMN than G-TAD. This may indicate that G-TAD can better utilize context information with the graph structure, and capture the subtle differences of actions. By adding our collected negative samples in training, the mAP of the model improves by 1.9% for G-TAD and 13.3% for BMN, which demonstrates the benefits of introducing negative samples in training for accurate temporal location regression.

4.3. Spatial-temporal action detection

Experiment setup. We employ YOWO [20] as the baseline model. YOWO is a single-stage end-to-end spatial-temporal action detection framework. It contains two branches, 2D CNN for spatial feature extraction and 3D CNN for temporal modeling, as well as a Channel Fusion and Attention Mechanism (CFAM) to fuse the two branches. The model is first trained with frame-level detection. We take 16 frames and resize them to 224×224 for input. During the testing stage, we follow [10] for linking score calculation, and apply the Viterbi algorithm for finding the optimal path to generate tubes.

For negative video samples in our dataset, in order to

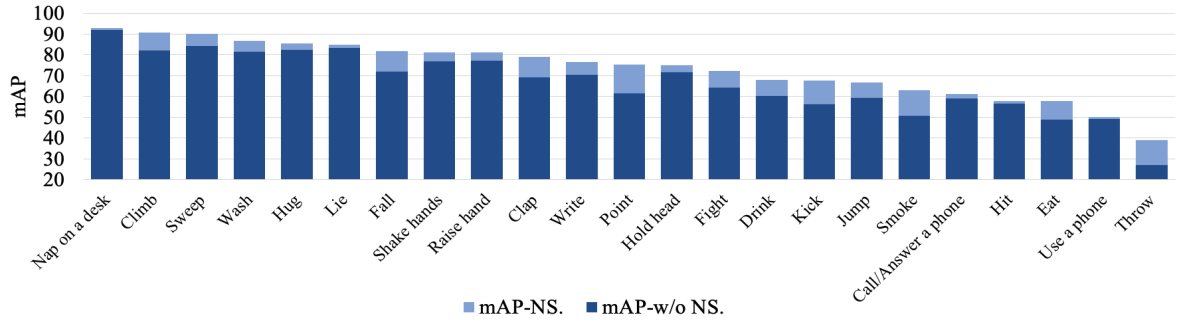
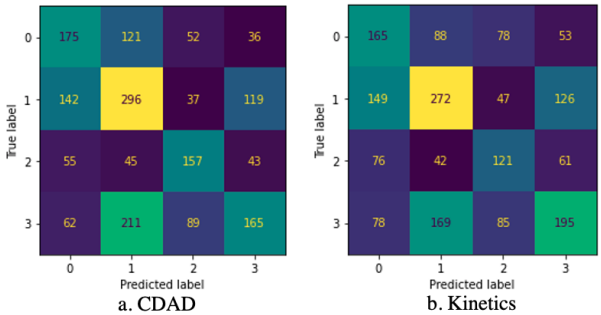
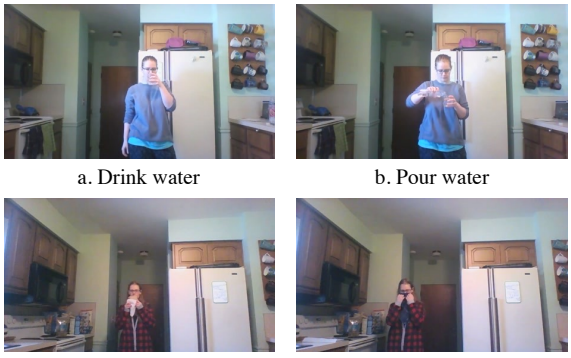


Figure 5. Multi-label classification results of different classes by training with or without collected negative samples (NS).



a. CDAD b. Kinetics

Figure 6. Confusion matrix on Charades by models pretrained on CDAD and Kinetics.



a. Drink water b. Pour water

c. Eat a sandwich d. Rub mouth with towel

Figure 7. Samples of actions from Charades for feature analysis.

keep consistency between positive and negative samples, we sample 1 frame/second for both training and testing.

Evaluation metrics. We follow [20] and employ two popular metrics for evaluation. 1) *Frame-mAP@0.5-IOU*. This metric measures the area under the detection precision-recall curve for each frame at 0.5 IOU threshold. 2) *Video-mAP@(0.1,0.2,0.5)-IOU*. It focuses on action tubes. The detected tube is regarded as a correct action instance if the mean per frame IOU with the ground truth across all frames of the action segment is greater than a threshold, and the action label is correctly predicted.

