communications engineering

Article

https://doi.org/10.1038/s44172-024-00272-7

Localization and recognition of human action in 3D using transformers

Check for updates

Jiankai Sun [1] ✉, Linjiang Huang[2], Hongsong Wang[3], Chuanyang Zheng[4], Jianing Qiu [5] ✉, Md Tauhidul Islam [6], Enze Xie[7], Bolei Zhou[8], Lei Xing [1,6], Arjun Chandrasekaran[9] & Michael J. Black[9]

Understanding a person's behavior from their 3D motion sequence is a fundamental problem in computer vision with many applications. An important component of this problem is 3D action localization, which involves recognizing what actions a person is performing, and when the actions occur in the sequence. To promote the progress of the 3D action localization community, we introduce a new, challenging, and more complex benchmark dataset, BABEL-TAL (BT), for 3D action localization. Important baselines and evaluating metrics, as well as human evaluations, are carefully established on this benchmark. We also propose a strong baseline model, i.e., Localizing Actions with Transformers (LocATe), that jointly localizes and recognizes actions in a 3D sequence. The proposed LocATe shows superior performance on BABEL-TAL as well as on the large-scale PKU-MMD dataset, achieving state-of-the-art performance by using only 10% of the labeled training data. Our research could advance the development of more accurate and efficient systems for human behavior analysis, with potential applications in areas such as human-computer interaction and healthcare.

Understanding a person's behavior in the 3D world is a fundamental and important problem[1–9]. A momentous step towards solving this is identifying what actions a person is performing, as well as when, where, and why the actions are performed. 3D Temporal Action Localization (3D-TAL) involves recognizing actions that a person is performing in a 3D motion sequence, and locating precise start and end times of each action, as illustrated in Fig. 1a. This research has various potential applications[10–12]. For example, 3D action localization may enable the automatic retrieval of semantically relevant movements for graphic artists and game designers who use large databases of 3D data to animate virtual characters. In contrast to RGB temporal action localization, 3D-TAL focuses on the body motion alone, factoring out effects of the image texture and lighting variation, which can result in bias in action recognition[9,13]. Compared to RGB temporal action localization, 3D-TAL task using motion-captured data provides several advantages. It offers accurate spatial and temporal information[14], robustness to occlusions[15,16], invariance to viewpoint changes[17], and enables fine-grained analysis of body dynamics[18,19]. These benefits make it a valuable approach for a range of applications, including Human-Computer Interaction (HCI), animation, and AR/VR systems[20–22]. Furthermore, there is a potential for integrating Large Language Models (LLMs) with 3D-TAL, opening up new possibilities for analysis and synthesis[23–25].

Despite the decreasing cost of 3D sensors, e.g., Kinects, phones, and iPads, and the growing accessibility to 3D human movement data[26–29], progress in 3D-TAL[30–33] has nearly stagnated in recent years. We argue that the lack of a suitable benchmark dataset limits the progress on this task. Current benchmarks are recorded in constrained indoor environments, leading to a lack of data diversity. Taking the large-scale PKU-MMD[32] dataset as an example, although the number of action categories is relatively large, the intra-class variance is actually low. A proof-of-concept experiment shows a mean Average Precision (mAP) of 91.9% could be reached by using 10% of the labeled training data.

In light of the limitations of current benchmarks and the necessity of pushing forward the development of 3D-TAL, we introduce a new benchmark, namely BABEL-TAL (BT). This dataset is carefully built on the recent BABEL dataset[34]. With the aim of thoroughly assessing the model's performance, we systematically categorize the classes within BABEL-TAL into two tiers of granularity, encompassing 20 and 60 classes, respectively. As a result, two distinct subsets are created: BABEL-TAL-20 (BT-20) and BABEL-TAL-60 (BT-60). Moreover, a comprehensive dataset,

[1]School of Engineering, Stanford University, Stanford, CA, USA. [2]Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong, Hong Kong SAR. [3]Department of Computer Science and Engineering, Southeast University, Nanjing, Jiangsu, China. [4]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, Hong Kong SAR. [5]Department of Biomedical Engineering, The Chinese University of Hong Kong, Hong Kong, Hong Kong SAR. [6]Department of Radiation Oncology, Stanford University, Stanford, CA, USA. [7]Department of Computer Science, The University of Hong Kong, Hong Kong, Hong Kong SAR. [8]Department of Computer Science, University of California, Los Angeles, Los Angeles, CA, USA. [9]Perceiving Systems Department, Max Planck Institute for Intelligent Systems, Tübingen, Germany. ✉e-mail: jksun@cs.stanford.edu; jianingqiu@cuhk.edu.hk

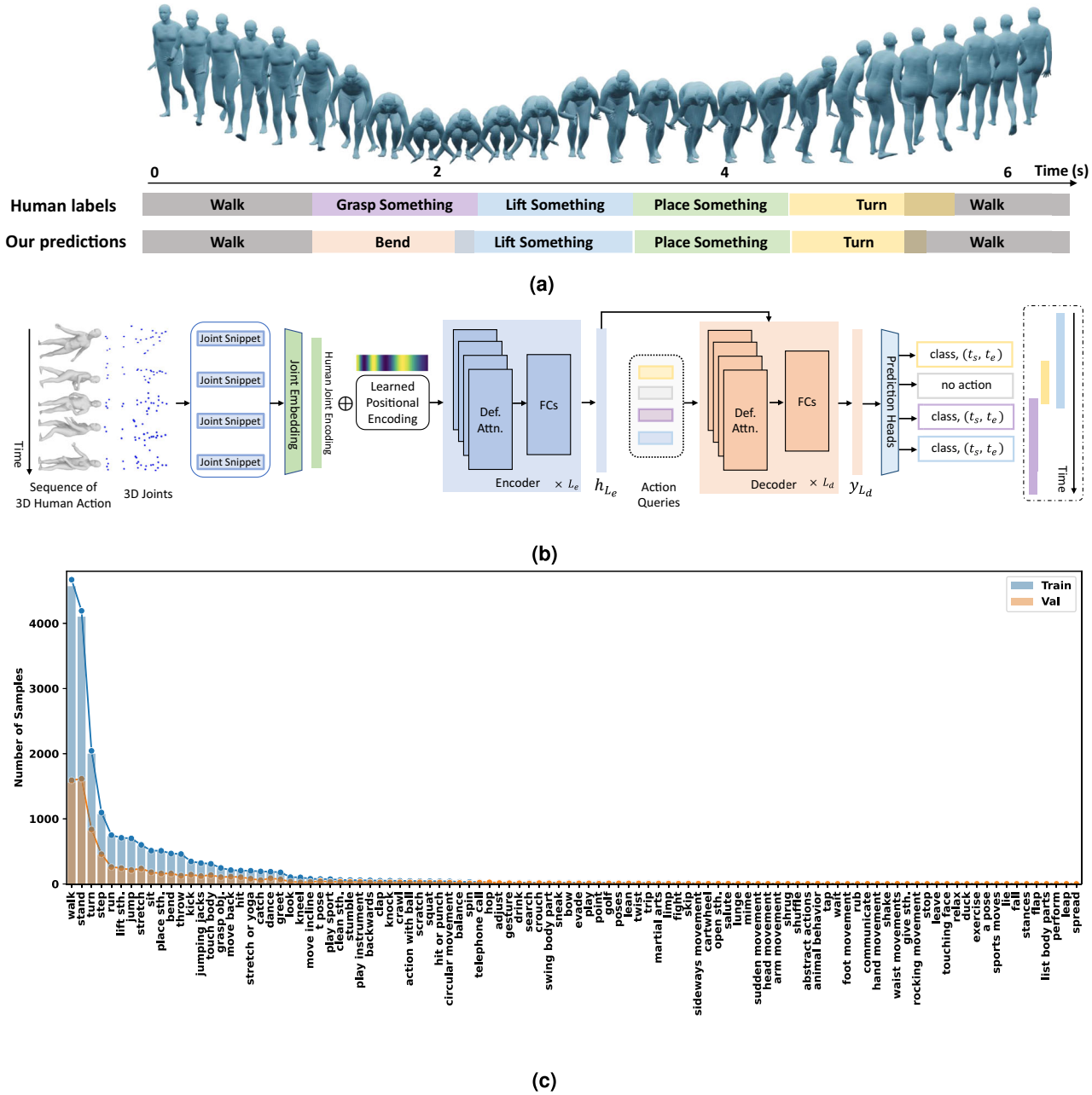


Fig. 1 | Overview of 3D-TAL, LocATe, and BT-ALL. a 3D-TAL Task Description: 3D Temporal Action Localization (3D-TAL) involves identifying actions and their precise spans (start and end times) in a 3D motion sequence. We compare the human-provided labels for the 3D motion with predictions from our proposed LocATe approach. Predictions from LocATe correlate well with human labels, including simultaneous actions (visualized as temporal overlaps between different action spans). LocATe produces accurate localizations, and meaningful actions (even when disagreeing with the human label, e.g., "Grasp Something" vs. "Bend"). **b** LocATe Framework: Given a sequence of human poses, LocATe outputs a set of action spans via an encoding-decoding paradigm. **c** Class frequency distributions of the introduced BABEL-TAL-ALL (BT-ALL): This dataset offers a rich spectrum of action labels and demonstrates intra-class diversity. Additionally, the distribution of action data closely follows a long-tailed pattern, mirroring real-world scenarios.

encompassing more fine-grained action classes, is designated as BABEL-TAL-ALL (BT-ALL). In applications involving human-computer interaction, precise action localization is crucial for accurately interpreting and responding to human actions. For example, a domestic robot equipped with fine-grained action recognition capabilities could accurately locate and retrieve household items, assist with cooking or cleaning tasks, and provide support for elderly or disabled individuals in daily activities. Our datasets specifically address these objectives by providing fine-grained annotations and a diverse range of activities, which we believe will be instrumental for future research and applications.

Actions in the BT dataset vary widely in complexity, and samples have large intra-class variances. As shown in Table 1, compared with previous datasets, BT contains more complex movement sequences, with a long-tailed distribution of actions. In fact, experimental results have demonstrated that BT-20 poses a challenge for existing 3D-TAL methods. We observe that a previous method[31] that could achieve mAP of 81.1% on PKU-MMD[32], could only achieve a surprisingly low mAP of 11.4% on BT-20. In order to further the research in this field, it is essential to curate new challenging datasets, such as BT, that encompass a broader range of complex scenarios and diverse contexts, spanning various application domains, to

Table 1 | Comparison of existing 3D action localization datasets

Datasets	Classes	Sequences	Instances	Subjects	Modalities	Year	Duration
G3D[46]	20	210	1467	10	RGB, D, S	2012	—
CAD-120[76]	20	120	~1200	4	RGB, D, S	2013	—
Comp. Act[49].	16	693	2529	14	RGB, D, S	2014	—
Watch-N-Patch[48]	21	458	~ 2500	7	RGB, D, S	2015	230 min
OAD[77]	10	59	~ 700	—	RGB, D, S	2016	216 min
PKU-MMD[32]	51	1076	21545	66	RGB, D, IR, S	2017	3000 min
Wei et al.[51]	35	201	—	—	RGB, D, S	2020	—
BABEL-TAL-20	20	5727	6244	346	3D Mesh, S	2024	—
BABEL-TAL-60	60	6808	7332	584	3D Mesh, S	2024	—
BABEL-TAL-ALL	102	8808	9617	925	3D Mesh, S	2024	2580 min

The BABEL-TAL (BT) dataset stands out from existing 3D action localization datasets in several key ways. Firstly, it pioneers in using 3D motion-capture data to provide precise body joint movements for temporal action localization. Secondly, this dataset comprises an extensive range of action labels and showcases substantial intra-class diversity. Thirdly, the action data adheres to a long-tailed distribution, mirroring real-world scenarios. Lastly, the dataset contains continuous actions in extended motion sequences, free from environmental or actor constraints. *D* Depth, *S* Skeleton, *IR* Infrared Radiation.

