JRDB-Act: A Large-scale Dataset for Spatio-temporal Action, Social Group and Activity Detection

Mahsa Ehsanpour[1], Fatemeh Saleh[2*], Silvio Savarese[3], Ian Reid[1], Hamid Rezatofighi[4]
[1]The University of Adelaide, [2]Samsung AI Center, [3]Stanford University, [4]Monash University
https://jrdb.erc.monash.edu/

Abstract

The availability of large-scale video action understanding datasets has facilitated advances in the interpretation of visual scenes containing people. However, learning to recognise human actions and their social interactions in an unconstrained real-world environment comprising numerous people, with potentially highly unbalanced and long-tailed distributed action labels from a stream of sensory data captured from a mobile robot platform remains a significant challenge, not least owing to the lack of a reflective large-scale dataset. In this paper, we introduce JRDB-Act, as an extension of the existing JRDB, which is captured by a social mobile manipulator and reflects a real distribution of human daily-life actions in a university campus environment. JRDB-Act has been densely annotated with atomic actions, comprises over 2.8M action labels, constituting a large-scale spatio-temporal action detection dataset. Each human bounding box is labeled with one pose-based action label and multiple (optional) interaction-based action labels. Moreover JRDB-Act provides social group annotation, conducive to the task of grouping individuals based on their interactions in the scene to infer their social activities (common activities in each social group). Each annotated label in JRDB-Act is tagged with the annotators' confidence level which contributes to the development of reliable evaluation strategies. In order to demonstrate how one can effectively utilise such annotations, we develop an end-to-end trainable pipeline to learn and infer these tasks, i.e. individual action and social group detection. The data and the evaluation code will be publicly available at https://jrdb.erc.monash.edu/.

1. Introduction

Understanding and predicting human actions and intentions are essential tasks in tackling many real-world problems such as autonomous driving, robot navigation safety, human-robot interaction, and detection of perilous behaviors in surveillance systems. Developing an AI model performing these tasks is challenging due to the high variations of human actions in an unconstrained real-world environment. Moreover, dealing with daily actions which resembles a highly unbalanced, long-tailed distribution poses new challenges for many existing approaches.

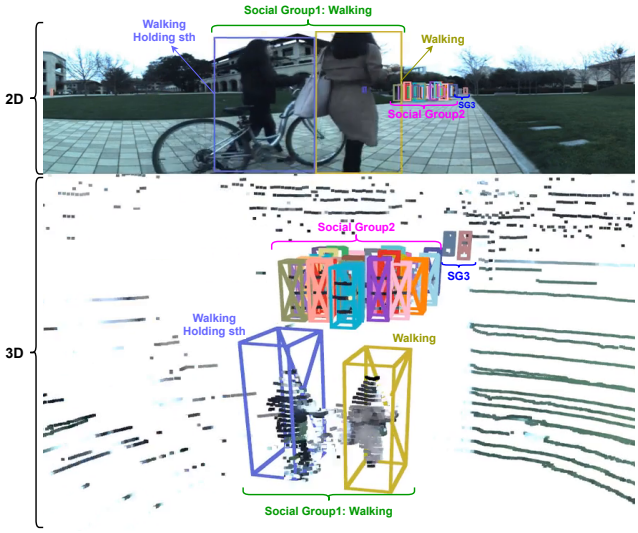

Figure 1. An illustration of a single frame of the JRDB-Act dataset. As shown, the data captured with a 2D and 3D multi-modal sensory platform is accompanied with a new set of annotations including individual actions and social group formation leading to infer social activities (common activities in each social group) to further complement the 2D and 3D detection and tracking annotation in the JRDB.

Recently, great progress has been made to create large-scale video datasets for human activity understanding [3, 9, 20, 26, 45]. While these popular datasets have contributed significantly to the recent advances in human activity understanding from visual data, their primary application is not targeting robotics domain and therefore rarely reflect the challenges in problems such as human-robot interaction and robot navigation in human crowded environments, *e.g.*

*Work done while at the Australian National University (ANU).

shopping malls, university campus, *etc*. Such environments include not only many individuals, but also often groups of people connected to each other through some form of interaction, *e.g*. engaging in common activities or goals, which form the concept of social groups and activities. Moreover, in many robotics problem, *e.g*. for safe navigation and collision risk prediction in human environments, it is essential to anticipate every individual's action and intention way ahead of time, considering their social interactions. To this end, the availability of a spatio-temporally dense annotated human action data is indispensable for the development and evaluation of a robotic perception system.

With this motivation, we introduce JRDB-Act, a large-scale dataset captured from a mobile robot platform, containing dense spatio-temporal individual action and social group annotation. JRDB-Act is an extension of the recently introduced JRDB [35, 43]. We now elaborate the unique characteristics of JRDB-Act and our proposed method.

New Annotations. We provide a set of atomic action labels for each person at each frame from the three categories of human pose, human-human, and human-object interactions, as shown in Fig. 1. Our action vocabulary contains common daily human actions including 11 human pose, 3 human-human, and 12 human-object interaction classes. Since these action labels are densely annotated over space and time, JRDB-Act contains over 2.8M action labels, making it one of the large-scale spatio-temporal action datasets publicly available. Furthermore, the dataset provides new unique annotations, *i.e*. social group labels, by assigning a group ID to each person in each frame such that individuals with the same ID represent a social group. We further provide social activity annotation for each group by inferring it from the annotated individual actions and social groups. Another novel aspect of JRDB-Act is the difficulty level annotation, *e.g*. easy, moderate, and difficult, for each annotated label which reflects the confidence level of annotators for the corresponding label. The provided difficulty level can be conducive to more reliable evaluation paradigms.

Unique Challenges. The sequences in JRDB-Act are captured from human daily-life in different indoor and outdoor places of a university campus as an unconstrained environment [35] by a mobile robotic platform. Thus, they reflect the highly unbalanced distribution of human actions in real-world scenarios. Moreover, the sequences naturally include diverse levels of human population density. The average number of people per frame in JRDB-Act is 30, which is significantly higher than most popular action datasets. Further, the robot motion and the perspective view of the captured sequences makes this dataset challenging. Considering the aforementioned compelling attributes, dense annotations, and natural complexities, JRDB-Act introduces means to study new problems and challenges in human understanding for computer vision and robotics community.