Results and analysis. As we can see from Table 5, the negative samples at the testing stage have a huge influence on spatial-temporal action detection. The frame-mAP@0.5-IOU and Video-mAP@0.1-IOU drop by 53.6% and 36.9%, respectively. Using negative samples during training improves both regression and classification accuracy. The frame-mAP improves by 5.12% and video-mAP@0.1-IOU improves by 15.7%, which indicates the effectiveness of negative samples. However, the model performance still has much room to be improved, raising new challenges for future research on spatial-temporal action detection.

Methods	N-train	N-test	V-mAP			F-mAP
			0.1	0.2	0.5	
YOWO	×	×	48.41	48.07	43.80	70.13
	×	✓	11.54	11.44	9.94	16.55
	✓	✓	**27.27**	**27.16**	**25.39**	**21.67**

Table 5. Spatial-temporal detection experiments on CDAD.

5. Conclusion and discussions

In this work, we established a new action dataset, namely common daily action dataset (CDAD), which contains 57,824 videos for 23 types of common daily actions. For every positive sample of the target action, we collected several associated negative support samples, which were proven very crucial for learning the desired class boundaries. Rich annotations were collected for CDAD, including multi-class action labels, start-end time of the actions, and spatial bounding boxes of the actions. We conducted extensive experiments to provide baseline results and validated the importance of the collected negative samples in three tasks: multi-label action recognition, temporal action localization and spatial-temporal action detection. The proposed CDAD dataset provided an important benchmark for human daily action understanding. In particular, it allowed researchers to investigate in-depth the roles and effects of negative examples, which are highly similar to positive ones, on robust action understanding model learning.

Though the proposed CDAD has large volumes of video number, total video time, action instances, and collected negative samples, it has a relatively limited number of categories. In addition, the background scenes in CDAD are not rich enough, either. We will consider enriching the categories and background scenes of CDAD in the future, facilitating the research of practical daily action understanding.

References

[1] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015. 1, 3

[2] J. Carreira and A. Zisserman. In *CVPR*, pages 4724–4733, 2017. 1, 2

[3] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A Ross, Jia Deng, and Rahul Sukthankar. Rethinking the faster r-cnn architecture for temporal action localization. In *CVPR*, pages 1130–1139, 2018. 2

[4] R Christoph and Feichtenhofer Axel Pinz. Spatiotemporal residual networks for video action recognition. *NIPS*, pages 3468–3476, 2016. 2

[5] 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: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 2021. 2, 3

[6] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, pages 720–736, 2018. 2

[7] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, pages 6202–6211, 2019. 1, 2

[8] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, pages 1933–1941, 2016. 2

[9] Yixin Gao, S Swaroop Vedula, Carol E Reiley, Narges Ahmidi, Balakrishnan Varadarajan, Henry C Lin, Lingling Tao, Luca Zappella, Benjamın Béjar, David D Yuh, et al. Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In *MICCAI workshop: M2cai*, volume 3, page 3, 2014. 2

[10] G. Gkioxari and J. Malik. Finding action tubes. In *CVPR*, pages 759–768, 2015. 7

[11] 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 *ICCV*, pages 5842–5850, 2017. 1, 2, 3

[12] 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 *CVPR*, pages 6047–6056, 2018. 1, 3

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5

[14] Rui Hou, Chen Chen, and Mubarak Shah. Tube convolutional neural network (t-cnn) for action detection in videos. *ICCV*, 2017. 1, 2

[15] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 155:1–23, 2017. 2

[16] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *ICCV*, pages 3192–3199, Dec. 2013. 3

[17] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. `http://crcv.ucf.edu/THUMOS14/`, 2014. 3

[18] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014. 2

[19] 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, 2, 3

[20] Okan Köpüklü, Xiangyu Wei, and Gerhard Rigoll. You only watch once: A unified cnn architecture for real-time spatiotemporal action localization. 2019. 1, 2, 7, 8

[21] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *CVPR*, pages 780–787, 2014. 2, 3

[22] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *ICCV*, pages 2556–2563. IEEE, 2011. 2, 3

[23] H. W. Kuhn and Bryn Yaw. The hungarian method for the assignment problem. *Naval Res. Logist. Quart*, pages 83–97, 1955. 3

[24] Yan Li, Bin Ji, Xintian Shi, Jianguo Zhang, Bin Kang, and Limin Wang. Tea: Temporal excitation and aggregation for action recognition. In *CVPR*, pages 909–918, 2020. 2

[25] Chuming Lin, Jian Li, Yabiao Wang, Ying Tai, Donghao Luo, Zhipeng Cui, Chengjie Wang, Jilin Li, Feiyue Huang, and Rongrong Ji. Fast learning of temporal action proposal via dense boundary generator. In *AAAI*, volume 34, pages 11499–11506, 2020. 2

[26] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *ICCV*, pages 7083–7093, 2019. 1, 2, 5

[27] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *ICCV*, pages 3889–3898, 2019. 1, 2, 7

