EgoTracks: A Long-term Egocentric Visual Object Tracking Dataset

Hao Tang[1], **Kevin J Liang**[1], **Kristen Grauman**[1,2], **Matt Feiszli**[1,*], **Weiyao Wang**[1,*]
Meta AI[1], UT Austin[2], Equal Contribution[*]
{haotang, kevinjliang, grauman, mdf, weiyaowang}@meta.com

Abstract

Visual object tracking is key to many egocentric vision problems. However, the full spectrum of challenges of egocentric tracking faced by an embodied AI is underrepresented in many existing datasets, which tend to focus on short, third-person videos. Egocentric video has several distinguishing characteristics from those commonly found in past datasets: frequent large camera motions and hand interactions with objects commonly lead to occlusions or objects exiting the frame, and object appearance can change rapidly due to widely different points of view, scale, or object states. Embodied tracking is also naturally long-term, and being able to consistently (re-)associate objects to their appearances and disappearances over as long as a lifetime is critical. Previous datasets under-emphasize this re-detection problem, and their "framed" nature has led to adoption of various spatiotemporal priors that we find do not necessarily generalize to egocentric video. We thus introduce EgoTracks, a new dataset for long-term egocentric visual object tracking. Sourced from the Ego4D dataset, EgoTracks presents a significant challenge to recent state-of-the-art single-object trackers, which we find score more poorly on our new dataset than existing popular benchmarks, according to traditional tracking metrics. We further show improvements that can be made to the STARK tracker to significantly increase its performance on egocentric data, resulting in a baseline model we call EgoSTARK. We publicly release our annotations and benchmark (https://github.com/EGO4D/episodic-memory/tree/main/EgoTracks), hoping our dataset leads to further advancements in tracking.

1 Introduction

First-person or "egocentric" computer vision aims to capture the real-world perceptual problems faced by an embodied AI; it has drawn strong recent interest as an underserved but highly relevant domain of vision, with important applications ranging from robotics [63, 18] to augmented and mixed reality [2, 65, 28]. Visual object tracking (VOT), long a fundamental problem in vision, is a core component to many egocentric tasks, including tracking the progress of an action or activity, (re-)association of objects in one's surroundings, and predicting future states of the environment. Yet, while the VOT field has made many significant advancements over the past decade, tracking in egocentric video remains underexplored. This lack of attention is in large part due to the absence of a large-scale egocentric tracking dataset for training and evaluation. While the community has proposed a number of popular tracking datasets in recent years, including OTB [76], TrackingNet [57], GOT-10k [32], and LaSOT [21], we find that the strong performance that state-of-the-art trackers achieve on these benchmarks does not translate well to egocentric video, thus establishing a strong need for such a tracking dataset.

We attribute this performance gap to the many unique aspects of egocentric views compared to the more traditional third-person views of previous datasets. In contrast to intentionally "framed" video, egocentric videos are often uncurated, meaning they tend to capture many attention shifts between

37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks.

Figure 1: A video from the proposed EgoTracks dataset, with yellow clip segments marking when the object (`blowtorch`) is visible. Note the frequent disappearances and reappearances of the object over an 8 minute video, with lengthy absences, necessitating re-detection to track accurately without false positives. The egocentric nature of the video includes the camera-wearer interacting with the object (Occurrence 2), resulting in significant hand occlusions and dramatic changes in pose.

activities, objects, or locations. Due to the first-person perspective, large head motions from the camera wearer often result in objects repeatedly leaving and re-entering the field of view; similarly, hand manipulations of objects [64] leads to frequent occlusions, rapid variations in scale and pose, and potential changes in state or appearance. Furthermore, egocentric video tends to be long (sometimes representing the entire life of an agent or individual), meaning the volume of the aforementioned occlusions and transformations scales similarly. These characteristics all make tracking objects in egocentric views dramatically more difficult than scenarios commonly considered in prior datasets, and their absence represents an evaluation blindspot.

Head motions, locomotion, hand occlusions, and temporal length lead to several challenges. First, frequent object disappearances and reappearances causes the problem of *re-detection* within egocentric tracking to become especially critical. Many previous tracking datasets primarily focus on short-term tracking in third-person videos, providing limited ability to evaluate many of the challenges of long-term egocentric tracking due to low quantity and length of target object disappearances. As a result, competent re-detection is not required for strong performance, leading many recent short-term trackers to neglect it, instead predicting a bounding box for every frame, which may lead to rampant false positives or tracking the wrong object. Additionally, the characteristics of short-term third-person video have also induced designs relying on gradual changes in motion and appearance. As we later show (Section 5.2), many of the motion, context, and scale priors made by previous short-term tracking algorithms fail to transfer to egocentric video.