Our Proposed Method. In order to showcase the potential research directions and challenges required to be tackled in JRDB-Act, we develop an end-to-end trainable pipeline for both individual action and social group detection tasks. Our method uses the panoramic video clips as input and adopts a similar backbone as [13] to extract spatio-temporal individual features. However, we fuse additional pair-wise geometrical features and incorporate a novel eigenvalue-based loss function to improve the social group detection performance compared to [13]. We also suggest a simple, yet effective strategy to handle the unbalanced nature of action labels by partitioning and balancing action loss functions based on the occurring frequency of action classes in the dataset.

2. Related Work

Datasets. Over the last decade, multiple video action datasets such as HMDB-51 [28] and UCF101 [47] have been introduced which consist of short clips for the video classification task [18, 34, 42]. Since these datasets are not large and diverse enough to train deep models, large-scale video datasets such as Sports1M [25], YouTube-8M [1], Something-something [19] and Kinetics [26] have been introduced for the task of video action classification. Some other video datasets such as ActivityNet [3], THUMOS [23], MultiTHUMOS [53], Charades [45] and HACS [57] contain untrimmed videos for the task of temporal action localization. Few datasets, such as CMU [27], MSR Actions [54], UCF Sports [40] and JHMDB [24] provide spatial as well as temporal localization. The small number of action categories and the limited number of short video clips motivated the community to introduce AVA [20] and AVA-Kinetics [29], two large-scale spatio-temporal action detection datasets. In AVA, spatio-temporal action labels are provided for one frame per second, in which every person is annotated with a bounding box and at least one action. The AVA-Kinetics dataset extends Kinetics with AVA-style annotation. There are also a number of video datasets such as SOA [37] and HVU [11] that provide multi-label annotation by providing scene, object, event, attribute and concept labels in a video, still limited to the video classification task. As another group of datasets, instructional video analysis datasets [2, 9, 41, 49] have been released which are focused on a specific domain such as cooking or furniture assembly. Volleyball [22] and Collective Activity Dataset (CAD) [7] have been introduced with a focus on group activity recognition. In these datasets, actors are annotated with an action label and the whole scene is annotated with one group activity label. However, a real scene generally comprises several groups of people with potentially different social activities. Recently, CAD has been extended in terms of annotations to Social-CAD [13], in which different social groups and their corresponding social activities have been annotated. While Social-CAD is

the first attempt to tackle spatio-temporal action and social group detection tasks, it only contains 44 sequences with limited labels. Although all these datasets have vastly contributed to the recent advances in human action understanding in videos, they are not capable of reflecting challenges in robotics applications in human crowded environments. To target such specific application domains *e.g*. social robot navigation and human-robot interaction, we propose JRDB-Act, a large-scale spatio-temporal human action, social group and per-group social activity detection dataset captured from a mobile robot which has been annotated densely in space and time.

Action Analysis Frameworks. Over the past few years, there have been extensive studies on video classification [4, 12,46,55] and temporal action detection [44,45,52,58]. Recently, by introducing spatio-temporally annotated datasets such as AVA [20], the spatio-temporal action detection task [15–17, 30, 48, 50] has received considerable attention. In parallel, there also exist works focusing on group activity recognition on datasets, *e.g*. Volleyball [22] and CAD [7], where the aim is to predict a single group activity label for the entire scene [5, 6, 8, 31, 32]. Although these approaches try to recognise the interactions between people for group activity recognition, they are not capable of inferring social groups. Recently, CAD [7] has been augmented with social group and social activity label per group in [13] and a corresponding framework is proposed to detect individuals' action, social groups, and social activity for each group in the scene. However, as substantiated by our experiments, this framework does not generalize well to the natural complexities of JRDB-Act. To improve upon this framework's performance, we *(i)* exploit the bounding box locations to derive pair-wise geometrical features and further incorporate an eigenvalue-based loss to enhance the social group detection task and, *(ii)* suggest a simple strategy, *i.e*. a loss partitioning approach, to handle the unbalanced nature of action labels in the dataset.

3. The JRDB-Act Dataset

The multi-modal JRDB dataset [35] is composed of 64 minutes of sensory data captured by the mobile JackRabbot robot, containing 54 sequences of indoor and outdoor scenes in a university campus environment, covering different human poses and social interactions. JRDB provides 1) over 2.4 million 2D bounding boxes for all the people visible in the five stereo RGB cameras, capturing a panoramic cylindrical 360°image view, 2) over 1.8 million 3D oriented bounding boxes in the point-clouds captured from the two 16-array LiDAR sensors, 3) association of all the 3D bounding boxes with the corresponding 2D boxes, and 4) track ID of all the 2D and 3D boxes over time. While the provided annotations are useful for human localization and tracking, JRDB lacks sufficient information for social human activity analysis. Therefore we propose JRDB-Act by providing additional individuals' human action and social group annotation on top of the existing JRDB. All these annotations make JRDB-Act the only available dataset for multi-task learning of human detection, tracking, individual action, social group and per-group social activity detection. JRDB-Act is manually annotated by a group of annotators, instructed for each task to ensure consistency over the dataset. Then, it has been inspected by another group, instructed for the quality assessment of the provided annotations. The rest of this section provides details of JRDB-Act regarding annotations, benchmarking, and statistics.

A. Action Vocabulary. Since JRDB is collected in a university campus environment, our action vocabulary includes common daily human actions. Through a comprehensive inspection of the dataset, we concluded 11 pose-based, 3 human-human interaction and 12 human-object interaction action labels. Fig. 2 demonstrates the list of existing action labels per category in JRDB-Act.