Table 2 | Quantitative evaluations on the BABEL-TAL-20 (BT-20) dataset

Method	tIoU									mAP
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	
Beyond-Joints[31]	14.3	13.6	13.3	12.3	11.4	10.5	8.9	6.2	4.1	10.5
ASFD[72]	24.2	23.1	22.6	22.2	21.9	20.4	18.9	12.2	9.0	19.3
SRN[51]	25.1	24.0	22.7	21.7	20.1	18.2	16.7	15.9	10.4	19.4
TSP[73]	26.9	25.6	24.1	23.0	22.5	20.4	17.1	13.0	10.1	20.3
G-TAD[37]	25.1	24.1	23.9	23.0	22.1	21.1	18.5	14.1	11.8	20.4
AGT[38]	27.3	26.0	25.7	24.5	23.4	21.5	19.4	15.9	12.4	21.9
ActionFormer[39]	30.4	27.1	25.3	25.1	24.5	22.7	20.6	16.1	12.0	22.6
LocATe	43.5	41.1	41.0	38.2	35.1	30.5	23.7	16.4	9.99	31.1
LocATe w/ tricks	46.6	45.5	43.0	40.2	36.0	30.5	23.7	15.9	9.78	32.0

We report the AP with the tIoU in the range [0.1, 0.9] as well as the mAP. LocATe represents our single-stage transformer-based approach, while LocATe w/ tricks refers to our method enhanced with tricks, including iterative bounding box refinement and a two-stage decoder[40]. Notably, our approach LocATe outperforms the previous method Beyond-Joints, with particularly substantial improvements at lower tIoU thresholds when compared to other benchmark methods. *AP* Average Precision, *tIoU* threshold IoU, *mAP* mean Average Precision.

provide a benchmark database for a holistic evaluation of different methods. This dataset will serve to validate algorithms' robustness and generalization capabilities, inspiring researchers to develop more robust and adaptable methods. In order to establish a robust baseline model that can serve as a foundation for future research on BT, we draw upon the recent advancements in Transformer architecture and present our approach, called Localizing Actions with Transformers (LocATe). LocATe is designed to address the challenges in 3D-TAL by effectively capturing global correlations among frames in extended sequences. Despite its simplicity, LocATe proves to be highly competitive, demonstrating promising results in the task at hand. Unlike other vision transformers[35,36] based on image patches or features, LocATe takes a sequence of 3D human joint positions as the input. Given the 3D skeleton sequence, we formulate action localization and recognition as an action-span set prediction problem which enables the proposed method to trivially model simultaneous actions. In a nutshell, LocATe is an end-to-end approach for 3D action localization, much simpler and more accurate than existing multi-stage methods.

Contributions. (1) We introduce a large-scale and challenging dataset BABEL-TAL (BT) for 3D action localization and establish important baselines on this dataset. (2) We empirically find that there exist human label disagreements in annotations for 3D action localization and devise a human evaluation method to complement the automatic mAP evaluation. (3) We present an end-to-end baseline model named LocATe, which leverages the Transformer architecture and incorporates deformable attention mechanisms. Our proposed model achieves superior performance on the BT-20 dataset and other public benchmark datasets, surpassing the state-of-the-art methods, with a remarkable mean Average Precision (mAP) of 91.9% attained using just 10% of the labeled training data on the PKU-MMD dataset.

Results

In this section, we evaluate and analyze the performance of existing 3D-TAL models as well as LocATe on BT, focusing primarily on the BT-20 benchmark, given its exemplar nature within BT series and the complexities that current methods face when dealing with the more difficult BT-60 and BT-ALL benchmarks. To attain further insights into the proposed BT benchmarks and the baseline model LocATe, we also compare with the results on existing popular datasets. Analyses and discussions are provided. The mean Average Precision (mAP) at various temporal IoU thresholds is used to measure the performance.

Experimental Setup

Our model is implemented using PyTorch and trained on an NVIDIA A40 GPU. We compare our approach with 3D-TAL baselines such as Beyond-Joints[31] and Cui et al.[30]. Additionally, we adapted 2D-TAL methods as baselines but found that the pre-trained backbones they used on 2D video data were not suitable for 3D-TAL. Therefore, we replaced them with similarly sized CNN or Transformer architectures. We first train both the

Table 3 | Comparison of the proposed method with previous methods on the PKU-MMD dataset (mAP@tIoU = 0.5)

Method	Cross-view	Cross-subject
JCRRNN[77]	53.3	32.5
TAP-B-M[78]	48.6	35.2
Beyond-Joints[31]	91.1	81.1
Cui et al.[30]	93.3	83.5
LocATe	94.6	93.2

LocATe attains state-of-the-art performance and stands out by surpassing the previous methods by almost 10% in the context of cross-subject evaluation. *PKU-MMD* Peking University Multi-Modal Dataset, *mAP* mean Average Precision, *tIoU* threshold IoU.

baselines and our LocATe model from scratch on the BT-20 dataset and evaluate their performance (Table 2). To better understand the differences between BT-20 and previous 3D benchmarks, we also train and evaluate the baselines and LocATe from scratch on PKU-MMD, a popular 3D-TAL benchmark prior to BT-20 (Table 3).

Comparative analysis

Performance of baselines. Table 2 shows the experimental results of baseline models and LocATe on the BT-20 dataset. Firstly, it is evident that all baseline models yield notably low mAPs when evaluated on the BT-20 dataset. All models, including the baselines and our LocATe, are trained and evaluated from scratch on BT-20. For instance, considering the Beyond-Joints[31] model, which achieves an impressive 91.1% mAP in a cross-view setting on the PKU-MMD dataset, it completely fails on the BT-20 dataset, with a mAP of only 10.5%. This substantial drop in performance underscores the considerable challenge posed by the BT benchmarks. Besides, we observe that the G-TAD baseline[37] outperforms the RNN-based prior work Beyond-Joints[31]. This suggests that modeling global context, including background, is more effective than only modeling local temporal neighborhoods. Graph-based approaches such as AGT[38], which can also model long-term context, outperform Beyond-Joints by larger margins. While ActionFormer[39] utilizes Local Self-Attention, we believe LocATe's Deformable Attention is more effective, as it captures local information with variable receptive fields. Furthermore, the multiscale approach of ActionFormer is less applicable to 3D data, which is different from 2D images and does not require scaling.

Performance of LocATe. LocATe w/ tricks denotes the proposed method using tricks of iterative bounding box refinement and a two-stage decoder[40], whereas LocATe is the plain version. We can observe that our approach LocATe outperforms the prior method Beyond-Joints. The performance increases of LocATe over other comparative methods are evident for lower threshold IoU (tIoU) thresholds. For downstream tasks such as recognition and retrieval, a detection model with a high recall is more important than achieving high precision. Furthermore, we observe that iterative bounding box refinement and two-stage decoders can further improve the performance of the algorithm. The two-stage decoder first proposes action spans, which are then provided as action query features to the decoder. These improvements present an interesting trade-off between model complexity and performance. It also suggests a potential direction for improvement by learning better action features.

Analysis of Different Actions

Temporal action localization involves two sources of error: localization and recognition. To further understand the recognition error, we visualize the confusion matrix of action classification on predicted spans in Fig. 2. We find that the action "stand" is confused with many other actions. This is understandable since in the ground-truth data, a person is rarely labeled as "stand" when other actions are present. This shows that our dataset is challenging and diverse, and our annotation is fine-grained and accurate. Besides, we also find that the actions "throw" and "catch" are often confused by the model, as they frequently occur in succession in the motion sequence. The recognition error may be largely due to the localization error, which causes misalignment between predicted and ground-truth action spans. We find that despite the low precision, "grasp something" is in fact, predicted frequently, and often confused with "lift something" because they may share similar motion patterns, features, and temporal context.

Figure 3 shows how the Average Precision (AP) for each action varies across different tIoU thresholds. We observe a large variance in the AP of each class, both in terms of absolute values and sensitivity to tIoU settings. For example, we find that "run" and "walk" achieve the highest AP scores, while "touch body part" has the lowest AP scores. This may be because the former actions are easier to distinguish from others, while the latter actions are more ambiguous and complex. "exercise/yoga" and "touch body part" are highly context-dependent, meaning their correct identification relies heavily on the surrounding context. Without clear contextual cues, the model may struggle to accurately predict these actions. We also notice that increasing the number of samples per class improves the performance, as it provides more training data for the model. However, other contributing factors, such as the intra-class variance, also affect the AP of each class. For instance, LocATe achieves reasonable AP for "place something", despite having fewer labeled data than some other classes.

Human Experiments

Different from labeling the bounding box of an object, labeling an action and marking its precise start and end frames in a video is a highly subjective task. Invariably, there exists disagreement between human labels for the same motion sequence. However, the definition of "ground-truth" labels typically only considers an annotation from one person. This implies that although the upper bound of the mAP metric is 100.0%, an approach might perform quite well even if its mAP $\ll$ 100.0%.

Experiments of Human Evaluation. To address the above limitation of automatic evaluation of the mAP metric for 3D action localization, we use an evaluation method that directly compares the performance of two approaches via a head-to-head human evaluation. We simultaneously present an evaluator with two videos of the same motion sequence. Human evaluators are recruited from the crowdsourcing platform Amazon Mechanical Turk. The task interface is illustrated in Fig. 4. The evaluator answers the following question – Which labels better match the motion (left or right)? The task interface for the human study is provided in the Supplementary human_eval/task.html. Note that the sequence and labels are random in the provided sample. The order in which the labels from the different methods appear (left or right) is randomized across trials. As an example, we compare LocATe with Beyond-Joints[31] and G-TAD[37], respectively. We collect votes from 5 unique evaluators for each motion sequence to account for subjectivity. We observe that human evaluators prefer labels from LocATe 69.20% of the time, **higher** than 30.80% of Beyond-Joints. We also observe that LocATe outperforms G-TAD with 55.31% vs. 44.69%. It is encouraging that the mAP metric and human evaluation appear to be correlated. The differences between Beyond-Joints and LocATe are larger than the differences between G-TAD and LocATe under both metrics.

Results on PKU-MMD

PKU-MMD[32] is a large-scale dataset for 3D action localization and is recorded using Kinect v2. It contains 1076 long videos performed by 66 subjects from 3 different camera views. Each video has more than 20 action instances from 51 action categories. We follow the same cross-subject and cross-view evaluation settings[32] and compute the mAP of different actions. Table 3 shows the comparison of the proposed method with previous methods. LocATe achieves state-of-the-art performance and in particular, it outperforms prior arts by nearly 10% in cross-subject evaluation.