Notably, re-detection, occlusions, and longer-term tracking have long been recognized as difficult for VOT as a field, leading to recent benchmark construction efforts [51, 11, 55, 68, 33, 70] emphasizing these aspects. We argue that egocentric video provides a natural source for these challenges at scale while also representing a highly impactful application for tracking, therefore constituting a significant opportunity. We thus present **EgoTracks**: a large-scale long-term egocentric visual object tracking dataset for training and evaluating long-term trackers. Seeking a realistic challenge, we source videos from Ego4D [28], a large-scale dataset consisting of unscripted, in-the-wild egocentric videos of daily-life activities. The result is a large-scale dataset to evaluate the tracking and re-detection ability of SOT models, with more than 22,028 tracks from 5708 average 6-minute videos. This constitutes the first large-scale dataset for visual ob-

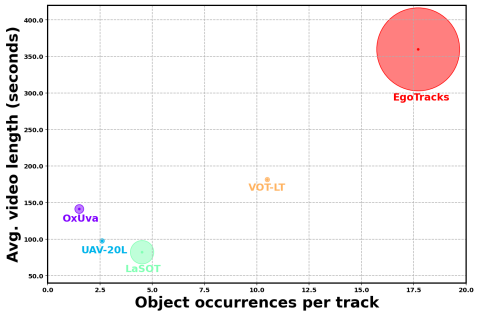

Figure 2: EgoTracks is an order of magnitude larger than past long-term VOT datasets, with significantly more tracks (circle area) and object disappearances/appearances in longer videos.

ject tracking in egocentric videos in diverse settings, providing a new, significant challenge compared with previous datasets.

We perform a thorough analysis of our new dataset and its new characteristics relative to prior benchmarks, demonstrating its difficulty and the need for further research to develop trackers capable of handling long-term egocentric vision. Our experiments reveal remaining open problems and insights towards promising directions in egocentric tracking. Leveraging these intuitions, we propose multiple simple yet effective changes, such as adjustment of spatiotemporal priors, egocentric data finetuning, and combining multiple templates. We apply these proposed strategies on a state-of-the-art (SOTA) STARK tracker [79], training a strong tracker dedicated to long-term egocentric tracking: **EgoSTARK**. We hope Ego-STARK can serve as a strong baseline and facilitate future research.

We make the following contributions:

1. We present EgoTracks, the first large-scale long-term object tracking dataset with diverse egocentric scenarios. We analyze its uniqueness in terms of evaluating the re-detection performance of trackers.
2. We conduct comprehensive experiments to understand the performance of many state-of-the-art trackers on the EgoTracks validation set and observe that due to the biases and evaluation blindspots of existing third-person datasets, they tend to struggle.
3. We perform an analysis of what makes a good tracker for long-form egocentric video. Applying these learnings to the STARK tracker [79], we produce a strong baseline we call EgoSTARK, which achieves significant improvements (+15% F-score) on EgoTracks.

2 Related work

2.1 Visual object tracking datasets

Visual object tracking studies the joint spatial-temporal localization of objects in videos. From a video and a predefined taxonomy, multiple object tracking (MOT) models simultaneously detect, recognize, and track multiple objects. For example, MOT [54] tracks humans, KITTI [25, 50] tracks pedestrians and cars, and TAO [15] tracks a large taxonomy of 833 categories. In contrast, single object tracking (SOT) follows a single object via a provided initial template of the object, without any detection or recognition. Thus, SOT is often taxonomy-free and operates on generic objects. The community has constructed multiple popular benchmarks to study this problem, including OTB [76], UAV [56], NfS [36], TC-128 [45], NUS-PRO [41], GOT-10k [32], VOT [38], and TrackingNet [57].

While these SOT datasets mainly consist of short videos (e.g. a few seconds), long-term tracking has been increasingly of interest. Tracking objects in longer videos (several minutes or more) poses unique challenges, e.g. significant transformations, displacements, disappearances, and reappearances. On top of localizing the object when visible, the model also must produce no box when the object is absent, and then re-localize the same object when it reappears. OxUvA [68] is one of the first to benchmark longer videos (average 2 minutes), with 366 evaluation-only videos. LaSOT [21] scales this to 1400 videos with more frequent object reappearances. Concurrently, VOT-LT [37] includes frequent object disappearances and reappearances in 50 purposefully selected videos.

Our EgoTracks focuses on long-term SOT and presents multiple critical and unique attributes: 1) significantly larger scale, with **5708** videos of an average **6 minutes** (Figure 2); 2) more frequent disappearances & reappearances (avg. **17.7** times) happening in natural, real-world scenarios; 3) data sourced from egocentric videos shot in-the-wild, involving unique challenging situations, such as large camera motions, diverse perspective changes, hand-object interactions, and frequent occlusions.

2.2 Single object tracking methodologies

Many modern approaches use convolutional neural networks (CNNs), either with Siamese network [43, 71, 42] or correlation-filter based [13, 3, 8, 53, 4] architectures. With recent successes in vision tasks like classification [17] and detection [5], Transformer architecture [69] for tracking have also become popular. For example, TransT [6] uses attention-based feature fusion to combine features of the object template and search image. More recently, several works utilize Transformers as direct predictors to achieve a new state of the art, such as STARK [79], ToMP [52] and SBT [77]. These models tokenize frame features from a ResNet [30] encoder, and use a Transformer to predict the bounding box and object presence score with the feature tokens. These methods are often developed on short-term SOT datasets and assume that the target object stays in the field of view with minimum occlusions. On the other hand, long-term trackers [70, 33, 11] are designed to cope with

Table 1: **Object tracking datasets comparison.** In addition to larger scale than previous datasets, the scenarios captured by EgoTracks represent a significantly harder challenge for SOTA trackers, suggesting room for improved tracking methodology.