B. Action Annotation. Action annotation is densely provided per-frame (7 fps) and per-box for both the LiDAR and video sequences. However, the panoramic videos are used to annotate the action labels. During the annotation process, we utilised JRDB annotated 2D-bounding boxes and track-IDs; for each bounding box, one (mandatory) pose-based and an arbitrary number of (optional) interaction-based action labels were selected from the available list of action vocabulary. If none of the classes in the list were descriptive for a bounding box, annotators were able to tag the box as *miscellaneous-[description]* for each label category, and the descriptions were later used to expand the action vocabulary with the newly discovered labels. Annotators also tagged each action label with its corresponding difficulty level indicating the annotator's confidence level for the corresponding label. There are scenarios where 1) action label is obvious, 2) there is uncertainty in the action label but we can take a probable guess, and 3) the person is far away from the camera or occluded, however the action could be inferred from some evidences such as its past history and its current movement. We respectively tag these scenarios as *easy, moderate,* and *difficult*. In some cases, it is not possible to infer the action as the bounding box is fully-occluded in the duration of the video or the person is very far from the robot. Here, both the pose-based action label and its corresponding difficulty level are tagged as *impossible*. The provided difficulty level can be conducive to more reliable and fair evaluation protocols.

C. Social Group Annotation. People in a scene may form different social groups [13], while each group is engaged with a social activity. To provide group annotation, a unique group ID is assigned to those belonging to the same social group in each frame and this assignment may vary over time. Each group label is tagged with a difficulty level to re-

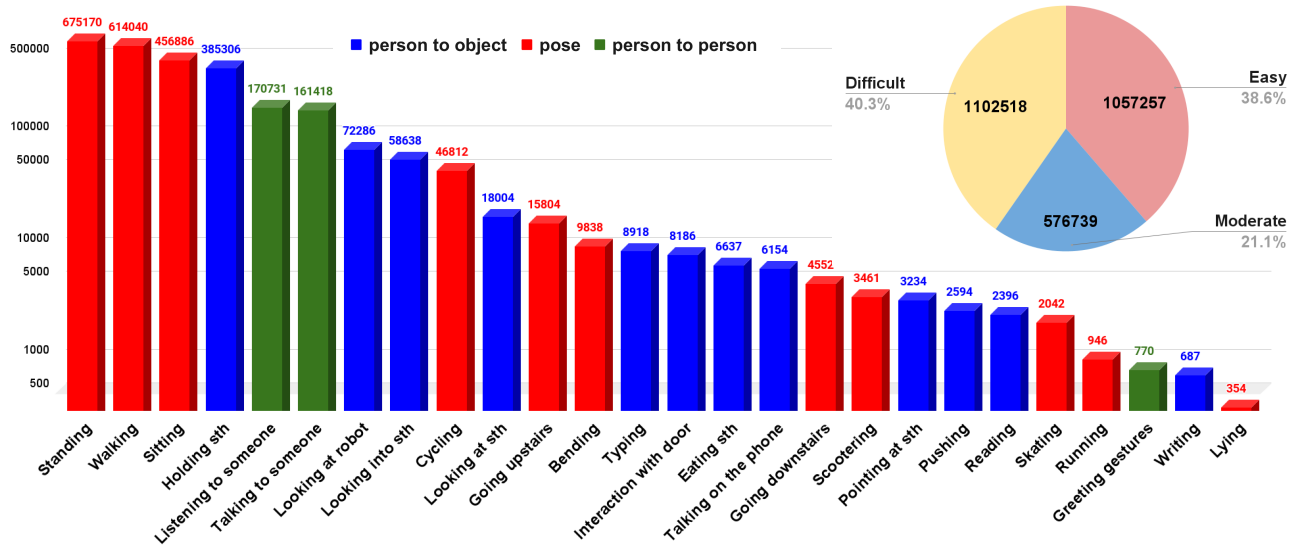

Figure 2. Left: The distribution of action classes in ***log-scale*** sorted by descending order, with colors indicating action types. Right: The distribution of different difficulty levels in action label annotations.

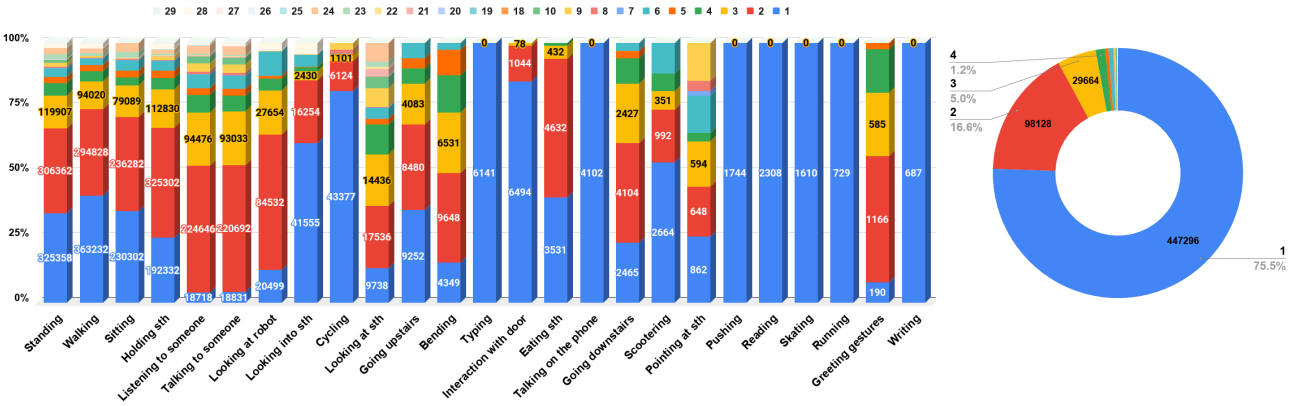

Figure 3. Left: The distribution of social group size for each social activity label shown in individual bars, with colors indicating the size. Right: The distribution of social group size for the entire dataset.

flect the annotator's confidence level. We used *easy, moderate*, and *difficult* for the cases where the group membership was respectively 1) easily recognisable, 2) could be estimated based on some visual and temporal cues, 3) could not be inferred due to the distance from the camera or occlusion. Given the annotated social groups and individuals' action labels in each frame, we generated a *pseudo groundtruth* social activity label for each group using the most frequent individual action labels in that group. We also assign a difficulty level to the inferred social activity labels by averaging the corresponding individual actions' difficulty levels.