We further conduct experiments that only use a fraction of training data to train 3D action localization models. We compare the results with Beyond-Joints[31] under the same settings in Fig. 5. We find that the

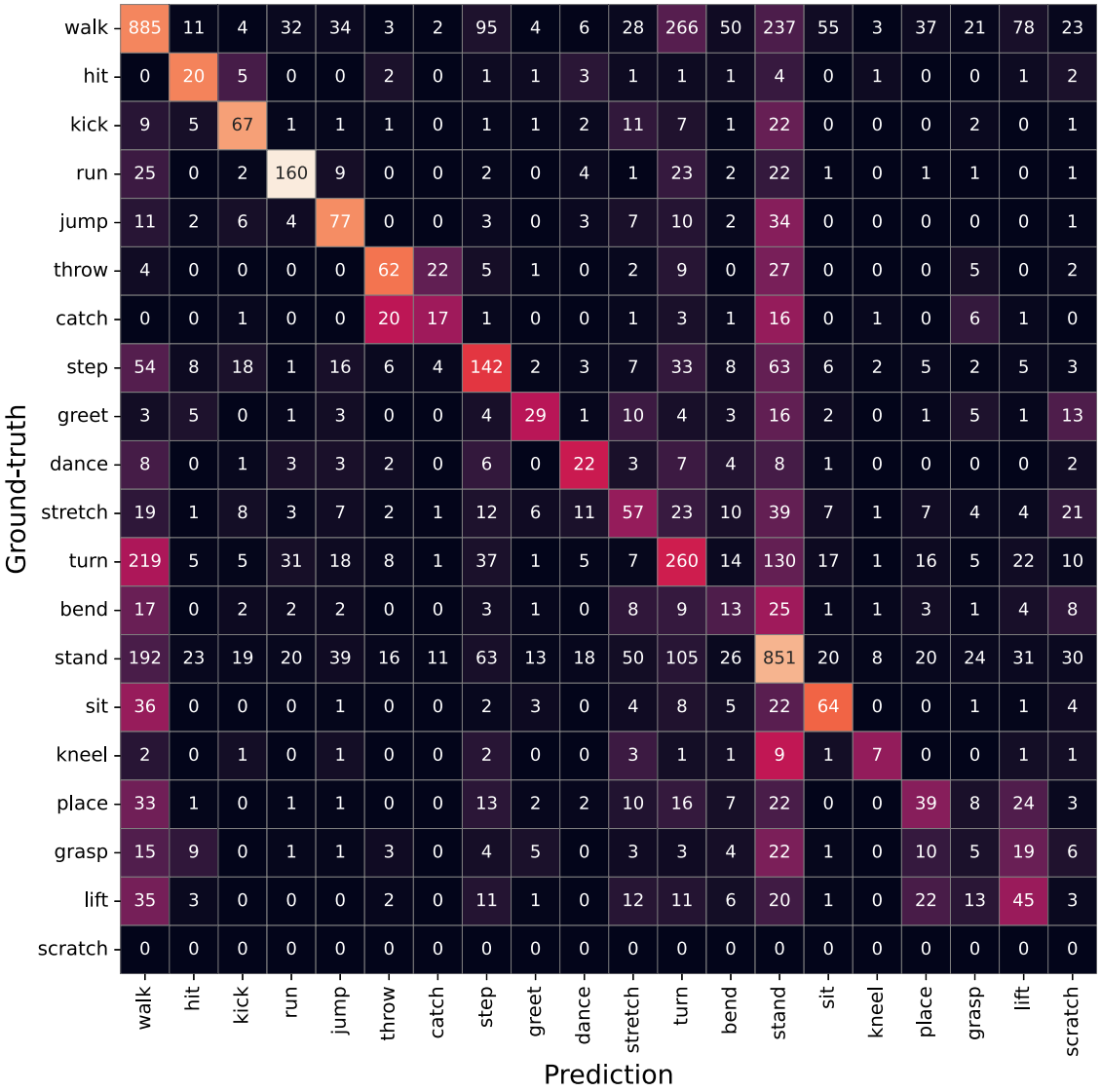
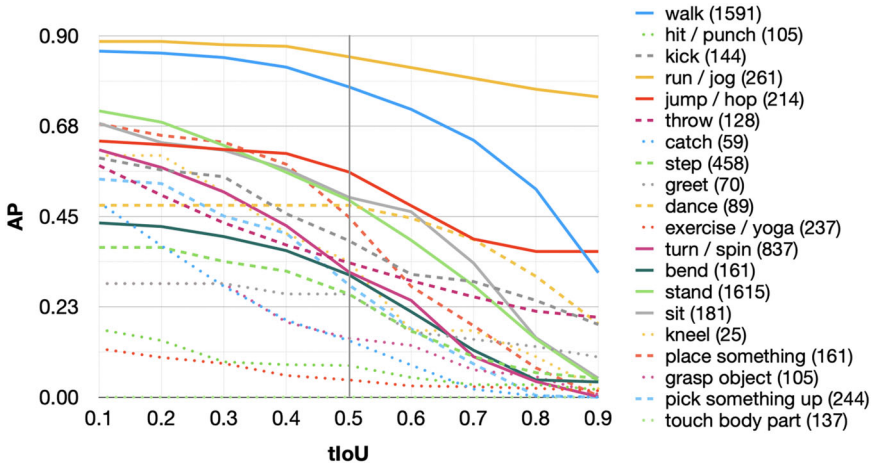


Fig. 2 | Confusion matrix for action recognition on the predicted spans on the BABEL-TAL-20 (BT-20) dataset. A cell contains the number of samples that are wrongly predicted in action classification. The color of a cell along the diagonal represents the precision of a particular class. Notably, we observe that the action "stand" is frequently mistaken for several other actions, highlighting the challenging and diverse nature of our dataset, as well as the precision of our annotations. Additionally, we have noticed that the actions "throw" and "catch" are often misclassified by the model, which can be attributed to their frequent occurrence in sequence, likely leading to localization errors causing misalignment between predicted and ground-truth action spans. Furthermore, despite its lower precision, the action "grasp" is frequently predicted and commonly confused with "lift" because they may share similar motion patterns, features, and temporal context.

Fig. 3 | Average Precision (AP) for each action across varying threshold IoU (tIoU) thresholds on BABEL-TAL-20 (BT-20). The number of samples per class is annotated in parentheses beside actions in the legend. We notice a disparity in the AP scores for each class, both in terms of their absolute values and their sensitivity to changes in tIoU settings. Specifically, we observe that "run" and "walk" achieve the highest AP scores, whereas "touch body part" exhibits the lowest AP scores. This divergence might stem from the former actions being relatively more distinguishable from other actions, while the latter ones are characterized by greater ambiguity and complexity.



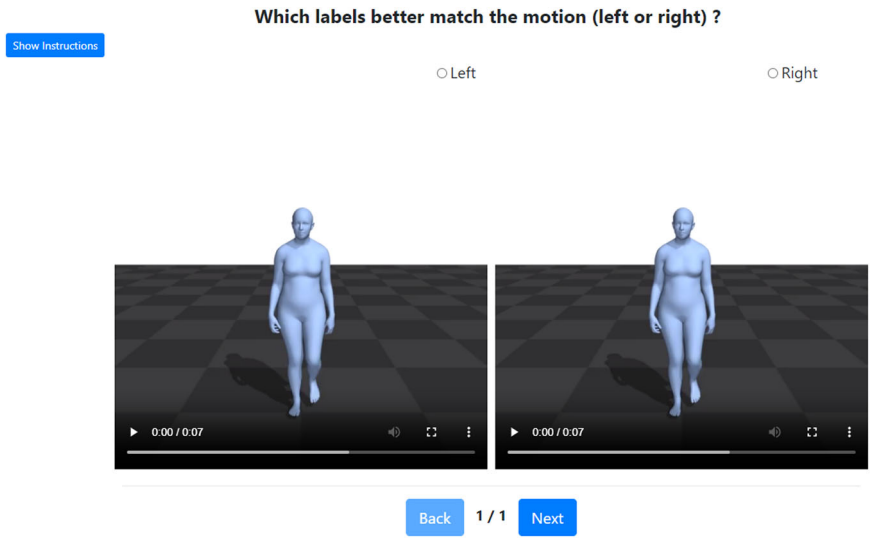


Fig. 4 | The interface of head-to-head human evaluation for two 3D action localization approaches. The evaluator responds to the question: "Which labels better match the motion, left or right?" We conduct a head-to-head comparison between the labels generated by LocATe and those produced by other methods through a human study. It is important to note that the sequence and labels in the provided sample are randomized, and the arrangement of labels from various methods, whether on the left or right, is also randomized across trials.

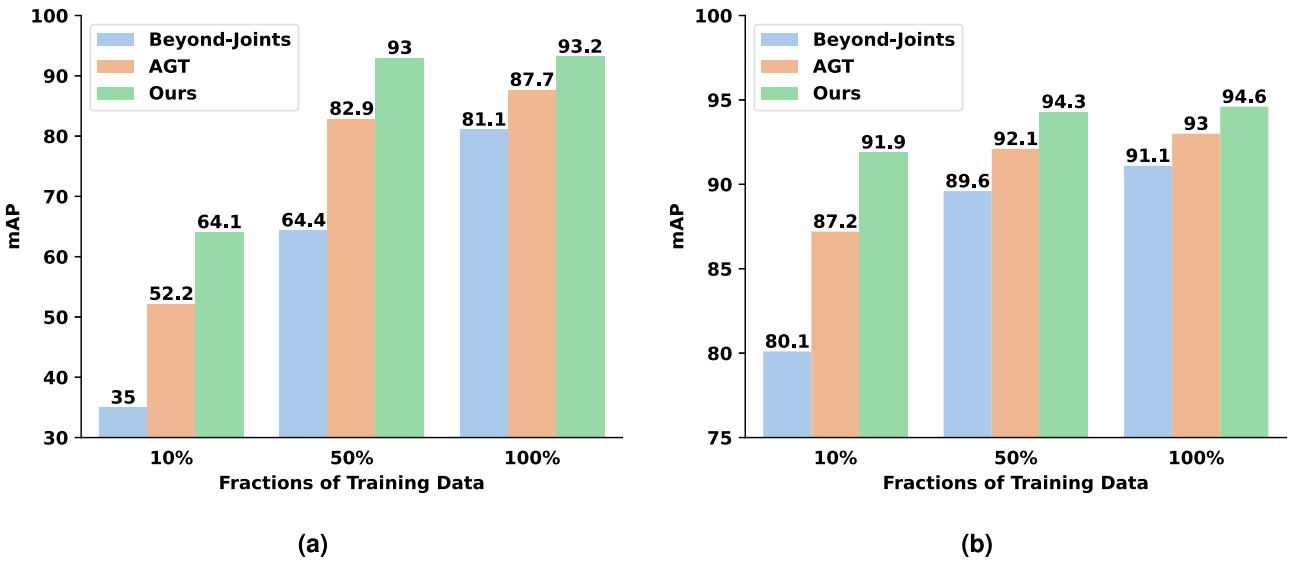


Fig. 5 | Performances using fractions of training data on the PKU-MMD dataset. We observe that the performance of our approach remains robust even when utilizing only half of the training data. Impressively, with just 10% of the training data, our approach maintains a performance level of 91.9% in the relatively straightforward cross-view evaluation. In contrast, the performance of Beyond-Joints deteriorates when working with reduced fractions of training data. This experiment highlights the scalability of our proposed approach, demonstrating its efficacy even when only a limited amount of labeled data is accessible. It also underscores the low intra-class variance among the action categories within the PKU-MMD dataset. Thus, the creation of a new and challenging 3D action localization benchmark becomes imperative to drive future research in this domain. **a** Cross-subject evaluation. **b** Cross-view evaluation. PKU-MMD: Peking University Multi-Modal Dataset.

performance of our approach does not even decrease when using only half of the training data. When using 10% for training, our approach still achieves 91.9% mAP in cross-view evaluation. In contrast, the performance of Beyond-Joints drops sharply when using small fractions of training data. This experiment demonstrates the consistency and scalability of the proposed approach when only a small amount of labeled data is available. It also indicates the low intra-class variance among the action categories in the PKU-MMD dataset, suggesting a new and challenging 3D action localization benchmark is needed to advance future research in this area.