Dataset	Video Hours	Avg. Length (s)	Ann. FPS	Ann. Type	Egocentric	SOTA (P/AO)*
ImageNet-Vid [62]	15.6	10.6	25	mask	No	
YT-VOS [78]	5.8	4.6	5	mask	No	-/83.6 [31]
DAVIS 17 [61]	0.125	3	24	mask	No	-/86.3 [7]
TAO [15]	29.7	36.8	1	mask	No	
UVO [74]	2.8	10	30	mask	No	-/73.7 [58]
EPIC-KITCHENS VISOR [14]	36	12**	0.9	mask	**Yes**	-/74.2 [58]
GOT-10k [32]	32.8	12.2	10	bbox	No	-/75.6 [9]
OxUvA [68]	14.4	141.2	1	bbox	No	
LaSOT [21]	31.92	82.1	30	bbox	No	80.3/- [9]
TrackingNet [57]	125.1	14.7	28	bbox	No	86/- [9]
TREK-150 [19, 20]	0.45	10.81	60	bbox	**Yes**	-/50.5 [19, 20]
EgoTracks (Ours)	**602.9**	**367.9**	5	bbox	**Yes**	45/54.1

*: P: Precision, AO: average overlap (J-Score for mask-based datasets). **: Original video is 720s.

the problem of re-detecting objects in their reappearances. Designed to be aware of potential object disappearances, these approaches search the whole image for its reappearance.

2.3 Tracking in egocentric videos

Multiple egocentric video datasets have been introduced in the past decades [12, 28, 39, 66, 60, 23], offering a host of interesting challenges, many of which require associating objects across frames: activity recognition [35, 44, 80, 75, 26], anticipation [22, 24, 27], video summarization [16, 39, 40, 49], human-object interaction [14, 47], episodic memory [28], visual query [28], and camera-wearer pose inference [34]. To tackle these challenges, tracking is leveraged in many methodologies [28, 14, 48, 40, 47], yet few works have been dedicated to this fundamental problem on its own. [19, 20] have started to recognize the challenges of egocentric object tracking and might be the most related work to ours. The major difference, however, is the scale of the dataset: [19, 20] contain 150 tracks intended only for evaluation, while EgoTracks is 100× larger (see Table 1), containing 20k tracks with training and evaluation splits. Also, while past efforts have sourced videos from the kitchen-heavy EPIC-KITCHEN [12], EgoTracks sources videos from Ego4D [28], which has more diverse scenarios. EgoTracks provides a unique, large-scale testbed for developing tracking methods dedicated to egocentric videos; our improved baseline EgoSTARK also serves as a potential plug-and-play module to solve other tasks where object association is desired.

In egocentric video understanding, Ego4D [28] and EPIC-KITCHENS VISOR [14] are closely related. Ego4D contains the largest collection of egocentric videos in-the-wild; EgoTracks is annotated on a subset of Ego4D. In addition, Ego4D proposes many novel tasks, such as Episodic Memory, with tracking identified as a core component. VISOR was introduced concurrently, annotating short-term (12 sec on average) videos from EPIC-KITCHENS [12] with instance segmentation masks. We believe EgoTracks offers multiple unique values complementary to EPIC-VISOR: long-term tracking (6 min vs. 12 sec), significantly larger-scale (5708 video clips vs. 158), and more diversified video sources (80+ scenes vs. kitchen-only; see Appendix A).

3 The EgoTracks dataset

We present EgoTracks: a large-scale long-term egocentric single object tracking dataset, consisting of a total of 22028 tracks from 5708 videos. We follow the same data split as the Ego4D Visual Queries (VQ) 2D benchmark.(See Supplementary for details).

3.1 Ego4D visual queries (VQ) benchmark

Ego4D [28] is a massive-scale egocentric video dataset, consisting of 3670 hours of diverse daily-life activities of consenting participants in an in-the-wild format; the videos have personally identifiable information removed and were screened for offensive content. The dataset is accompanied by multiple benchmarks, but the most relevant task for our purposes is episodic memory's 2D VQ task: Given an egocentric video and a cropped image of an object, the goal is to localize when and where the object was last seen in the video, as a series of 2D bounding boxes in consecutive frames. This task is closely related to long-term tracking: finding an object in a video given a visual template is identical to the re-detection problem in long-term tracking. Moreover, Ego4D's baselines rely heavily on tracking methods: Siam-RCNN [70] and KYS [4] for global and local tracking, respectively.

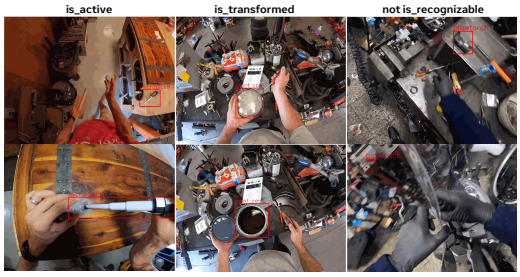

Figure 3: **EgoTracks tracklet attribute examples.** *Left*: Micropipette on a bench (top) versus actively used (bottom). *Middle*: A paint can (top) is opened (bottom). *Right*: A blowtorch (top) requiring context from other frames to identify due to distance and motion blur (bottom).

Table 2: **Track attributes** in train/val sets.