D. JRDB-Act splits. Following JRDB splits, JRDB-Act is divided into training, validation, and test sets at video level, thus, all the frames of a video sequence appear in one specific split. The 54 video sequences are split into 20 training, 7 validation, and 27 test videos. For the purpose of consistency with the standard evaluation of other relevant datasets, we evaluate all the task on the key-frames, which are sampled every one second, resulting in 1419 training, 404 validation, and 1802 test samples.

E. Benchmark and Metrics. Our evaluation is performed on the key-frame level following the standard practice in [20]. We adopt the widely used average precision (AP) using an IoU threshold of 0.5, following the standard PASCAL VOC [14] challenge, and customize it to report the performance of each task. To report the performance of social grouping on a set of detected bounding boxes, we first calculate a list of true positive boxes (TP) for each detection confidence threshold. Then, similar to [13, 51], we determine a correspondence between the predicted and truth group IDs by solving an ID assignment between the refined prediction (TP) list and the groundtruth list. Finally

we re-calculate the final number of true positives considering the group IDs and use AP to report the final results. Mean AP (mAP) is also used to report the performance of individual action and social activity detection tasks, following the same practice as in [20]. See supp. material for a comprehensive explanation of our evaluation strategy.

F. JRDB-Act Statistics. Fig. 2 shows the JRDB-Act's distribution of annotated individuals' action labels in log-scale representing a long-tailed distribution in the dataset. Further, in the pie chart of Fig. 2, difficulty level distribution in action labels is reflected in which only 61.4% of action labels are annotated with respect to the visual cues (tagged as *easy* and *moderate*) and the remaining 38.6% are inferred based on the bounding box history or movement (tagged as *difficult*). Fig. 3 demonstrates the distribution of social activity labels with respect to the size of social groups. The donut chart in Fig. 3 indicates the distribution of social group size in the dataset. As illustrated, 75.5%, 16.6%, 5%, 1.2% of social groups consist of one, two, three, and four members respectively and only 1% of the data contain groups with five or more members (maximum 29 members).

4. Proposed Baseline

We propose an end-to-end trainable baseline for spatio-temporal detection of individuals' actions, social groups, and social activities per group in videos. The architecture of our model is illustrated in Fig. 4. We utilise the same backbone $f_\theta(x)$ as in [13] including the I3D feature extractor, the self-attention, and the graph attention modules to extract rich spatio-temporal feature map for each individual in which social interactions are encoded. To further enhance the social grouping performance and to reduce the discrepancy between train and inference compared to [13], we propose to incorporate an eigenvalue-based loss function [10] on the similarity matrix extracted from the visual features and geometrical relations between the detected bounding boxes. Further, in order to overcome the highly unbalanced nature of action labels in the data, we propose to utilise softmax/sigmoid loss partitioning approach inspired by [33].

Learning Social Group Formation. Social groups in a scene can be shown as a graph in which nodes are the individuals and the edges indicate the connectivities between them. The graph of the groundtruth social groups can be presented by a matrix $\hat{A}$ consisting of 0 and 1 in which $\hat{A}_{i,j}$ indicates whether the pair (i, j) belongs to the same social group. A_θ is formed by the model in which for each pair of bounding boxes i and j, the normalised GIoU [39], $D_G(i, j)$, representing a geometrical similarity between each pair is calculated such that 0 and 1 represent far and close boxes, respectively. The normalised similarity between the visual features (extracted from $f_\theta(x)$) of two bounding boxes i and j is also calculated as $D_V(h^i_\theta, h^j_\theta)$. The final $A^{i,j}_\theta$ is then attained by the concatenation of $D_V(h^i_\theta, h^j_\theta)$ and $D_G(i, j)$ and utilising a MLP layer to project the 2-dim vector to a 1-dim vector. The training objective in learning social groups is to reduce the discrepancy between the predicted A_θ and $\hat{A}$. To this end, we utilise a binary cross entropy loss between the elements of A_θ and $\hat{A}$ denoted by L_{BCE} in Eq. 2. Further, since the number of connected components (social groups) in the groundtruth matrix $\hat{A}$ is equal to the number of zero eigenvalues of its laplacian matrix $\hat{\mathrm{L}}$, we want the laplacian matrix of A_θ denoted by L_θ to have the same number of zero eigenvalues as in $\hat{\mathrm{L}}$. To this end, we utilise $L_{eig}(\theta)$ denoted by Eq. 1,

$$L_{eig}(\theta) = \hat{e}^T \mathrm{L}^T_\theta \mathrm{L}_\theta \hat{e} + \alpha \exp(-\beta tr(\bar{\mathrm{L}}^T_\theta \bar{\mathrm{L}}_\theta)) \quad (1)$$

in which $\hat{e}$ is the groundtruth eigenvector corresponding to the zero eigenvalue, L_θ is the laplacian matrix corresponding to the predicted similiraty matrix A_θ and α and β are coefficients. The proof of Eq. 1 is stated in the supp. material. The loss in Eq. 1 is inspired by the fully differentiable, eigendecomposition-free loss proposed in [10] to train a deep network whose loss depends on the eigenvector corresponding to the single zero eigenvalue of a matrix predicted by the network. We extend it to our problem with multiple zero eigenvalues indicating the number of social groups. To learn the number of social groups, as a cardinality loss, we utilise a mean square error function between the groundtruth number of social groups and the 1-dim learned feature from the concatenation of h_θ (max-pool of boxes' visual features) and the summation of the A_θ elements denoted by L_{MSE} in Eq. 2.

$$L_G = L_{BCE}(A_\theta, \hat{A}) + L_{eig}(\mathrm{L}_\theta, \hat{\mathrm{L}}) + L_{MSE}((h_\theta || \sum_i A^i_\theta), GT_{cardinality}) \quad (2)$$