Visualization of Predictions

We provide html files in the Supplementary viz/ folder that visualize results from different approaches on the BT-20 benchmark. Note that these are fully self-contained files that do not contain external links and do not collect any information about the user.

We render videos from the 3D mocap sequences from AMASS[26], and overlay action labels on the videos. The human labels (ground-truth) derived from BABEL[34], are visualized in viz/human_label.html. The predictions from the prior RNN-based approach Beyond-Joints[31] are in viz/beyond_joints.html.

In viz/LocATe_opt.html, we visualize the predictions from our approach where the spans correspond to the optimal match with the human labels. While the raw predictions from LocATe are reasonable, they include many short, overlapping spans that are false positives. To enable use in downstream applications, we apply Non-maximum Suppression (NMS) to the predictions from LocATe. The results from LocATe with NMS are visualized in viz/LocATe_nms.html.

In addition to visualizing results from different methods, we also visualize the different human labels for the same sequence, in viz/human_var.html. We observe that the human-provided labels vary both in terms of the actions that are named and the spans of actions. For instance, some complex activities like 'stretch or yoga' are composed of simpler actions like 'stand', 'turn', or 'bend'. While some human annotators label the simpler actions as well as the overall activity, other annotators only label the latter. This variance in semantic labels motivates us to consider human evaluation as an alternative to automatic metrics like mAP.

Non-maximum Suppression (NMS)

The 'raw' predictions from LocATe consist of many short, and overlapping spans of the same action. This form of output is acceptable when computing performance using automatic metrics (like mAP). However, the overlapping predictions could potentially distract a downstream human user, and severely affect the usability of the system. In the object detection literature, overlapping predictions are handled via NMS. We apply a simple algorithm for the NMS of overlapping spans. NMS eliminates low-confidence predictions for an action that overlaps with a higher confidence span of the same action. Note that the spans themselves are predicted by the regression head, and do not have confidence estimates associated with them. We use the recognition confidence estimate for NMS.

We observe in visualizations (see viz/LocATe_nms.html) that NMS eliminates many short spans, resulting in more coherent predictions. As a consequence, the overall recall of LocATe is lower with NMS than without NMS. In our human studies, we utilize the predictions from LocATe after applying NMS.

Sources of Error

We presented the different loss functions that comprise our overall objective. Temporal Action Localization involves solving two sub-tasks – localization and recognition. Hence errors in the final performance can propagate from either of these tasks.

The first question regarding the overall objective is about the relative weighting of the losses for the two tasks. We present the overall objective as simply a sum of the recognition loss and localization loss. In experiments with an earlier LocATe w/ GAT model, we attempt to train the model with larger weights on the classification loss because we observed some room for improvement in recognition. Specifically, we tried weights in the range [1, 10], but this did not improve performance. However, we note that a more thorough hyper-parameter search could indeed improve the overall performance of LocATe.

Another consequence of using multiple loss functions is that models with different classification and localization losses can achieve similar overall performance as measured by mAP. When comparing Graph Attention (GAT) and Deformable Attention (DA) in LocATe, the relative performance improvement with DA is more evident due to better recognition rather than better temporal localization (refer to Table 4).

While there does exist a correlation between task loss and mAP – larger recognition or localization losses imply lower mAP – at a fine-grained level, this relationship is imprecise. This is to be expected, given the procedure to calculate mAP.

Influence of Input Features

2D Features. As a sanity check, we first ask the following question – given 3D data, does 3D action localization really shows better performance than 2D-TAL on the same data? To answer this question, we consider a baseline model[38] that takes 2D videos as input. We utilize the rendered 2D videos of the mocap sequences in BT-20 as input. The human bodies are animated using the SMPL[41] body model. We extract I3D[42] features for the BT-20 videos, using an I3D model that was pretrained for activity recognition on real videos from the Kinetics[42] dataset. To obtain I3D features corresponding to an input video with T frames, we first divide the video into short overlapping segments of 8 frames with an overlap of 4 frames resulting in T' chunks. In other words, we extract features in a sliding-window fashion, with a filter size = 8 frames and stride = 4 frames. We obtain a tensor of size $T' \times 2048$ as features for these T' chunks. Each video feature sequence is rescaled to 100×2048 (input size of the transformer) using linear interpolation along the temporal dimension.

3D Action Recognition Features. State-of-the-art 2D-TAL methods[38] utilize features from a video recognition backbone (e.g., I3D[42]) as input to the model. Compared to raw pixel values from the videos, these features are lower-dimensional and semantically more meaningful. Although using the representation from a feature extractor increases the computation and memory requirements compared to using the 'raw' joint positions as input, we attempt to determine if action recognition (AR) features to improve performance in 3D action localization. First, we train a popular action recognition model, 2S-AGCN[43], to classify the 20 actions in the BT-20 dataset. To maximize the discriminativeness of the input feature, we train the recognition model with 8 frames – the size of a single input snippet to the LocATe transformer encoder. The model achieves a top-1 accuracy of 63.41% and a top-5 accuracy of 85.83%, demonstrating that it successfully captures some semantic information, despite being far from perfect. We then experiment with features extracted from two different layers of the 2S-AGCN model – after the second graph-convolution layer (Early-layer AR feat.) and after the last graph-convolution layer (Later-layer AR feat.).

Results. Table 5 shows the results from our experiments with different input representations. Unsurprisingly, the 2D features extracted from the rendered videos, underperform the 3D joint features. This is because the 3D representation contains more information than 2D, which demonstrates the superiority of 3D representation. We observe that the 3D action recognition features do not improve performance compared to joint position information. This implies that 3D input feature representation is still an open problem to explore. Note that we performed the experiments with LocATe w/ GAT, i.e., LocATe with sparse Graph Attention, which is an earlier model. Since we observed the best performance with joint positions as input, we employed the same in our final model LocATe.

Table 4 | We investigate the source of improvement with Deformable Attention in LocATe, compared to Graph Attention (GAT)

Method	Rec. loss ↓	Loc. loss ↓	mAP ↑
LocATe w/ GAT	0.671	0.395	23.4
LocATe	0.362	0.374	36.0

We break down the sources of error into two components: recognition loss (column 2) and localization loss (column 3). It is essential to recognize that lower loss values indicate superior performance. We gauge performance using the metric mAP@tIoU = 0.5, where higher values indicate better performance. *mAP* mean Average Precision, *tIoU* threshold IoU, *Rec* recognition, *Loc* localization.

Table 5 | Ablation Study

Method	mAP @ 0.5 tIoU
2D Features[38]	14.5
Joint pos.	23.4
Early-layer AR feat.	21.3
Later-layer AR feat.	20.4
Joint pos. + Early + Later AR feat.	21.4

Effect of different 3D Human Representations on BABEL-TAL-20 (BT-20). The 2D features obtained from the rendered videos exhibit inferior performance compared to the 3D joint features. This discrepancy arises from the 3D representation's ability to encapsulate a richer information set compared to the 2D representation, underscoring the superiority of the 3D approach. *mAP* mean Average Precision, *tIoU* threshold IoU.

Discussion

3D action localization is a pivotal computer vision task with diverse practical applications. In this study, we have explored various aspects of 3D action localization, from dataset challenges and model performance to the impact of different evaluation metrics.

Dataset Challenges and Diversity

Our investigation into existing datasets, such as PKU-MMD, revealed several challenges in 3D action localization. Notably, we observed a variance in class distribution, with certain actions being more easily distinguishable than others. Actions like "run" and "walk" consistently achieved high AP scores, while more complex actions like "touch body part" posed challenges. Actions like "touch body part" are extremely difficult for LocATe, which demonstrates poor performance across all tIoU thresholds. Activities like "dance" typically exhibit larger intra-class variance than other "simpler" actions like "run". Learning to accurately recognize and localize actions that have limited data and large variance, is an open challenge. Poor performance in classes like "touch body part" suggests that our current data representation – joint positions in a skeleton – lacks the expressiveness to capture this subtle action. Improving performance on understanding this action in 3D, is an interesting future direction, as social touch is an important component of human interactions[44,45]. These findings emphasize the importance of fine-grained and accurate annotations, as well as the need for challenging and diverse datasets that better reflect real-world scenarios.

Model Performance

Our study involved the evaluation of different approaches for 3D action localization. Notably, the comparison between our proposed LocATe approach and existing techniques yielded promising results. LocATe exhibited state-of-the-art performance, with a particularly remarkable improvement of nearly 10% in cross-subject evaluation compared to the previous method Beyond-Joints. These results highlight the potential for advancements in this field, as well as the importance of robust algorithms and models.

Data Scaling and Low Annotation Availability

An intriguing aspect of our investigation was the scalability of our proposed approach. We demonstrated that our method remained robust even when trained with a reduced amount of labeled data, showing its adaptability and effectiveness under resource constraints. This finding is of particular importance in scenarios where obtaining large labeled datasets is challenging. It also underscores the need for a new and challenging benchmark dataset to foster future research in 3D action localization.

The Superiority of 3D Representation

One of the key takeaways from this study is the advantage of 3D representation over 2D features. We found that 3D joint features consistently outperformed 2D features, underscoring the value of capturing three-dimensional spatial information for accurate action localization. This highlights the need for more advanced techniques in leveraging 3D data for this task.

In conclusion, in light of the saturating performance of existing 3D human action recognition and localization methods on simple benchmarks, we introduce BABEL-TAL, an unconstrained and substantially more challenging, complex benchmark, to further the research in this field. We also present LocATe, a Transformer-based single-stage method that learns to jointly perform localization and recognition. Representative strong baselines are examined on BT-20 and LocATe outperforms all of them. Further analyses of the confusion matrix and performances on different actions indicate that there is room for improvement for 3D action localization methods on our challenging BT-20 benchmark. One limitation of our work is that we have not yet explored the integration of RGB and motion-captured data for motion analysis. In future research, we plan to investigate the potential benefits of leveraging these multiple modalities together. We believe that the dataset, method, and findings in this work will be beneficial to the community and our contributions will have practical applications in fields such as human-computer interaction, animation, AR/VR systems, and LLM-integrated 3D human behavior understanding. Studying 3D action localization is not just an academic pursuit but has a direct social impact. For example, augmenting healthcare and wellness. With the ability to precisely track and understand human motion, our research can contribute to advancements in telerehabilitation, monitoring elderly populations, and providing early intervention in medical conditions. The scalability of our proposed approach, even with limited data, holds promise in regions with limited healthcare resources, improving access to healthcare and enhancing the quality of life. For individuals with disabilities, 3D action localization can be useful. The ability to understand and interpret gestures and movements provides the foundation for assistive technologies that empower those with limited mobility or communication abilities. Whether it is controlling a wheelchair, operating household appliances, or communicating with others, these technologies can improve the independence and quality of life of people with disabilities. By advancing the accuracy, scalability, and understanding of human actions, it offers us with new opportunities for human-machine interaction and healthcare enhancement. As researchers in this field, our responsibility is not only to achieve technological milestones but to channel our discoveries toward the broader benefits of human life and society. It is through this lens that we advocate for continued innovation and research in 3D action localization.