	Total number	Percentage
All Tracks	17593	100%
`is_active`	3963	22.52%
`is_transformed`	1080	6.13%
`is_recognizable`	17557	99.79%

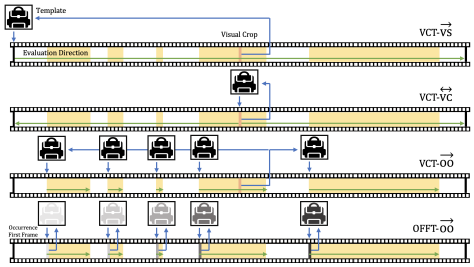

Table 3: Evaluation protocols visualization.

Shortcomings. While highly related, the VQ dataset is not immediately suitable long-term tracking. In particular, the VQ annotation guidelines were roughly the following: 1) identify three different objects that appear multiple times in the video; 2) annotate a query template for each object, which should contain the entire object without any motion blur; 3) annotate an occurrence of the object that is temporally distant from the template. Thus, these annotations are not exhaustive over time (they are quite sparse), limiting their applicability to tracking. On the other hand, the selection criteria result in a strong set of candidate objects, which we leverage to build EgoTracks.

3.2 Annotating VQ for long-term tracking

We thus start with the VQ visual crop and response track, asking annotators to first identify the object represented by the visual crop, the response track, and object name. From the video's start, we instruct the annotators to draw a bounding box around the object each time it appears. Because annotators must go through each video in its entirety, which contain an average of $\sim$1800 frames at 5 frames per second (FPS), this annotation task is labor-intensive, taking roughly 1 to 2 hours per track. An important aspect of this annotation is its exhaustiveness: the entire video is densely annotated for the target object, and any frame without a bounding box is considered as a negative. Being able to reject negatives examples is an important component to re-detection in real-world settings, as false positives can impact certain applications as much as false negatives.

Quality Assurance. All tracks are quality checked by expert annotators after the initial annotations. To measure the annotation quality, we employ multi-review on a subset of the validation set. Three independent reviewers are asked to annotate the same video. We find the overlaps between these independent annotations are high (> 0.88 IoU). Further, since EgoTracks has a focus on re-detection, we check the temporal overlap of object presence and find it to be consistent across annotators. In total, the entire annotation effort represented roughly 86k worker-hours of effort.

3.3 Tracklet attributes

In addition to the bounding box annotations, we also label certain relevant attributes to allow for different training strategies or deeper analysis of validation set performance. We annotate the following three attributes per occurrence (see Figure 3 for examples and Table 2 for statistics):

- `is_active`: In Ego4D, the camera wearer often interacts with relevant objects with their hands. Objects in the state of being handled pose a challenge for tracking algorithms due to frequent occlusion and rapid changes in pose.
- `is_transformed`: Objects in Ego4D may undergo transformations, such as deformations and state changes. Such instances require being able to quickly adapt to the tracked object having a new appearance.
- `is_recognizable`: Due to occlusions, motion blur, scale, or other conditions, some objects in Ego4D can be extremely difficult to recognize without additional context. We thus annotate if the object is recognizable solely based on its appearance, without using additional context information (*e.g.* other frames).

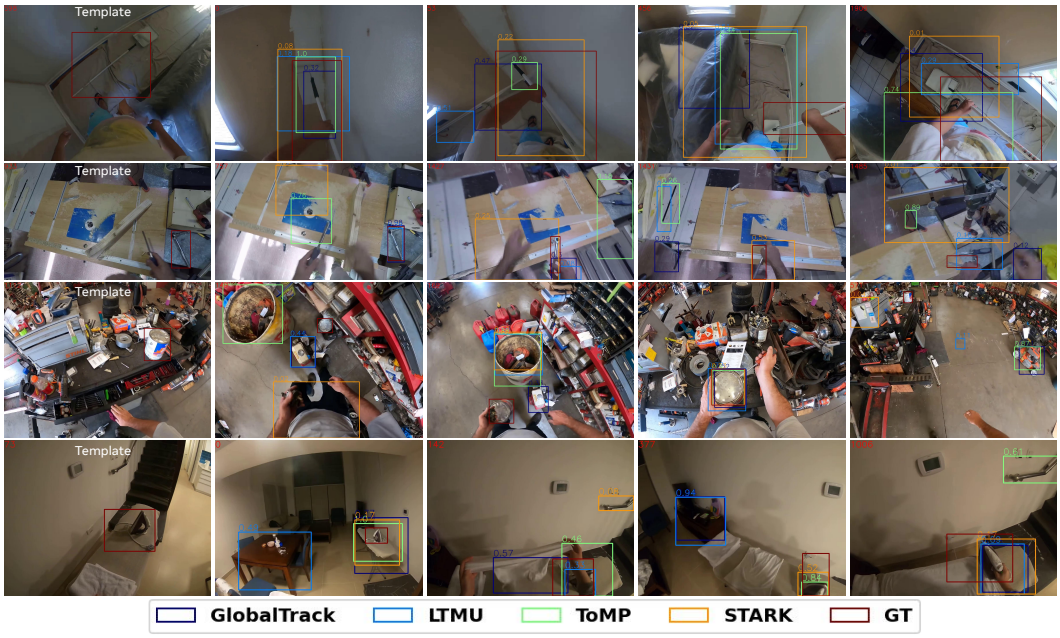

Figure 4: Qualitative results of different trackers. EgoTracks presents significant challenges for all trackers, due to drastic viewpoint changes, occlusions, changes in scale, head motion etc.