Learning Actions. Each bounding box is annotated with one pose-based and an arbitrary number of interaction-based action labels and the occurrence of action classes is highly unbalanced in the dataset. One naive way to learn actions is to use a cross entropy loss to learn pose-based and a binary cross entropy loss to learn interaction-based actions. However, we empirically observe that action classifier's performance is highly harmed by the unbalanced nature of action labels. To overcome this problem, we divide the pose-based and interaction-based action classes into several disjoint partitions. The number of samples of the least frequent class in each partition, is greater than 0.1 of the number of samples of the highest frequent class in that partition. In each partition excluding the last one, we add an "Other" class which shows the presence of an action class in the less frequent partitions. We have 3 and 4 partitions for pose-based and interaction-based partitions respectively. The list of action labels in each partition is provided in the supp. material. We then train each pose-based and interaction-based partition separately by using

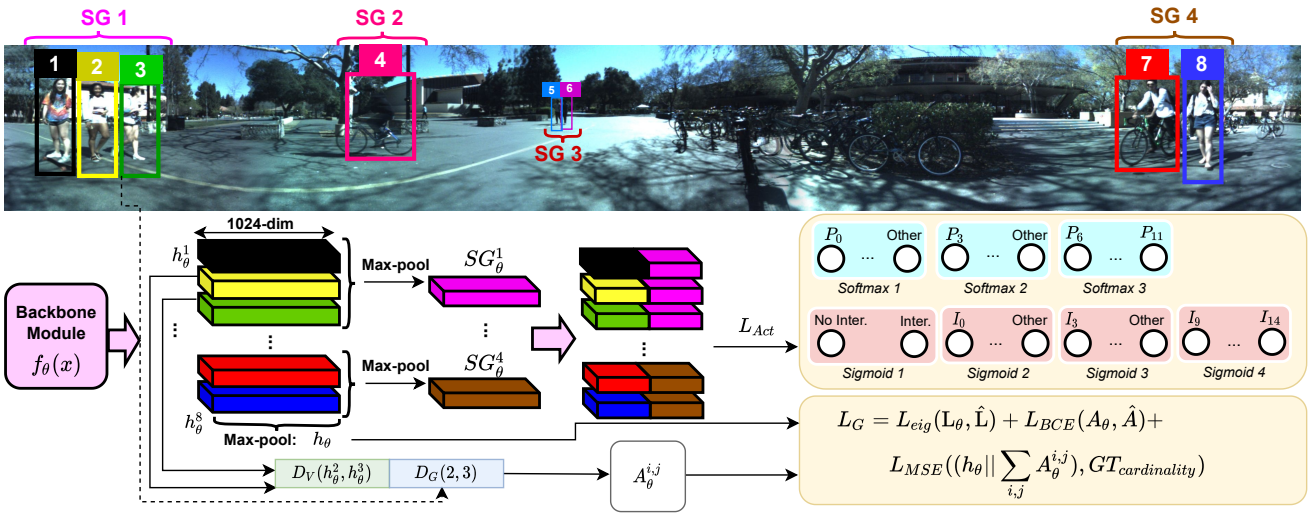

Figure 4. Overview of our framework during training. Given the spatio-temporal feature representation of the individuals denoted by h_θ^i in the key-frame, we optimize two tasks. First, to learn the individual actions, we compute the individual's feature map by concatenating the individual's visual feature and its corresponding social group's feature map (SG_θ^i) obtained by max-pooling the feature maps of its members. Then, to compute L_{Act}, we compute cross entropy and binary cross entropy losses for each pose-based (P) and interaction-based (I) action groups. Second, to learn the social group formation and the social group cardinality, we calculate the similarity matrix A_θ between individuals based on their pair-wise geometrical ($D_G(i,j)$) and feature distance extracted from the backbone ($D_V(h_\theta^i, h_\theta^j)$) and utilise it along with the extracted spatio-temporal feature (h_θ) to compute different loss terms as in L_G.

cross entropy and binary cross entropy losses respectively as in Eq. 3. Further, to maintain the balance, we only train partitions with an existing groundtruth label for each training sample. An illustration of our action learning strategy is shown in Fig. 5.

$$L_{Act} = \sum_{i=0}^{2} \lambda_i L_{CE}(P_\theta^i, P^i) + \sum_{j=0}^{3} \lambda_j L_{BCE}(I_\theta^j, I^j) \quad (3)$$

In Eq. 3, λ is a coefficient, P_θ^i and I_θ^j are the predicted pose-based and interaction-based actions, and P^i and I^j are the corresponding groundtruth labels respectively.

Training. Our model takes as input a video clip with the key-frame located at the end. The input clip is then fed to the backbone to obtain spatio-temporal feature map of the individuals in the key-frame denoted by h_θ^i. The similarity matrix A_θ between individuals is calculated based on their pair-wise geometrical and feature distance. The calculated similarity matrix and the extracted spatio-temporal features are then utilised to learn the social grouping loss L_G denoted by Eq. 2. Given the groundtruth social connections in training, we obtain each social group's feature map by max-pooling the features of its members. Each individual's feature representation is concatenated with its social group feature map. Individual's obtained feature map are utilised to learn the action loss L_{Act} as in Eq. 3. As shown in Fig. 5, For each training sample we only activate the terms of L_{Act} in which there exists a groundtruth label and set the other terms to zero to avoid training with groundtruth vectors of all zeros. The total training objective is stated in Eq. 4.