[28] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: Boundary sensitive network for temporal action proposal generation. In *ECCV*, pages 3–19, 2018. 1, 2

[29] Qinying Liu and Zilei Wang. Progressive boundary refinement network for temporal action detection. In *AAAI*, volume 34, pages 11612–11619, 2020. 2

[30] Fuchen Long, Ting Yao, Zhaofan Qiu, Xinmei Tian, Jiebo Luo, and Tao Mei. Gaussian temporal awareness networks for action localization. In *CVPR*, pages 344–353, 2019. 2

[31] Brais Martinez, Davide Modolo, Yuanjun Xiong, and Joseph Tighe. Action recognition with spatial-temporal discriminative filter banks. In *ICCV*, pages 5482–5491, 2019. 2

[32] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *TPAMI*, 42(2):502–508, 2019. 2

[33] AJ Piergiovanni and Michael S Ryoo. Representation flow for action recognition. In *CVPR*, pages 9945–9953, 2019. 2

[34] Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2847–2854, 2012. 3

[35] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *ICCV*, pages 5533–5541, 2017. 1, 2

[36] M. D. Rodriguez, J. Ahmed, and M. Shah. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, pages 1–8, 2008. 3

[37] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele. A database for fine grained activity detection of cooking activities. In *CVPR*, pages 1194–1201, 2012. 2, 3

[38] Marcus Rohrbach, Anna Rohrbach, Michaela Regneri, Sikandar Amin, Mykhaylo Andriluka, Manfred Pinkal, and Bernt Schiele. Recognizing fine-grained and composite activities using hand-centric features and script data. *IJCV*, 119(3):346–373, 2016. 2

[39] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *CVPR*, pages 2616–2625, 2020. 1, 2, 3

[40] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, pages 510–526. Springer, 2016. 2, 3

[41] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, pages 568–576, 2014. 2

[42] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2, 3

[43] Mohammad S Sorower. A literature survey on algorithms for multi-label learning. *Oregon State University, Corvallis*, 18:1–25, 2010. 6

[44] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497, 2015. 2

[45] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Val Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016. 1, 2, 5, 6

[46] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, pages 20–36. Springer, 2016. 2

[47] Philippe Weinzaepfel, Xavier Martin, and Cordelia Schmid. Human action localization with sparse spatial supervision. *arXiv preprint arXiv:1605.05197*, 2016. 2

[48] Philippe Weinzaepfel and Grégory Rogez. Mimetics: Towards understanding human actions out of context. *International Journal of Computer Vision*, 129(5):1675–1690, 2021. 7

[49] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krähenbühl, and Ross Girshick. Long-Term Feature Banks for Detailed Video Understanding. In *CVPR*, 2019. 1, 2

[50] Jianchao Wu, Zhanghui Kuang, Limin Wang, Wayne Zhang, and Gangshan Wu. Context-aware rcnn: A baseline for action detection in videos. In *ECCV*, 2020. 1, 2

[51] Mengmeng Xu, Chen Zhao, David S. Rojas, Ali Thabet, and Bernard Ghanem. G-tad: Sub-graph localization for temporal action detection. In *CVPR*, June 2020. 1, 7

[52] Mengmeng Xu, Chen Zhao, David S Rojas, Ali Thabet, and Bernard Ghanem. G-tad: Sub-graph localization for temporal action detection. In *CVPR*, pages 10156–10165, 2020. 2

[53] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. Temporal pyramid network for action recognition. In *CVPR*, pages 591–600, 2020. 2

[54] Xitong Yang, Xiaodong Yang, Ming-Yu Liu, Fanyi Xiao, Larry S Davis, and Jan Kautz. Step: Spatio-temporal progressive learning for video action detection. In *CVPR*, 2019. 1, 2

[55] Da-Hye Yoon, Nam-Gyu Cho, and Seong-Whan Lee. A novel online action detection framework from untrimmed video streams. *Pattern Recognition*, 106:107396, 2020. 2

[56] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. Hacs: Human action clips and segments dataset for recognition and temporal localization. 2017. 1, 3

[57] Peisen Zhao, Lingxi Xie, Chen Ju, Ya Zhang, Yanfeng Wang, and Qi Tian. Bottom-up temporal action localization with mutual regularization. In *ECCV*, pages 539–555. Springer, 2020. 2

[58] Yue Zhao, Yuanjun Xiong, and Dahua Lin. Recognize actions by disentangling components of dynamics. In *CVPR*, pages 6566–6575, 2018. 2

[59] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *ICCV*, pages 2914–2923, 2017. 2

[60] Yin-Dong Zheng, Zhaoyang Liu, Tong Lu, and Limin Wang. Dynamic sampling networks for efficient action recognition in videos. *TIP*, 29:7970–7983, 2020. 2