4 Analysis of state-of-the-art SOT trackers

We compare the performance of several off-the-shelf tracking models on EgoTracks's validation set. Identifying STARK [79] as the one with the best performance, we conduct further ablation studies under different settings using STARK to further understand its behavior.

4.1 Evaluation protocols and metrics

Evaluation Protocols. We introduce several evaluation protocols for EgoTracks, consisting of different combinations of the initial template, evaluated frames, and the temporal direction in which the tracker is run. For the initial template, we consider two choices:

- **Visual Crop Template (VCT)**: The visual crop images were specifically chosen to be high-quality views of the target and served as our annotators' references for identifying the object throughout the videos. Thus, they make ideal candidates for initializing a tracker.
- **Occurrence First Frame Template (OFFT)**: The tracker is initialized with the first frame of each occurrence (see $\overrightarrow{\text{OO}}$ below). While this may result in a lower quality view of the object, temporal proximity to subsequent frames means it may be closer in appearance.

Note that we exclude the template frame from the calculation of any evaluation metrics. We also consider several choices for the evaluated frames and temporal direction:

- **Video Start Forward ($\overrightarrow{\text{VS}}$)**: The tracker is evaluated on every frame of the video in causal order, starting from the first frame. This represents a tracker's ability to follow an object through a long video.
- **Visual Crop Forward/Backward ($\overleftrightarrow{\text{VC}}$)**: The tracker is run on the video twice, once starting at the visual crop frame and running forward and time, and a second time running backwards. This represents an alternative way of covering every frame in the video, but with closer visual similarity between **VCT** initialization and the first frames encountered by the tracker.
- **Occurrences Only Forward ($\overrightarrow{\text{OO}}$)**: The tracker is only evaluated on the object occurrences, when the object is visible. This simplifies the tracking task and allows us to dis-entangle the challenge of re-detection from that of simply tracking in an egocentric clip.

We specify protocols by concatenating the appropriate descriptors. We primarily consider **VCT-$\overrightarrow{\text{VS}}$**, **VCT-$\overleftrightarrow{\text{VC}}$**, **VCT-$\overrightarrow{\text{OO}}$**, and **OFFT-$\overrightarrow{\text{OO}}$** (Table 3) in our experiments.

Table 4: **EgoTracks performance comparison.** Off-the-shelf, all trackers perform poorly, demonstrating the new challenges of EgoTracks. Higher performance from tracking by detection methods + Oracle imply that instance association, not detection, is one of the primary challenges.

Method	AO	F-score	Precision	Recall	FPS
KYS [4]	16.09	13.09	12.50	13.74	20
DiMP [3]	16.45	11.84	10.31	13.91	43
GlobalTrack [33] †	23.63	20.40	31.28	15.14	6
LTMU [11] †	29.33	27.46	37.28	21.74	13
ToMP [52]	30.93	20.95	19.63	22.46	24.8
Siam-RCNN [70] †	37.48	35.38	52.80	26.67	4.7
MixFormer (MixViT-L, ConvMAE) [9, 10]	27.93	25.54	28.30	23.27	10
STARK [79] - Res50	35.99	30.48	34.70	27.17	41.8
STARK [79] - Res101	35.03	30.18	35.30	26.35	31.7
Tracking by Detection †					
Mask R-CNN [29]+Oracle	60.00	-	-	-	
GGN [73]+Oracle	75.92	-	-	-	
GGN+InstEmb	15.19	9.92	11.75	8.58	

†: trackers with re-detection

Table 5: Comparing tracker initializations. (Left) Comparison of trackers initialized from the first frame in each occurrence and tracking only that single occurrence (oracle re-detection). (Right) STARK whole-video performance, starting from video start frame vs. the visual crop frame.

Method	AO	Success	Pre	Pre_{norm}
KYS [4]	33.92	34.87	31.22	34.87
DiMP [3]	34.80	35.70	32.13	38.98
ToMP [52]	45.17	45.93	41.74	47.88
STARK [79]	50.01	50.64	45.76	51.91

Method	AO	F-score	Precision	Recall
STARK - **VCT-$\overrightarrow{\text{VS}}$**	35.99	30.48	34.70	27.17
STARK - **VCT-$\overleftrightarrow{\text{VC}}$**	40.01	34.02	38.31	30.60

Metrics. We adopt common metrics in object tracking, including F-score, precision, and recall; details can be found in [51]. Trackers are ranked mainly by the F-score. We additionally consider average overlap (AO), success, precision, and normalized precision as short-term tracking metrics [67].

4.2 SOT trackers struggle on EgoTracks

We compare the performance of select trackers on EgoTracks with the **VCT-$\overrightarrow{\text{VS}}$** evaluation protocol. Given the breadth of tracking algorithms, we do not aim to be exhaustive but select high-performing representatives of different tracking principles. KYS [4] and DiMP [3] are two short-term tracking algorithms that maintain an online target representation. ToMP [52], STARK [79] and MixFormer [9, 10] are three examples of the SOTA short-term trackers based on Transformers. GlobalTrack [33] is a global tracker that searches the entire search image for re-detection. LTMU [11] is a high performance long-term tracker that combines a global tracker (GlobalTrack) with a local tracker (DiMP). Siam R-CNN [70] leverages dynamic programming to model a full path of history for long-term. The performance of these trackers on EgoTracks are summarized in Table 4. AO in this table is equivalent to recall at the probability threshold 0. Qualitative results are shown in Figure 4.