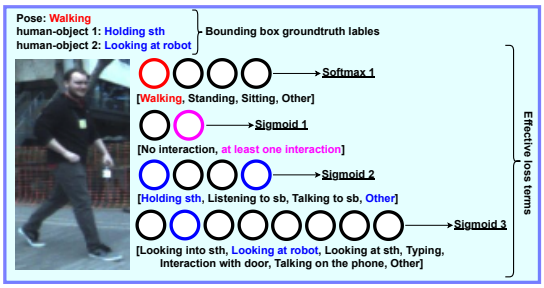

Figure 5. Illustration of different softmax and sigmoid terms of L_{Act} for a training sample. As shown, there are 3 groundtruth actions in this sample including one from the pose-based and two from the human-object interaction categories. For the pose-based action, only one softmax is activated as "Walking" belongs to "Softmax 1". The first sigmoid determines whether there is an interaction-based action. The subsequent sigmoids specifically determine the present interaction-based action labels. Here, "Holding sth" belongs to "Sigmoid 2" and "Looking at robot" falls into the "Other". Thus, the third sigmoid is activated to recognise the "Looking at robot" action.

$$L_{total} = L_G + L_{Act} \quad (4)$$

Inference. At test time, for individual action prediction, we perform softmax operation on the predictions of each cross entropy and sigmoid operation on predictions of each binary cross entropy functions. We then choose the predicted action labels based on a hierarchical approach starting from the first partition and going to the next one in the hierarchy

Method	grouping loss	Cardinality	Geo feature	G1 AP↑	G2 AP↑	G3 AP↑	G4 AP↑	$G5^+$ AP↑	overall AP↑
Baseline1 [13]	BCE	H	-	8.0	29.3	37.5	**65.4**	**67.0**	41.4
Baseline2	BCE	H	✓	26.1	57.0	**61.2**	63.0	53.7	52.2
Baseline3	BCE	MSE	✓	79.6	63.0	43.7	56.9	40.7	56.8
Ours	BCE+EIGEN	MSE	✓	**81.4**	**64.8**	49.1	63.2	37.2	**59.2**

Table 1. Social grouping ablation study on JRDB-Act validation-set using groundtruth bounding boxes. G1, G2, G3, G4, $G5^+$ indicate social groups with 1, 2, 3, 4, 5 or more members.

Method	Action mAP↑
[CE+BCE]	8.0
[W-CE+W-BCE]	8.1
[M-CE+M-BCE] [Ours]	**9.0**

Table 2. Individual action detection ablation study on JRDB-Act validation-set using groundtruth bounding boxes.

only if the "Other" class is predicted. For social group prediction, we perform graph spectral clustering [56] on the obtained similarity matrix between individuals and by utilising the predicted number of social groups. Since the social activity label of each group is the most frequent action labels of its members, we follow the same strategy and infer the activity of each predicted social group from the predicted action labels of its individuals.

5. Experiments

In this section, we provide implementation details of our framework, evaluate different aspects of it, and present a comparison against the existing method proposed in [13].

Implementation Details. The backbone setup and hyper-parameters are identical to [13]. We utilize ADAM optimizer with $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$. α and β in Eq. 1 are set to 1. Since the training objective includes learning social groups and actions and to effectively learn both tasks, we train the model in two stages. In the first stage, we train the model with L_G for 50 epochs with a mini-batch size of 1 and an initial learning rate of 10^{-4}. We then fine-tune the network with L_{total} for 50 epochs. The learning rate is reduced by a factor of 10^{-1} on validation loss plateau. Input video clips to the model are 15 frames long with the annotated key-frame at the end. See the supp. material for more implementation details.

Ablation Studies. All the ablations in Tab. 1 and Tab. 2 are performed using groundtruth bounding boxes on validation-set to remove the effect of detection performance from the experiments. Further, evaluation for each task is performed by considering the corresponding groundtruth labels with easy and moderate difficulty tags and difficult labels are removed from the evaluation. Labels with difficult tags however are used for evaluation on test-set in Tab. 4.

A. Social Group Formation: We compare our framework against three baselines for predicting social groups in terms of grouping AP for groups with different number of members and the average of obtained grouping APs. Within our suggested framework, the network learns to estimate the number of social groups by minimizing a mean square error loss during training; we indicate this by MSE in the cardinality column in Tab. 1. On the contrary, the graph clustering approach used in [13] requires the number of social groups to be known in advance and thus, relies on a heuristic [36] to infer this number; we indicate this with H in the cardinality column. Accordingly, we define three baselines in Tab. 1. [Baseline1] [13], addresses the group formation task with grouping loss consists of a single binary cross entropy based on the visual features of individuals, indicated by BCE in the grouping loss column, and the graph spectral clustering utilizes the heuristic to infer the number of social groups. As validated by our experiments, this heuristic underestimates the number of groups; *i.e.,* spectral clustering tends to group everyone into few or even a single group. Thus, the performance of this baseline for groups with sizes 4, 5 and above is fortuitously better than the other methods while it performs significantly worse on the lower group size categories. [Baseline2] extends [Baseline1] by exploiting geometrical features in addition to the visual features. Evidently, the geometrical features lead to better performance in identification of small-sized social groups. Similarly, [Baseline3] extends [Baseline2] by learning the social group cardinality instead of adopting the heuristic; this significantly boosts the group formation performance for small-sized groups. Finally, [Ours] shows the effect of utilizing the eigen-value loss in our framework which yields the highest overall group formation results.

B. Action and Social Activity Prediction: We show the effectiveness of our proposed strategy (*i.e.* loss partitioning) to deal with highly unbalanced individuals' action labels in Tab. 2. [Baseline1], utilizes a single cross entropy loss and a single binary cross entropy loss [CE+BCE] to learn pose-based and interaction-based action classes respectively. [Baseline2] [13], utilizes the cross entropy loss and the binary cross entropy loss functions in a weighted manner [W-CE+W-BCE]. Normalized weights of action labels is calculated based on the inverse of their occurrence frequency in train and validation sets. Finally, in [Ours], we utilize the loss partitioning strategy using multiple cross entropy and binary cross entropy losses [M-CE+M-BCE] as elaborated in Sec. 4. As validated by our experiments, weighting strategy does not address the unbalanced distribution of action classes in the data, whereas the proposed

Method	G1 AP↑	G2 AP↑	G3 AP↑	G4 AP↑	G5^{+} AP↑	overall AP↑	Action mAP↑	G-Act mAP1↑	G-Act mAP2↑
[13]+Faster-RCNN	9.5	24.3	21.2	**39.8**	10.8	21.1	4.4	3.5	1.3
[13]+MMPAT	11.8	27.5	22.4	38.8	**24.6**	25.0	4.9	3.5	1.3
Ours+Faster-RCNN	42.5	**40.8**	23.1	25.6	13.4	29.1	5.3	4.4	3.4
Ours+MMPAT	**56.6**	39.5	**24.3**	22.4	14.8	**31.5**	**5.4**	**4.7**	**3.4**

Table 3. Final results of our model against [13] on JRDB-Act test-set using two different sets of detection bounding boxes (Faster-RCNN [38] and MMPAT [21]) and by considering labels with Easy and Moderate difficulty tags in evaluation.