Methods

Related Work

3D Action Localization Datasets. Over the years, datasets have driven progress in 3D action localization. G3D[46], one of the earliest datasets of this area, aims at understanding the gestures of a person in real-time for video-game applications. CAD-120[47] contains daily activities performed in different environments, but is composed of simpler actions. Watch-n-Patch[48] is a dataset of daily human activities, and can be used for the application called "action patching". Action patching primarily aims for unsupervised action segmentation and recognition, and also serves to detect forgotten actions during long-term activities. Similarly, Lillo et al.[49] focus on modeling spatial and temporal compositions of simple actions to detect complex activities. Unlike these datasets, SBU Kinetic interaction[50] contains eight two-person interactions such as "approaching", "hugging", etc. PKU-MMD[32] consists of 51 simple actions such as "putting on glasses", "shaking hands" specifically performed by actors. Wei et al.[51] curate an expanded 3D dataset for concurrent activity detection, encompassing skeletal sequences and corresponding RGB-D videos that encompass a broader spectrum of simultaneous activities. In contrast, the recently introduced BABEL dataset[34] is constructed by amalgamating motion capture (mocap) sequences from the extensive AMASS data archive[26], which encompasses a diverse compilation of mocap datasets. As opposed to previous datasets, BABEL-TAL exhibits a higher degree of complexity and features a long-tailed distribution of actions, and hence it is more challenging for 3D-TAL. We emphasize the significance of developing and utilizing relatively challenging datasets to foster advancements in the field of 3D-TAL. Such datasets serve as indispensable catalysts for pushing the boundaries of research and the invention of more robust and effective algorithms to tackle real-world scenarios.

Transformers for 3D Action Recognition. Transformers[52] have recently been extended for 3D skeleton-based human action recognition. Shi et al.[53] present a decoupled spatial-temporal attention network based on the self-attention mechanism and combine spatial transformer with temporal convolution. Stacked Relation Networks (SRN)[51] use a specialized relation network for decompositional design to enhance the expressiveness of instance-wise representations via inter-instance relationship modeling. Sequential Correlation Network (SCN)[54] combines a recurrent neural network and a correlation model hierarchically to model the complex correlations and temporal dynamics of concurrent activities.

Plizzari et al.[55] propose a two-stream Transformer-based model by employing self-attention on both the spatial and temporal dimensions. Zhang et al.[56] introduce a spatial-temporal specialized transformer to model the skeleton sequence in spatial and temporal dimensions, respectively. Recently, Pang et al.[57] design a Transformer-based model for skeleton-based human interaction recognition by learning the relationships of interactive persons from both semantic and distance levels. There are also some Transformer-based approaches for unsupervised or self-supervised skeleton-based action recognition. Chen et al.[58] design a pre-training scheme to train a hierarchical Transformer-based encoder for skeleton sequences. Kim et al.[59] propose a transformer architecture for unsupervised skeleton-based action representation learning with global and local attention mechanisms to model joint dynamics and temporal contexts, respectively.

BABEL-TAL

Compared with 3D action recognition which relies on trimmed videos, 3D action localization from the original untrimmed video aligns much better with realistic scenarios. However, the research in this area is still at an early stage, partly due to the shortage of appropriate and large-scale benchmarks. Reviewing existing action and pose-related datasets, we find these datasets are either inapplicable for 3D-TAL or need further improvement to meet the demands of real-world applications. For instance, Human3.6M[60] is suitable for human pose estimation, but cannot be used for temporal action localization. Besides, it is recorded in constrained environments. Likewise, PKU-MMD[32] is collected in indoor environments with limited subjects, views, and classes. NTU RGB+D[61] is a much larger dataset for human action recognition, but only contains a single action for a video and is also captured in constrained environments. As a result, neither action sequences nor categories are varied enough to meet the real demand. There are also some datasets that are captured in the wild but can only be used for 2D action localization, such as THUMOS[62], MultiTHUMOS[63], and ActivityNet[64].

To facilitate the research progress, we aim to introduce a new large-scale dataset for 3D action localization with a wide variety of action categories and high intra-class diversities. Here, we base on the recently released BABEL dataset[34] to carefully construct a benchmark BABEL-TAL (BT) for 3D action localization, which includes three sets, BABAL-TAL-ALL, BABEL-TAL-60, and BABAL-TAL-20. We will introduce BT from the following three aspects: data processing, labels, and data distributions.

Data Processing. The mocap sequences in BABEL-TAL are derived from AMASS dataset[65], which are generated by mocap actors and collected through their performances. To create a robust temporal localization dataset, we first filter those sequences of extremely short duration, which often lack sufficient temporal context. We use the modified VIA annotation software[66] provided by BABEL to verify and refine the frame annotation of the motion sequences to maintain accurate temporal references. As a result, after careful data processing and refinement, we obtain the BABEL-TAL-ALL.

Labels. BABEL-TAL exhibits diversity in the types of actions. After data processing, BABEL-TAL includes 102 diverse action classes, which contain both commonplace activities such as "walk", "jump" and specialized activities such as "yoga" and "cartwheel". The variability in action types across datasets within BABEL-TAL permits a broad spectrum of action diversity, catering to various domains and levels of proficiency, but also poses a formidable challenge to methods' robustness and generalization. To provide a more realistic training and evaluation environment for 3D-TAL, we select the most frequent 60 actions in BABEL to build a subset of BABEL-TAL-ALL, called BABEL-TAL-60 (BT-60).

Nevertheless, according to our experiments, BT-60 is still challenging for existing 3D-TAL methods, leading to only 3.0% mAP for a recent method of AGT[38]. We find that there are a few fine-grained and semantically similar categories among the most frequent actions. For instance, "run" and "jog", which depict the same action, belong to different categories in the BABEL. The detailed category merge process is presented in Table 6. By organizing the raw labels into sets of higher-level labels, BT-20 provides researchers with a more manageable and meaningful label set that strikes a balance between specificity and practicality. This organization enhances the feasibility of action localization tasks and enhances the usability of the dataset. The resulting dataset BABEL-TAL-20 (BT-20) contains a total of 5727 sequences with 20 classes.

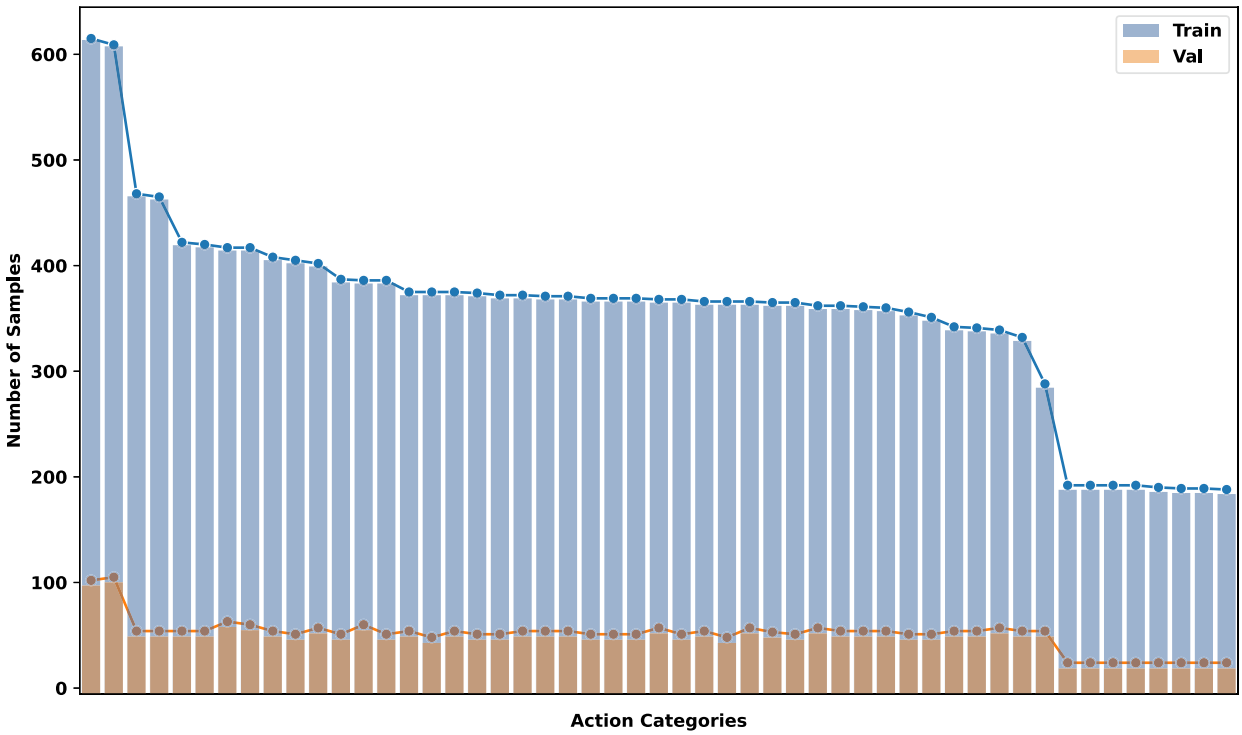
Table 6 | BABEL-TAL-20 (BT-20) has 20 actions (column 1), where each is a superset of BABEL actions (column 2)

BT-20	BABEL	# Train	# Validation
Turn	Turn, spin	2044	837
Run	Run, jog	750	261
Lift something	Take/pick something up, lift something	709	244
Jump	Jump, hop, leap	700	214
Stretch (exercise/yoga)	Stretch, yoga, exercise/ training	601	237
Scratch (touch body part)	Scratch, touching face, touching body parts	310	137
Hit	Hit, punch	206	105
Walk	Walk	4671	1591
Kick	Kick	347	144
Throw	Throw	460	128
Catch	Catch	193	59
Step	Step	1097	458
Greet	Greet	179	70
Dance	Dance	189	89
Bend	Bend	468	161
Stand	Stand	4193	1615
Sit	Sit	512	181
Kneel	Kneel	102	25
Place something	Place something	510	161
Grasp object	Grasp object	247	105