We highlight several observations. First, the object presence scores from most short-term trackers are not very useful, as can be seen from the low precision of KYS (12.5), DiMP (13.91), and ToMP (19.63), while long-term trackers like GlobalTrack, DiMP_LTMU and Siam R-CNN achieve higher precisions at 31.28, 37.28 and 52.8. This is expected as long-term trackers are designed to place more emphasis on high re-detection accuracy, though there clearly is still room for improvement. STARK achieves the second highest precision at 34.70, which is an exception as it has a second training stage to teach the model to classify whether the object is present. Second, more recent works such as MixFormer and STARK achieve better F-score than previous short-term trackers. This could be partially due to advances in training strategies, more data, and Transformer-based architectures. Surprisingly, we found recent MixFormer [10] does not outperform STARK, despite achieving new SOTA on its training dataset. This highlights a potential difficulty in generalization.

We also include results using the principle of Tracking by Detection [59, 1]: a detector proposes 100 bounding boxes, and we select the best using cosine similarity of box features. We observe that an open-world detector GGN [73] trained on COCO [46] generalize reasonably well with oracle matching, achieving 75.92 AO. However, the association problem is very challenging, bringing down the AO to 15.19. Implementation details can be found in Appendix B.

4.3 Re-detection and diverse views are challenging

We perform additional EgoTracks experiments following alternative evaluation protocols to gain further insights on tracker performance (Table 5). To decouple the re-detection problem from other

Table 6: **OFFT-$\overrightarrow{\text{OO}}$** AO of standard STARK model [79] for each attribute.

Attribute	True	False
`is_active`	49.65	55.73
`is_transformed`	49.19	55.31
`is_recognizable`	55.52	46.65

Table 7: Performance of trackers finetuned on EgoTracks.

Method	AO	F-score	Precision	Recall
ToMP	36.13	28.11	29.01	27.26
Siam-RCNN	45.67	41.41	56.11	32.81
STARK	44.25	38.20	42.06	34.99

Table 8: Train/test-time hyperparameters comparison.

	Method	AO	F-score	Precision	Recall
Data	STARK	35.99	30.48	34.70	27.17
	STARK - ft on VQ	38.94	33.53	39.13	29.33
	STARK - ft on EgoTracks	44.25	38.20	42.06	34.99
Augmentation	STARK - ft on VQ	38.94	33.53	39.13	29.33
	STARK - ft + multiscale	48.44	41.92	42.65	41.30
Search window	search_size = 320	35.99	30.48	34.70	27.17
	search_size = 480	48.21	39.69	43.95	36.19
	search_size = 640	52.09	42.39	46.23	39.15
	search_size = 800	54.08	43.74	47.60	40.45

Table 9: STARK with different context ratios. Bold row is the default setting. **CR**: context ratio, **SRR**: search region ratio, **SIS**: search image size (in image resolution).

Setting	Method CR	SRR	SIS	AO	F-score	Precision	Recall
Same SIS	1x	2.5x	320	28.22	26.81	28.68	25.16
	2x	**5x**	**320**	38.94	33.53	39.13	29.33
	3x	7.5x	320	44.70	36.03	40.28	32.59
	4x	10x	320	43.19	34.32	37.98	31.31
Same SRR	1x	5x	640	41.50	31.09	30.31	31.91
	3x	5x	208	39.87	35.36	41.54	30.79
Same CR	2x	7.5x	480	48.21	39.69	43.95	36.19
	2x	10x	640	52.09	42.39	46.23	39.15
	2x	12.5x	800	54.08	43.74	47.60	40.45

egocentric aspects of EgoTracks, we evaluate with the **OFFT-$\overrightarrow{\text{OO}}$** protocol, which ignores the negative frames of the video, thus obviating the need for re-detection. Unsurprisingly, all trackers do significantly better, emphasizing the challenging nature of re-detection in EgoTracks. We also run experiments in the **VCT-$\overleftrightarrow{\text{VC}}$** setting, where the initial template is temporally adjacent to the first tracked frames. Here we see a 3-4% improvement to AO, F-score, precision, and recall compared to the **VCT-$\overrightarrow{\text{VS}}$** protocol, illustrating that trackers like STARK are designed to expect gradual transitions in appearance. Both these experiments illustrate that the re-detection problem is a significant challenge for tracking and the need for better long-term benchmarks.

4.4 Attributes capture hard scenarios for tracking

We use the validation set tracklet attribute annotations described in Section 3.3 to further understand performance on our evaluation set. For each attribute, we split the tracklets into two groups, corresponding to the attribute being true and false. We then use a standard STARK tracker [79] and report AO for each group of tracklets using the **OFFT-$\overrightarrow{\text{OO}}$** evaluation protocol in Table 6. As might be expected, we find that when objects are being actively used by the user or in the midst of a transformation, AO tends to be lower, by roughly 6%, likely due to occlusions or changes in appearance. Additionally, STARK tends to have a harder time when the object is hard to recognize in the image, whether due to occlusions, blur, scale, or other conditions.

5 Egocentric tracking design considerations

Observing that existing trackers do not perform well on EgoTracks, we perform a systematic exploration of priors and other design choices for egocentric tracking. Though not specifically designed for long-term tracking, Section 4 suggests STARK [79] to be the most competitive tracker on EgoTracks. We focus on this tracker for additional analysis, suggesting improvements to egocentric performance.