[Ours]	G1↑	G2↑	G3↑	G4↑	G5^{+}↑	overall AP↑	Action mAP↑	G-Act mAP1↑	G-Act mAP2↑
[E,M,D]	34.9	37.3	18.3	16.4	7.6	22.9	4.4	3.5	2.7
[E,M]	42.5	40.8	23.1	25.6	13.4	29.1	5.3	**4.4**	3.4
[E]	**44.4**	**42.7**	**27.1**	**28.4**	**13.9**	**31.3**	**5.7**	**4.4**	**3.5**

Table 4. E:Easy, M: Moderate and D:Difficult. Performance of different tasks wrt the difficulty tag on JRDB-Act test-set.

loss partitioning approach shows improvement compared to the baselines. Finally, social activity labels are inferred from the predicted social groups and individuals' actions as the most frequent actions performed by the members of that group. Social activity labels are evaluated by ignoring social groups indicated by the G-Act mAP1 column (similar to individual actions evaluation) and by considering social groups indicated by G-Act mAP2. For G-Act mAP2, we consider a true positive as a box for which the social group and the social activity labels are correctly predicted.

Test-set Results. In Tab. 3, we show that our suggested framework outperforms [13], on JRDB-Act test-set using the public detection provided in the JRDB benchmark [35] obtained from Faster-RCNN [38] in each task. It is worth noting that the performance of Faster-RCNN on JRDB-Act test-set is 52.2 mAP which shows the complexity of the dataset in the detection task. To study the effect of detection on the performance of each task, we utilized MMPAT [21], a better-performing detection on JRDB, with 68.1 mAP on test-set in Tab. 3 and realized that more accurate detection boosts the grouping performance by a large margin. However, it performs almost on par with Faster-RCNN detection boxes on individual action and social activity detection tasks. This finding shows that understanding human actions in JRDB-Act is inherently complex due to the unique challenges in the data including robot motion and camera perspective. These challenges and results highlight the need of the existing research methods, including human activity detection frameworks, to support this new application in these type of environments which are underrepresented in existing datasets. In Tab. 4, we further investigate the effect of the annotated difficulty tag for each provided label including social group, individual action and social activity labels in the evaluation of each task. As observed, using only easy labels indicated by E in evaluation, yields the best performance. Using easy and moderate tags [E,M], also used in ablation studies, perform worse with a relatively small gap compared to [E] and using all the labels with easy, moderate and difficult tags [E,M,D] performs the worst with a large gap compared to [E,M].

Limitation and Discussion. The model's performance in the given tasks relies on the detector performance in predicting individuals' boxes as well as the model's performance in classification and clustering of the detected boxes. The current low action mAP in Tab. 2 using groundtruth bounding boxes, evaluated on easy and moderate action labels as well as the negligible effect of utilising more accurate detected bounding boxes as validated in Tab. 3, show the inherent complexity and challenges of JRDB-Act in understanding human actions due to the motion of the robot, camera perspective, and highly unbalanced action distribution with different difficulty levels. Thus, this dataset may challenge the existing action localization frameworks, demanding further research in this direction to tackle the associated unique complexities. Moreover, JRDB-Act is a multi-modal dataset and provides annotation for 3D data which can potentially contribute to the overall performance of the tackled tasks. However, utilising 3D input can mainly contribute to the downstream tasks, *e.g.* detection, tracking and extracting more accurate geometrical features. A better detection in turn, results in higher social grouping performance as substantiated in Tab. 3. Exploring the 3D sensor modality data and investigating sensor data fusion strategies can be considered as potential future work.

6. Conclusion

Learning to recognise human actions and their social groups in an unconstrained environment including crowded scenarios, with potentially highly unbalanced human daily actions from a stream of sensory data captured from a mobile robot, remains a challenge, due to the lack of a reflective large-scale dataset. In this paper, we introduced JRDB-Act, a dataset captured from a moving social robot platform, including spatio-temporal individual action and social group annotations conducive to the task of simultaneously detection of social groups, individual actions and social activities. We also developed an end-to-end trainable pipeline to serve as a baseline to tackle this multi-task problem. We believe the dense annotations, and natural complexities of JRDB-Act pose new challenges for future research in the vision and robotics community.

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. 2

[2] Yizhak Ben-Shabat, Xin Yu, Fatemeh Saleh, Dylan Campbell, Cristian Rodriguez-Opazo, Hongdong Li, and Stephen Gould. The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose. In *WACV*, pages 847–859, 2021. 2

[3] 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, 2

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 3

[5] Wongun Choi and Silvio Savarese. A unified framework for multi-target tracking and collective activity recognition. In *ECCV*, pages 215–230, 2012. 3

[6] Wongun Choi and Silvio Savarese. Understanding collective activitiesof people from videos. *IEEE transactions on pattern analysis and machine intelligence*, 36(6):1242–1257, 2013. 3

[7] Wongun Choi, Khuram Shahid, and Silvio Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *ICCVW*, pages 1282–1289, 2009. 2, 3

[8] Wongun Choi, Khuram Shahid, and Silvio Savarese. Learning context for collective activity recognition. In *CVPR*, pages 3273–3280, 2011. 3

[9] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, pages 720–736, 2018. 1, 2

[10] Zheng Dang, Kwang Moo Yi, Yinlin Hu, Fei Wang, Pascal Fua, and Mathieu Salzmann. Eigendecomposition-free training of deep networks with zero eigenvalue-based losses. In *ECCV*, pages 768–783, 2018. 5

[11] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Manohar Paluri, Jürgen Gall, Rainer Stiefelhagen, and Luc Van Gool. Large scale holistic video understanding. In *ECCV*, pages 593–610. Springer, 2020. 2