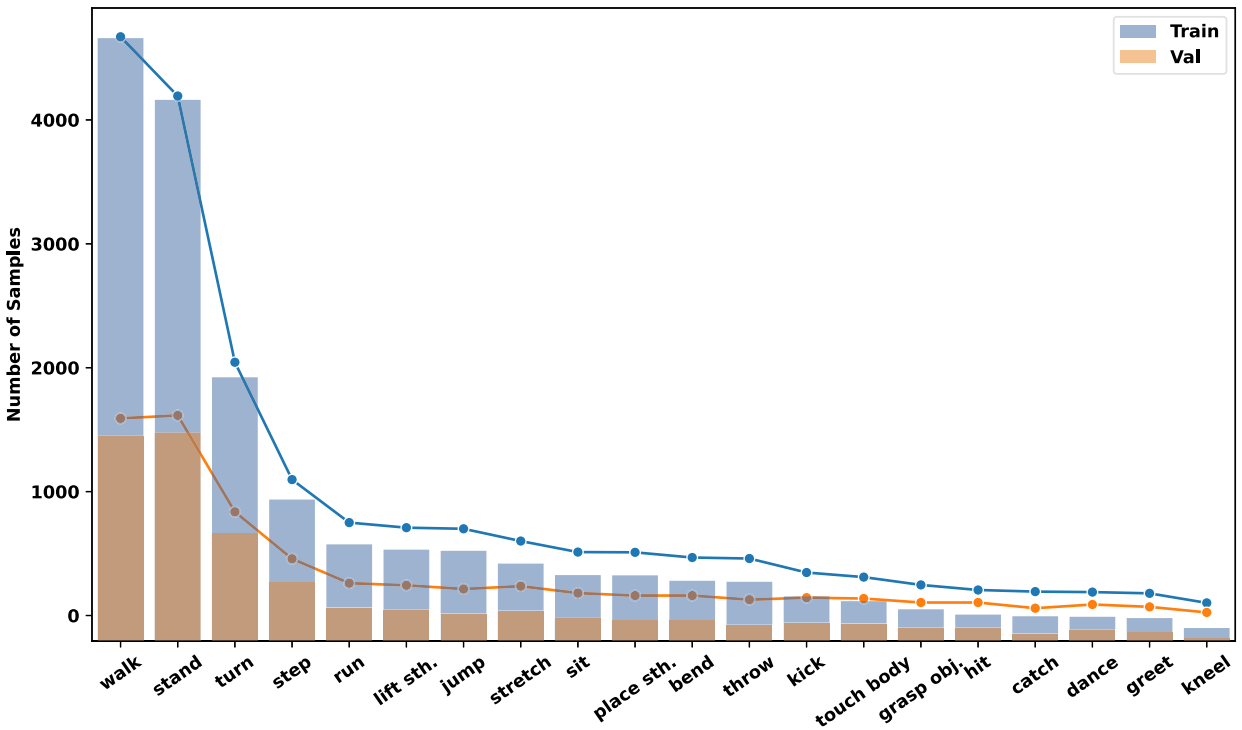
The number of samples in the training and validation set of BT-20 is in columns 3 and 4. The distribution of action categories in BT-20 samples closely mirrors real-world scenarios with a pronounced long-tailed pattern.

Data Distributions. Samples of action categories in BABEL-TAL follow a pronounced long-tailed distribution to closely resemble real-world scenarios. As illustrated in Table 6, for BT-20, the sample count for each class ranges from 102 to 4671 for the training set and from 25 to 1615 for the validation set. Furthermore, as shown in Fig. 6, compared with the existing dataset PKU-MMD[32], both training and validation samples in this new dataset exhibit more evident long-tailed class distribution, which fits visual phenomena in real-world applications.

To ensure ample availability of data for learning methods, we have developed the BABEL-TAL-20 (BT-20) Benchmark and expanded versions, namely BABEL-TAL-60 (BT-60) Benchmark and BABEL-TAL-ALL (BT-ALL) Benchmark. These benchmarks provide extensive datasets to facilitate the training and evaluation of various learning algorithms in the field of 3D action localization. By offering a diverse range of labeled sequences, these benchmarks aim to support the development of robust and effective techniques for understanding and analyzing human behavior. In the main part, we detailedly introduce the BABEL-TAL-20, in the following, we will introduce BABEL-TAL-ALL and BABEL-TAL-60.



(a)



(b)

Fig. 6 | Class frequency distributions of typical 3D action localization datasets. In contrast to the pre-existing dataset PKU-MMD[32], our new dataset BABEL-TAL-20 (BT-20) demonstrates a more pronounced long-tailed class distribution in both training and validation samples, aligning better with visual patterns encountered in real-world applications. **a** Class frequency distributions of PKU-MMD. **b** Class frequency distributions of the introduced BT-20. PKU-MMD: Peking University Multi-Modal Dataset.

BABEL-TAL-ALL (BT-ALL). After data processing on BABEL, we build a large and challenging dataset, namely BABEL-TAL-ALL (BT-ALL), for 3D action localization. In Fig. 1c, we plot the sorted distribution of samples per category in BT-ALL, the corresponding BABEL categories that constitute the action, and the number of samples in the training and validation set of BT-ALL. The comparison of existing 3D action localization datasets is shown in Table 1. The results demonstrate the considerable challenge posed by these benchmarks, primarily due to the inclusion of a larger number of action categories. The expanded set of action categories requires models to possess enhanced generalization capabilities and improved discrimination among diverse actions. The increased difficulty of these benchmarks encourages the development of more robust and sophisticated algorithms to address the complexities of 3D action localization tasks.

BABEL-TAL-60 (BT-60). We also create the BABEL-TAL-60 (BT-60) benchmark for 3D action localization with the 60-class action recognition subset of BABEL[34]. The action recognition task involves predicting the action class of the given "trimmed" span of movement. The movement spans in the BABEL-60 subset belong to the most frequent 60 actions in BABEL[34]. Given the full mocap sequence, the task is to localize and recognize all of the 60 actions. This is extremely challenging, especially due to many infrequent actions. Due to the expensive nature of (dense) annotation required for TAL tasks, we believe that learning efficiently from limited amounts of fully annotated data is an important problem for the community to solve in the long run. In Fig. 7, we plot the sorted distribution of samples per category in BT-60, the corresponding BABEL categories that constitute the action.

In summary, we propose a new benchmark for 3D-TAL, namely BABEL-TAL, including three sets: BT-20, BT-60, and BT-ALL. The BT dataset differs from existing 3D action localization datasets in at least four aspects. First, it is the first 3D motion-capture dataset with precise body joint movements for temporal action localization. Second, this dataset enjoys a wide variety of action labels and exhibits high intra-class diversity. Third, the action data follows a long-tailed distribution. Finally, continuous actions in the long motion sequence are unconstrained by environments or actors. Table 1 shows the comparison between previous datasets with our dataset.

LocATe

We propose a strong baseline model for temporal action localization, namely LocATe. 3D action localization involves predicting the action label, start time, and end time of every action occurring in a 3D motion sequence. Specifically, given a 3D motion sequence $\{\mathbf{x}_t\}_{t=1}^{T}$, the goal is to predict a set of N action spans $\Psi = \{(c^n, t_s^n, t_e^n)\}_{n=1}^{N}$, where c^n is the action category of the n-th span and (t_s^n, t_e^n) denote the start and end times of the n-th span. For simplicity, the 3D joint skeletons are used as the input. Figure 1b shows the overall architecture of LocATe. Given a sequence of human poses, LocATe functions as follows: First, a skeleton-based sampling strategy is used to convert the raw 3D skeleton sequence to a fixed set of joint snippets. Then, a projection of the 3D pose features is summed with positional (time) information and input to the transformer. Next, the transformer encoder models the global context across all temporal positions to produce the feature $\boldsymbol{h}_{L_e}$ that encodes action span information in the sequence. After that, the decoder transforms a fixed set of action queries into action query representations $\boldsymbol{y}_{L_d}$ based on the encoding $\boldsymbol{h}_{L_e}$. Finally, the prediction head outputs the temporal localization results.

Our Transformer-based architecture for action localization contains three main components: Transformer Encoder, Transformer Decoder, and Prediction Heads.

Deformable Attention. As shown in Fig. 1b, we utilize deformable attention (DA)[40] in our encoder and decoder. Given a feature $\boldsymbol{x}$ of a certain transformer layer at a particular position, DA aims to attend to a small set of relevant elements from $\boldsymbol{x}$. It learns K ($K \ll T$) relevant sampling locations from $\boldsymbol{x}$ based on a reference location.

Given a feature $\boldsymbol{x} \in \mathbb{R}^{C \times T}$ at a certain transformer layer and the feature at a particular position $\boldsymbol{z}_q \in \mathbb{R}^{C}$, DA aims to attend to a small set of relevant elements from $\boldsymbol{x}$. It identifies K relevant sampling locations from $\boldsymbol{x}$ based on a reference location p_q (the q-th position of the query feature $\boldsymbol{z}_q$), and $K \ll T$. The sampling locations are predicted relative to the reference

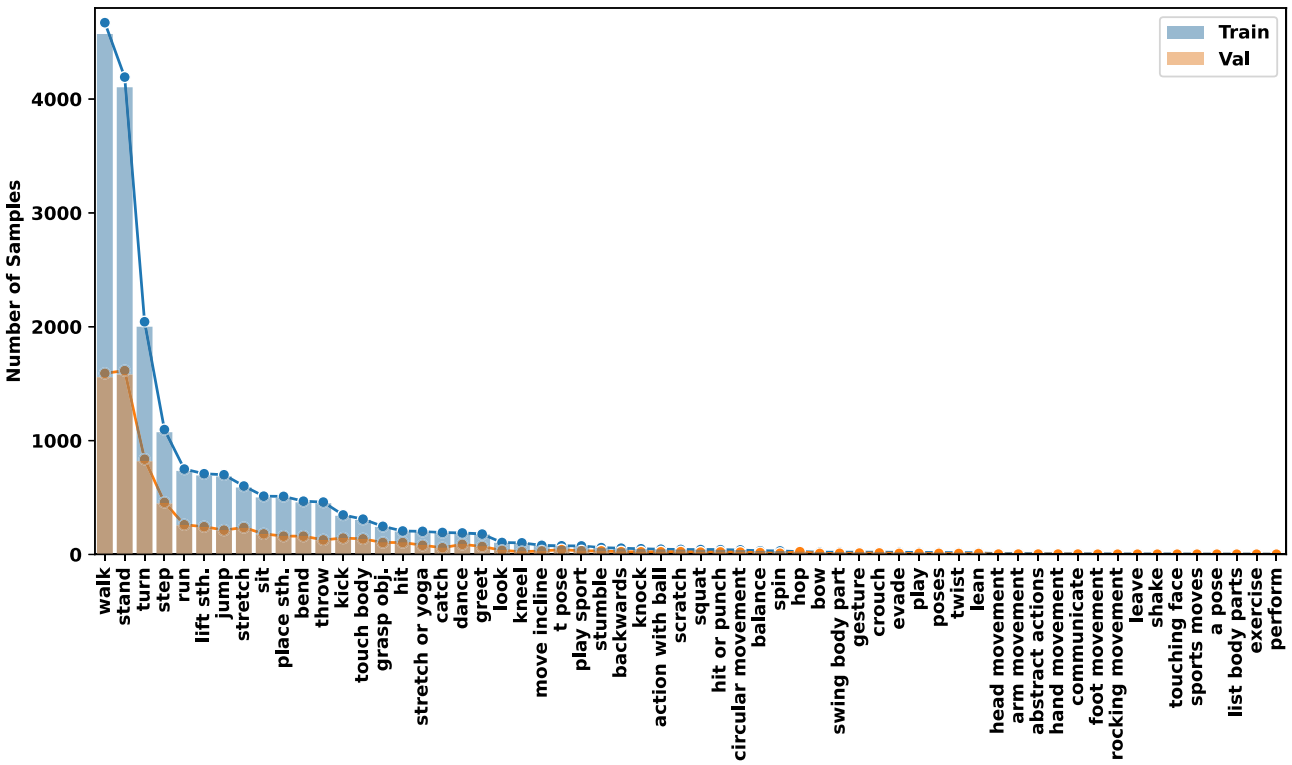


Fig. 7 | Class frequency distributions of the introduced BABEL-TAL-60 (BT-60).