5.1 Egocentric finetuning is essential

We first demonstrate how various trackers trained on third-person videos can significantly benefit from finetuning on EgoTracks. As shown in Table 7, all methods gain improvement on F-score ranging from 6% - 10%. In addition, as shown in Table 8, finetuning on the VQ response track subset improves the F-score from 30.48% to 33.53%, while using the full EgoTracks annotation further improves the F-score by 4.67% to 38.2%. This demonstrates that: 1) finetuning with egocentric data helps close the exocentric-egocentric domain gap; 2) training on full EgoTracks provides further gains, showing the value of our training set.

5.2 Third-person spatiotemporal priors fail

Modern SOTs find certain assumptions on object motion, appearance, and surroundings helpful on past datasets, but some of these design choices translate poorly to long-term egocentric videos.

Search window size. An example is local search. Many trackers assume the tracked object appears within a certain range of its previous location. Thus, for efficiency, these methods often search within a local window of the next frame. This is reasonable in high FPS, smooth videos with relatively

slow motion, commonly in previous short-term tracking data, but in egocentric videos, the object's pixel coordinates can change rapidly (frequent large head motions), and re-detection becomes a key problem. Therefore, we experiment with expanded search regions beyond what are common in past methods. As we expand search size from 320 to 800, we see dramatic improvements (Table 8): STARK is able to locate objects that were previously outside search window due to rapid movements.

Multiscale augmentations. The characteristics of egocentric video also affect common SOT assumptions of object scale. Many trackers are trained with the assumption that an object's scale is consistent with the template image and between adjacent frames. However, large egocentric camera motions, locomotion, and hand interactions with objects (*e.g.* bringing an object to one's face, as in eating) can translate to objects rapidly undergoing large changes in scale. We thus propose adding scale augmentations during training, randomly resizing the search image by a factor of $s \in [0.5, 1.5]$. While simple, we find this dramatically improves performance on EgoTracks, improving STARK's AO by nearly 10% and F-score by more than 8% (Table 8).

Context ratio. Past SOT works have found that including some background can be helpful for template image feature extraction, with twice the size of the object being common. We experiment with different context ratios to see if this rule of thumb transfers to egocentric videos. Because of the local window assumption, the sizes of the template and search images are related: $\frac{\text{Search Image Size(SIS)}}{\text{Search Region Ratio(SRR)}} = \frac{\text{Template Image Size}}{\text{Context Ratio(CR)}} = \text{Object Scale}$. The template image size is set to a fixed size 128×128. When changing the context ratio, we carefully control the other parameters for a fair comparison. The results are shown in Table 9. Among all three parameters - **CR**, **SRR**, and **SIS**, the search region size (determined by **SRR** and **SIS**) has the highest impact on the F-score. This is expected because there are frequent re-detections, which require the tracker to search in a larger area for the object, rather than just within the commonly used local window. Varying the **CR** has mixed results so we adhere to the common practice of using a **CR** of 2.

6 Future directions

Based on our experiments in Table 4 and Table 5, we found that re-detection to be a key challenge of long-term tracking, especially in egocentric video, where objects frequently go in and out of view, or are exposed to high motion blur. We see a few promising directions for future works:

a) Stronger features for associating objects should significantly improve re-detection; the impact of insufficiently discriminative feature embeddings can be clearly seen in the major gap in Tracking by Detection performance between the Oracle and InstEmb methods at the bottom of Table 4. Geometric keypoints, optical flow, or long-term trajectories [72] can also lead to large improvements here.

b) Leveraging spatial signals: camera trajectories can be estimated as additional signals to the tracker. For example, if an object remains static during the window where it is out-of-view, knowledge of camera location can help re-localize the position of this object.

c) Global, multi-view object representations: Egocentric videos, with their diverse camera trajectories and tendency to capture the camera wearer's interactions with objects, often offer significantly richer and more varied viewpoints of objects than traditional third-person tracking datasets. In the latter, object appearances tend to be more constant, so modern tracking methods have thus far been able to get away with using a single image template (optionally with an additional template from the latest frame). With a need for more robustness to the different viewpoints and occlusions offered by egocentric video, we believe that a challenging egocentric tracking dataset like EgoTracks represents an opportunity to develop trackers with more global, view-variant object representations learned in an online fashion. A simple version of this can be found in Section D of the supplementary material, where we augmented EgoSTARK by fusing multiple templates; we found that such a strategy indeed improved tracking results on EgoTracks.

7 Conclusion

We present EgoTracks, the first large-scale dataset for long-term egocentric visual object tracking in diverse scenes. We conduct extensive experiments to understand the performance of state-of-the-art trackers on this new dataset, and find that they struggle considerably, possibly in part due to overfitting to some of the simpler characteristics of existing benchmarks. We thus propose several adaptations for the egocentric domain, leading to a strong baseline that we call Ego-STARK, which has vastly

improved performance on EgoTracks. Lastly, we plan to organize a public benchmark challenge using a held-out test set with a test server as a testbed for new tracking algorithms. By publicly releasing this dataset and organizing the challenge, we hope to encourage advancements in the field of long-term tracking and draw more attention to the challenges of long-term and egocentric videos.