[12] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, pages 2625–2634, 2015. 3

[13] Mahsa Ehsanpour, Alireza Abedin, Fatemeh Saleh, Javen Shi, Ian Reid, and Hamid Rezatofighi. Joint learning of social groups, individuals action and sub-group activities in videos. In *ECCV*, 2020. 2, 3, 4, 5, 7, 8

[14] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015. 4

[15] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, pages 6202–6211, 2019. 3

[16] Rohit Girdhar, João Carreira, Carl Doersch, and Andrew Zisserman. A better baseline for ava. *arXiv preprint arXiv:1807.10066*, 2018. 3

[17] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *CVPR*, pages 244–253, 2019. 3

[18] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. *IEEE transactions on pattern analysis and machine intelligence*, 29(12):2247–2253, 2007. 2

[19] 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*, page 5, 2017. 2

[20] 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, 2, 3, 4, 5

[21] Yuhang He, Wentao Yu, Jie Han, Xing Wei, Xiaopeng Hong, and Yihong Gong. Know your surroundings: Panoramic multi-object tracking by multimodality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2969–2980, 2021. 8

[22] Mostafa S Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *CVPR*, pages 1971–1980, 2016. 2, 3

[23] 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

[24] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards understanding action recognition. In *ICCV*, pages 3192–3199, 2013. 2

[25] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, pages 1725–1732, 2014. 2

[26] 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

[27] Yan Ke, Rahul Sukthankar, and Martial Hebert. Efficient visual event detection using volumetric features. In *ICCV*, volume 1, pages 166–173, 2005. 2

[28] 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, 2011. 2

[29] Ang Li, Meghana Thotakuri, David A. Ross, João Carreira, Alexander Vostrikov, and Andrew Zisserman. The ava-kinetics localized human actions video dataset, 2020. 2

[30] Dong Li, Zhaofan Qiu, Qi Dai, Ting Yao, and Tao Mei. Recurrent tubelet proposal and recognition networks for action detection. In *ECCV*, pages 303–318, 2018. 3

[31] Shuaicheng Li, Qianggang Cao, Lingbo Liu, Kunlin Yang, Shinan Liu, Jun Hou, and Shuai Yi. Groupformer: Group activity recognition with clustered spatial-temporal transformer. In *ICCV*, 2021. 3

[32] Wenbo Li, Ming-Ching Chang, and Siwei Lyu. Who did what at where and when: simultaneous multi-person tracking and activity recognition. *arXiv preprint arXiv:1807.01253*, 2018. 3

[33] Yu Li, Tao Wang, Bingyi Kang, Sheng Tang, Chunfeng Wang, Jintao Li, and Jiashi Feng. Overcoming classifier imbalance for long-tail object detection with balanced group softmax. In *CVPR*, pages 10991–11000, 2020. 5

[34] Marcin Marszałek, Ivan Laptev, and Cordelia Schmid. Actions in context. Citeseer, 2009. 2

[35] Roberto Martín-Martín*, Mihir Patel*, Hamid Rezatofighi*, Abhijeet Shenoi, JunYoung Gwak, Nathan Dass, Alan Federman, Patrick Goebel, and Silvio Savarese. JRDB: A dataset and benchmark of egocentric robot visual perception of humans in built environments. *IEEE transactions on pattern analysis and machine intelligence*, 2021. 2, 3, 8

[36] Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *NIPS*, pages 849–856, 2002. 7

[37] Jamie Ray, Heng Wang, Du Tran, Yufei Wang, Matt Feiszli, Lorenzo Torresani, and Manohar Paluri. Scenes-objects-actions: A multi-task, multi-label video dataset. In *ECCV*, pages 635–651, 2018. 2

[38] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016. 8

[39] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*, pages 658–666, 2019. 5

[40] Mikel D Rodriguez, Javed Ahmed, and Mubarak Shah. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, pages 1–8, 2008. 2

[41] Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. A database for fine grained activity detection of cooking activities. In *CVPR*, pages 1194–1201, 2012. 2

[42] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local svm approach. In *ICPR*, volume 3, pages 32–36, 2004. 2

[43] Abhijeet Shenoi, Mihir Patel, JunYoung Gwak, Patrick Goebel, Amir Sadeghian, Hamid Rezatofighi, Roberto Martin-Martin, and Silvio Savarese. JRMOT: A real-time 3d multi-object tracker and a new large-scale dataset. In *IROS*, 2020. 2

[44] Gunnar A Sigurdsson, Santosh Divvala, Ali Farhadi, and Abhinav Gupta. Asynchronous temporal fields for action recognition. In *CVPR*, pages 585–594, 2017. 3

[45] 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, 2016. 1, 2, 3

[46] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *arXiv preprint arXiv:1406.2199*, 2014. 3

[47] 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

[48] Chen Sun, Abhinav Shrivastava, Carl Vondrick, Kevin Murphy, Rahul Sukthankar, and Cordelia Schmid. Actor-centric relation network. In *ECCV*, pages 318–334, 2018. 3

[49] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *CVPR*, pages 1207–1216, 2019. 2

[50] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *CVPR*, pages 284–293, 2019. 3

[51] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *ICML*, pages 478–487, 2016. 4

[52] Huijuan Xu, Abir Das, and Kate Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. In *CVPR*, pages 5783–5792, 2017. 3

[53] Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *International Journal of Computer Vision*, 126(2-4):375–389, 2018. 2

[54] Junsong Yuan, Zicheng Liu, and Ying Wu. Discriminative subvolume search for efficient action detection. In *CVPR*, pages 2442–2449, 2009. 2

[55] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *CVPR*, pages 4694–4702, 2015. 3

[56] Lihi Zelnik-Manor and Pietro Perona. Self-tuning spectral clustering. In *NIPS*, pages 1601–1608, 2005. 7

[57] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. Hacs: Human action clips and segments dataset for recognition and temporal localization. In *ICCV*, pages 8668–8678, 2019. 2

[58] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *ECCV*, pages 803–818, 2018. 3