location p_q, and denoted as Δp_{qk}, where k indexes the sampled keys. The model learns parameters W^V to project the features of the sampled keys $\boldsymbol{x}(p_q + \Delta p_{qk})$. Mathematically, DA for one attention head is expressed as

$$\mathrm{DA}(\boldsymbol{z}_q, p_q, \boldsymbol{x}) = \sum_{k=1}^{K} A_{qk} \cdot \boldsymbol{W}^V \boldsymbol{x}(p_q + \Delta p_{qk}), \tag{1}$$

where, A_{qk} denotes the scalar attention weight of the k-th sampled key, A_{qk} is in the range [0, 1] and normalized over K to sum to 1, and Δp_{qk} is a real valued number. Both A_{qk} and Δp_{qk} are obtained by a linear projection over $\boldsymbol{z}_q$. A bilinear interpolation is performed to compute $\boldsymbol{x}(p_q + \Delta p_{qk})$. For more details, please refer to the DA for object detection[40]. As presented in Eq. (1), the overall feature with different attention heads is a weighted sum over the DA representation.

Transformer Encoder. 3D human joints at each time step are represented by a joint vector x_t. A joint embedding which consists of a linear projection with the parameter of W is used to embed x_t to a feature space of size C. Each position (time-step) t is associated with a positional encoding $\boldsymbol{p}_t \in \mathbb{R}^C$, where $t \in \{1, \ldots, T\}$, obtaining a new sequence of joint representation $\{\tilde{x}_t\}_{t=1}^{T}$ which can be denoted as a matrix $\tilde{\mathbf{X}} \in \mathbb{R}^{C \times T}$, where $\tilde{\boldsymbol{x}}_t = W \cdot x_t + \boldsymbol{p}_t$.

The transformer encoder consists of L_e identical encoder layers. A single encoder layer E_ℓ involves the following operations: (1) Self-attention computation, (2) Concatenation of features from different heads, (3) Projection back into the original feature dimension, (4) Residual connection[67], (5) Layer Normalization[68]. Overall, the encoder feature $h_{L_e} = E_{L_e} \circ \ldots \circ E_1(\tilde{X})$ is the resulting output from all the layers.

The self-attention operation effectively models context across features in all T temporal positions. DA is properly used for both accuracy and efficiency. Each transformer head projects features from the previous position into a different feature space, capturing different signals among different scale levels. In an aggregation step, features from all heads are concatenated and projected back into the original feature dimension, followed by residual connection. Overall, the encoder feature $h_{L_e} = E_{L_e} \circ \ldots \circ E_1(\tilde{X})$ is the resulting output from all the layers.

Transformer Decoder. Inspired by previous works[36,40] that use object queries as input to the decoder for object detection, we introduce action queries $Q \in \mathbb{R}^{C \times N_a}$ as input to the decoder for action localization. The transformer decoder contains L_d decoder layers, and each decoder layer D_ℓ performs two attention computations: self-attention, and cross-attention. The decoder transforms action queries into action query representations y_{L_d}, which are then fed into prediction heads to obtain final localization results.

Prediction Heads. The Prediction head comprises two network branches: regression and recognition. The regression network consists of a 3-layer fully connected network (FC) with ReLU activation that predicts the start time $\hat{t}_s$ and end time $\hat{t}_e$ of the action. The recognition network is a single fully connected with a softmax function that scores the set of actions (including a "no action" class).

Loss Functions

Action localization is trained with two objectives: action classification and temporal boundary regression. Similar to Carion et al.[36], bipartite matching is used to match the ground truth and predictions. For classification, we utilize a class-balanced focal loss to address the issue of imbalance between action classes.

We use bipartite matching to match the ground-truth Ψ and predictions $\hat{\Psi}$. Since $|\Psi| < \hat{\Psi}, \hat{\Psi} = N_a$, we augment the ground-truth Ψ with "no action" spans such that the augmented ground-truth $\tilde{\Psi}_i$ has N_a spans. We then compute an optimal match between $\tilde{\Psi}_i$ and $\hat{\Psi}_i$. The optimal permutation σ^* among the set of all permutations Σ_{N_a} defined as

$$\sigma^* = \mathrm{argmin}_{\sigma \in \Sigma_{N_a}} \sum_{n=1}^{N_a} \mathscr{L}(\tilde{\Psi}^n, \hat{\Psi}^{\sigma(n)}). \tag{2}$$

The optimal permutation can be efficiently computed via the Hungarian algorithm. The formula of loss functions is described as follows.

Regression Loss. The regression loss $\mathscr{L}_\mathrm{r}$ measures the localization similarity between the predicted and ground-truth action spans and is a weighted combination of two terms. For the n-th pair in the matched permutations σ, the formulation is

$$\mathscr{L}_\mathrm{r} = \lambda_{iou} \mathscr{L}_{iou}(s^n, \hat{s}^{\sigma(n)}) + \lambda_{L_1} ||s^n - \hat{s}^{\sigma(n)}||_1, \tag{3}$$

where s^n is the start and end times $[t_s^n, t_e^n]$ of the n-th span, and λ_{iou} and λ_{L_1} are scalar hyperparameters.

Classification Loss. In an attempt to effectively model the heavy class imbalance, we exploit class-balanced focal loss $\mathscr{L}_c$[69] for action localization. Different from focal loss[70], the class-balanced focal loss incorporates a class-weighting term, which is a non-linear function of the class frequency. Given predicted class scores $\boldsymbol{z}$, we define $\tilde{z}^j$ as

$$\tilde{z}^j = \begin{cases} z^j, & \text{if } j = c^n \\ -z^j, & \text{otherwise} \end{cases} \tag{4}$$

Under the match permutation $\hat{p}_{\sigma(n)}(c^n)$, the loss $\mathscr{L}_c$ is computed between the ground truth-class c^n and the predicted class scores,

$$\mathscr{L}_c(c^n, \hat{p}_{\sigma(n)}(c^n)) = -\frac{1-\beta}{1-\beta^{f(c^n)}} \sum_{j=1}^{C} \left(1 - \hat{p}^j_{\sigma(n)}(c^n)\right)^\gamma \log\left(\hat{p}^j_{\sigma(n)}(c^n)\right), \tag{5}$$

where $\hat{p}^j = \mathrm{sigmoid}(\tilde{z}^j)$, $f(c^n)$ is the frequency of the ground-truth class c^n, $\beta \in [0, 1)$ and γ are scalar hyperparameters, and C is the total number of classes.

For class-balanced action localization, focal loss up-weights the cross-entropy loss for inaccurate predictions, resulting in a larger training signal for difficult samples.

Overall Objective. The bipartite matching loss $\mathscr{L}$ is a sum of the classification and regression losses:

$$\mathscr{L}(\tilde{\Psi}^n, \hat{\Psi}^{\sigma(n)}) = \mathscr{L}_c(c^n, \hat{p}_{\sigma(n)}(c^n)) + \mathscr{L}_\mathrm{r}(s^n, \hat{s}^{\sigma(n)}). \tag{6}$$

After obtaining the optimal permutation σ^*, the overall objective loss function $\mathscr{L}_F$ over all the matched pairs of action spans are defined as:

$$\mathscr{L}_F = \sum_{n=1}^{N_a} \mathscr{L}(\psi^n, \hat{\psi}^{\sigma^*(n)}). \tag{7}$$

Architecture and Hyper-parameters of LocATe

We implement LocATe using PyTorch 1.4, Python 3.7, and CUDA 10.2. There are 4 transformer encoder heads, and 4 decoder heads. For class-balanced focal loss, we set $\beta = 0.99$ and $\gamma = 2$. For LocATe, both the transformer encoder and decoder consist of four layers, i.e., $L_e = 4$, $L_d = 4$. The deformable attention has four heads in parallel. The default sequence length T before sending to the transformers is 100, and the feature dimension C is 256. The networks are trained with the Adam optimizer[71], and the learning rate is $4e^{-3}$.

Baselines

Here, we present baseline methods for comparison. To have a thorough evaluation, we revise some public implementations of RGB-based temporal action localization methods and slightly modify their input block to accommodate 3D skeletons as input. The baseline methods are listed as follows: (a) Beyond-Joints[31] is one of the state-of-the-state 3D action localization methods. It mainly consists of an RNN-based per-frame action classifier. We benchmark the performance with publicly available implementation. (b) SRN[51] uses a specialized relation network for decompositional design to enhance the expressiveness of instance-wise representations via inter-instance relationship modeling. (c) ASFD[72] is a purely anchor-free RGB temporal action localization method. We re-implement ASFD and adapt it to 3D skeleton input. (d) TSP[73] is a temporally sensitive supervised pretraining method for RGB video that considers global information to improve temporal sensitivity. We modified the original implementation to fit for 3D skeleton input. (e) G-TAD[37] is an RGB temporal action localization approach whose key idea is to effectively model the context around a short span of video. (f) AGT[38] is a Transformers-based RGB temporal action localization method that employs a Graph Attention (GAT) mechanism[74].

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The datasets are available for download from our project GitHub[75].

Code availability

Code is available on our project GitHub[75].

Received: 24 January 2024; Accepted: 23 August 2024;
Published online: 03 September 2024

References

1. Lee, I., Kim, D. & Lee, S. 3-d human behavior understanding using generalized ts-lstm networks. *IEEE Trans. Multimed.* **23**, 415–428 (2020).
2. Devanne, M.3d human behavior understanding by shape analysis of human motion and pose. Ph.D. thesis, Université Lille 1-Sciences et Technologies (2015).
3. Ortega, B. P. & Olmedo, J. M. J. Application of motion capture technology for sport performance analysis. *Retos: nuevas tendencias en educacion fisica, deporte y recreacion* 241–247 (2017).
4. Kanazawa, A., Black, M. J., Jacobs, D. W. & Malik, J. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)* (2018).
5. Huang, Y. et al. Towards accurate marker-less human shape and pose estimation over time. In *2017 international conference on 3D vision (3DV)*, 421–430 (IEEE, 2017).
6. Bogo, F. et al. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science (Springer International Publishing, 2016).
7. Jain, M., van Gemert, J., Jegou, H., Bouthemy, P. & Snoek, C. G. Action localization with tubelets from motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2014).
8. Nigam, S., Singh, R. & Misra, A. A review of computational approaches for human behavior detection. *Arch. Computational Methods Eng.* **26**, 831–863 (2019).
9. Pareek, G., Nigam, S. & Singh, R. Modeling transformer architecture with attention layer for human activity recognition. *Neural Computing and Applications* 1–14 (2024).
10. Slama, R., Wannous, H., Daoudi, M. & Srivastava, A. Accurate 3d action recognition using learning on the grassmann manifold. *Pattern Recognit.* **48**, 556–567 (2015).
11. Bhoi, A. Spatio-temporal action recognition: A survey. *arXiv preprint arXiv:1901.09403* (2019).
12. Aggarwal, J. K. & Xia, L. Human activity recognition from 3d data: a review. *Pattern Recognit. Lett.* **48**, 70–80 (2014).
13. Choi, J., Gao, C., Messou, J. C. & Huang, J.-B. Why can't I dance in the mall? Learning to mitigate scene bias in action recognition. *Adv. Neur. Inf. Process. Syst.* **32** (2019).
14. Moeslund, T. B., Hilton, A. & Krüger, V. A survey of advances in vision-based human motion capture and analysis. *Computer Vis. image Underst.* **104**, 90–126 (2006).
15. Pavllo, D., Porssut, T., Herbelin, B. & Boulic, R. Real-time finger tracking using active motion capture: A neural network approach robust to occlusions. In *Proceedings of the 11th ACM SIGGRAPH Conference on Motion, Interaction and Games*, 1–10 (2018).
16. Iwashita, Y., Kurazume, R., Hasegawa, T. & Hara, K. Robust motion capture system against target occlusion using fast level set method. In *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006*., 168–174 (2006).
17. Ji, X. & Liu, H. Advances in view-invariant human motion analysis: A review. *IEEE Trans. Syst., Man, Cybern., Part C. (Appl. Rev.)* **40**, 13–24 (2009).
18. Yenduri, S., Perveen, N. & Chalavadi, V. et al. Fine-grained action recognition using dynamic kernels. *Pattern Recognit.* **122**, 108282 (2022).
19. Zhu, X., Huang, P.-Y., Liang, J., de Melo, C. M. & Hauptmann, A. G. Stmt: A spatial-temporal mesh transformer for mocap-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1526–1536 (2023).
20. Menolotto, M., Komaris, D.-S., Tedesco, S., O'Flynn, B. & Walsh, M. Motion capture technology in industrial applications: A systematic review. *Sensors* **20**, 5687 (2020).
21. Li, J., Liu, K. & Wu, J. Ego-body pose estimation via ego-head pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17142–17151 (2023).
22. Araújo, J. P. et al. Circle: Capture in rich contextual environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21211–21221 (2023).
23. Tevet, G. et al. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations* (2023). https://openreview.net/forum?id=SJ1kSyO2jwu.
24. Shafir, Y., Tevet, G., Kapon, R. & Bermano, A. H. Human motion diffusion as a generative prior. In *The Twelfth International Conference on Learning Representations* (2024). https://openreview.net/forum?id=dTpbEdN9kr.
25. Qiu, J. et al. Large AI models in health informatics: Applications, challenges, and the future. *IEEE Journal of Biomedical and Health Informatics* (2023).
26. Mahmood, N., Ghorbani, N., Troje, N. F., Pons-Moll, G. & Black, M. J. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, 5442–5451 (2019).
27. Zheng, C. et al. Deep learning-based human pose estimation: A survey. *ACM Comput. Surv.* **56**, 1–37 (Association for Computing Machinery, New York, NY, 2023).
28. Wang, J. et al. Deep 3d human pose estimation: a review. *Computer Vis. Image Underst.* **210**, 103225 (2021).
29. Pavlakos, G. et al. Expressive body capture: 3d hands, face, and body from a single image. 10975–10985 (2019).
30. Cui, R., Zhu, A., Wu, J. & Hua, G. Skeleton-based attention-aware spatial–temporal model for action detection and recognition. *IET Computer Vis.* **14**, 177–184 (2020).
31. Wang, H. & Wang, L. Beyond joints: Learning representations from primitive geometries for skeleton-based action recognition and detection. *IEEE Trans. Image Process.* **27**, 4382–4394 (2018).
32. Liu, C., Hu, Y., Li, Y., Song, S. & Liu, J. PKU-MMD: A large scale benchmark for skeleton-based human action understanding. In

Proceedings of the Workshop on Visual Analysis in Smart and Connected Communities, 1–8 (Association for Computing Machinery, 2017).
33. Xu, L., Wang, Q., Lin, X. & Yuan, L. An efficient framework for few-shot skeleton-based temporal action segmentation. *Computer Vis. Image Underst.* **232**, 103707 (2023).
34. Punnakkal, A. R., Chandrasekaran, A., Athanasiou, N., Quiros-Ramirez, A. & Black, M. J. BABEL: Bodies, action and behavior with english labels. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 722–731 (2021).
35. Dosovitskiy, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, (2021). https://openreview.net/forum?id=YicbFdNTTy.
36. Carion, N. et al. End-to-end object detection with transformers. In *European Conference on Computer Vision*, 213–229 (Springer, 2020).
37. Xu, M., Zhao, C., Rojas, D. S., Thabet, A. & Ghanem, B. G-TAD: Sub-graph localization for temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10156–10165 (2020).
38. Nawhal, M. & Mori, G. Activity graph transformer for temporal action localization. *arXiv preprint arXiv:2101.08540* (2021).
39. Zhang, C.-L., Wu, J. & Li, Y. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, 492–510 (Springer, 2022).
40. Zhu, X. et al. Deformable DETR: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations* (2021).
41. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G. & Black, M. J. SMPL: A skinned multi-person linear model. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)* **34**, 248:1–248:16 (2015).
42. Carreira, J. & Zisserman, A. Quo vadis, action recognition? A new model and the Kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308 (2017).
43. Shi, L., Zhang, Y., Cheng, J. & Lu, H. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12026–12035 (2019).
44. Fieraru, M. et al. Three-dimensional reconstruction of human interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7214–7223 (2020).
45. Müller, L., Osman, A. A. A., Tang, S., Huang, C.-H. P. & Black, M. J. On self-contact and human pose. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recogßnition (CVPR)* (2021).
46. Bloom, V., Makris, D. & Argyriou, V. G3D: A gaming action dataset and real time action recognition evaluation framework. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 7–12 (IEEE, 2012).
47. Sung, J., Ponce, C., Selman, B. & Saxena, A. Unstructured human activity detection from RGBD images. *2012 IEEE International Conference on Robotics and Automation* 842–849 (2012).
48. Wu, C., Zhang, J., Savarese, S. & Saxena, A. Watch-n-patch: Unsupervised understanding of actions and relations. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4362–4370 (2015).
49. Lillo, I., Soto, A. & Niebles, J. C. Discriminative hierarchical modeling of spatio-temporally composable human activities. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 812–819 (2014).
50. Yun, K., Honorio, J., Chattopadhyay, D., Berg, T. L. & Samaras, D. Two-person interaction detection using body-pose features and multiple instance learning. *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* 28–35 (2012).
51. Wei, Y. et al. 3d single-person concurrent activity detection using stacked relation network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**, 12329–12337 (2020).
52. Vaswani, A. et al. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008 (2017).
53. Shi, L., Zhang, Y., Cheng, J. & Lu, H. Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition. In *Proceedings of the Asian Conference on Computer Vision*, 38–53 (Springer, 2020).
54. Wei, Y., Li, W., Chang, M.-C., Jin, H. & Lyu, S. Explainable and efficient sequential correlation network for 3d single person concurrent activity detection. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 8970–8975 (2020).
55. Plizzari, C., Cannici, M. & Matteucci, M. Skeleton-based action recognition via spatial and temporal transformer networks. *Computer Vis. Image Underst.* **208**, 103219 (2021).
56. Zhang, Y., Wu, B., Li, W., Duan, L. & Gan, C. Stst: Spatial-temporal specialized transformer for skeleton-based action recognition. In *Proceedings of the ACM International Conference on Multimedia*, 3229–3237 (2021).
57. Pang, Y., Ke, Q., Rahmani, H., Bailey, J. & Liu, J. Igformer: Interaction graph transformer for skeleton-based human interaction recognition. In *European Conference on Computer Vision*, 605–622 (Springer, 2022).
58. Chen, Y. et al. Hierarchically self-supervised transformer for human skeleton representation learning. In European Conference on Computer Vision, 185–202 (Springer, 2022).
59. Kim, B., Chang, H. J., Kim, J. & Choi, J. Y. Global-local motion transformer for unsupervised skeleton-based action learning. In *European Conference on Computer Vision*, 209–225 (Springer, 2022).
60. Ionescu, C., Papava, D., Olaru, V. & Sminchisescu, C. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.* **36**, 1325–1339 (2014).
61. Shahroudy, A., Liu, J., Ng, T.-T. & Wang, G. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1010–1019 (2016).
62. Jiang, Y.-G. et al. THUMOS challenge: Action recognition with a large number of classes. http://crcv.ucf.edu/THUMOS14/ (2014).
63. Yeung, S. et al. Every moment counts: Dense detailed labeling of actions in complex videos. *International Journal of Computer Vision* (2017).
64. Caba Heilbron, F., Escorcia, V., Ghanem, B. & Carlos Niebles, J. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, 961–970 (2015).
65. Mahmood, N., Ghorbani, N., Troje, N. F., Pons-Moll, G. & Black, M. J. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5442–5451 (2019).
66. Dutta, A. & Zisserman, A. The via annotation software for images, audio and video. In *Proceedings of the 27th ACM international conference on multimedia*, 2276–2279 (2019).
67. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (2016).
68. Ba, J. L., Kiros, J. R. & Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).
69. Cui, Y., Jia, M., Lin, T.-Y., Song, Y. & Belongie, S. Class-balanced loss based on effective number of samples. 9268–9277 (2019).
70. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2980–2988 (2017).
71. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. In *International Conference for Learning Representations* (San Diego, 2015).
72. Lin, C. et al. Learning salient boundary feature for anchor-free temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3320–3329 (2021).

73. Alwassel, H., Giancola, S. & Ghanem, B. TSP: Temporally-sensitive pretraining of video encoders for localization tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3173–3183 (2021).
74. Veličković, P. et al. Graph attention networks. In *International Conference on Learning Representations* (2018).
75. Sun, J. et al. Locate source code. https://github.com/locate-bench/locate (2024).
76. Sung, J., Ponce, C., Selman, B. & Saxena, A. Unstructured human activity detection from rgbd images. In *International Conference on Robotics and Automation*, 842–849 (IEEE, 2012).
77. Li, Y. et al. Online human action detection using joint classification-regression recurrent neural networks. In *European conference on computer vision*, 203–220 (Springer, 2016).
78. Song, S., Lan, C., Xing, J., Zeng, W. & Liu, J. Spatio-temporal attention-based lstm networks for 3d action recognition and detection. *IEEE Trans. Image Process.* **27**, 3459–3471 (2018).

Author contributions

J.S., L.H., H.W., A.C., and M.B. designed the study and conducted the experiments. J.S., C.Z., J.Q., and A.C. implemented the computational models. A.C. designed the dataset. J.S., E.X., B.Z., M.I., and L.X. analyzed the experimental results. J.Q., A.C., and M.B. supervised the work. All authors contributed to the writing of the paper and reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44172-024-00272-7.

Correspondence and requests for materials should be addressed to Jiankai Sun or Jianing Qiu.

Peer review information *Communications Engineering* thanks Liangchen Song, Swati Nigam and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2024