References

[1] S. Avidan. Ensemble tracking. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 494–501 vol. 2, 2005. 7

[2] Ronald T Azuma. A survey of augmented reality. *Presence: teleoperators & virtual environments*, 6(4):355–385, 1997. 1

[3] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6182–6191, 2019. 3, 7

[4] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Know your surroundings: Exploiting scene information for object tracking. In *European Conference on Computer Vision*, pages 205–221. Springer, 2020. 3, 4, 7

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 3

[6] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8126–8135, 2021. 3

[7] Ho Kei Cheng and Alexander G. Schwing. XMem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *ECCV*, 2022. 4

[8] Seokeon Choi, Junhyun Lee, Yunsung Lee, and Alexander Hauptmann. Robust long-term object tracking via improved discriminative model prediction. In *European Conference on Computer Vision*, pages 602–617. Springer, 2020. 3

[9] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. Mixformer: End-to-end tracking with iterative mixed attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13608–13618, 2022. 4, 7

[10] Yutao Cui, Cheng Jiang, Gangshan Wu, and Limin Wang. Mixformer: End-to-end tracking with iterative mixed attention, 2023. 7

[11] Kenan Dai, Yunhua Zhang, Dong Wang, Jianhua Li, Huchuan Lu, and Xiaoyun Yang. High-performance long-term tracking with meta-updater. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6298–6307, 2020. 2, 3, 7

[12] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018. 4

[13] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Atom: Accurate tracking by overlap maximization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4660–4669, 2019. 3

[14] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Ely Locke Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 4

[15] Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. Tao: A large-scale benchmark for tracking any object. In *European conference on computer vision*, pages 436–454. Springer, 2020. 3, 4

[16] Ana Garcia del Molino, Cheston Tan, Joo-Hwee Lim, and Ah-Hwee Tan. Summarization of egocentric videos: A comprehensive survey. *IEEE Transactions on Human-Machine Systems*, 47(1):65–76, 2017. 4

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[18] Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2022. 1

[19] Matteo Dunnhofer, Antonino Furnari, Giovanni Maria Farinella, and Christian Micheloni. Is first person vision challenging for object tracking? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2698–2710, 2021. 4

[20] Matteo Dunnhofer, Antonino Furnari, Giovanni Maria Farinella, and Christian Micheloni. Visual object tracking in first person vision. *International Journal of Computer Vision*, 131(1):259–283, 2023. 4

[21] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5374–5383, 2019. 1, 3, 4

[22] Yazan Abu Farha, Alexander Richard, and Juergen Gall. When will you do what? - anticipating temporal occurrences of activities. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5343–5352, 2018. 4

[23] Alircza Fathi, Jessica K. Hodgins, and James M. Rehg. Social interactions: A first-person perspective. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1226–1233, 2012. 4

[24] Antonino Furnari and Giovanni Farinella. Rolling-unrolling lstms for action anticipation from first-person video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP:1–1, 05 2020. 4

[25] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 3

[26] Deepti Ghadiyaram, Matt Feiszli, Du Tran, Xueting Yan, Heng Wang, and Dhruv Kumar Mahajan. Large-scale weakly-supervised pre-training for video action recognition. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12038–12047, 2019. 4

[27] R. Girdhar and K. Grauman. Anticipative video transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13485–13495, Los Alamitos, CA, USA, oct 2021. IEEE Computer Society. 4

[28] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 1, 2, 4, 14

[29] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 7

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[31] Zhiwei Hu, Bo Chen, Yuan Gao, Zhilong Ji, and Jinfeng Bai. 1st place solution for youtubevos challenge 2022: Referring video object segmentation, 2022. 4

[32] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1562–1577, 2019. 1, 3, 4

[33] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Globaltrack: A simple and strong baseline for long-term tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11037–11044, 2020. 2, 3, 7

[34] Hao Jiang and Kristen Grauman. Seeing invisible poses: Estimating 3d body pose from egocentric video. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3501–3509, 2017. 4

[35] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 4

[36] Hamed Kiani Galoogahi, Ashton Fagg, Chen Huang, Deva Ramanan, and Simon Lucey. Need for speed: A benchmark for higher frame rate object tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1125–1134, 2017. 3

[37] Matej Kristan, Aleš Leonardis, Jiří Matas, Michael Felsberg, Roman Pflugfelder, Luka Čehovin Zajc, Tomáš Vojíř, Goutam Bhat, Alan Lukežič, Abdelrahman Eldesokey, Gustavo Fernández, et al. The sixth visual object tracking vot2018 challenge results. In *ECCV 2018 Workshops*, Cham, 2019. Springer International Publishing. 3

[38] Matej Kristan, Jiri Matas, Aleš Leonardis, Tomas Vojir, Roman Pflugfelder, Gustavo Fernandez, Georg Nebehay, Fatih Porikli, and Luka Čehovin. A novel performance evaluation methodology for single-target trackers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(11):2137–2155, Nov 2016. 3

[39] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1346–1353, 2012. 4

[40] Yong Jae Lee and Kristen Grauman. Predicting important objects for egocentric video summarization. *Int. J. Comput. Vision*, 114(1):38–55, aug 2015. 4

[41] Annan Li, Min Lin, Yi Wu, Ming-Hsuan Yang, and Shuicheng Yan. Nus-pro: A new visual tracking challenge. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):335–349, 2015. 3

[42] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4282–4291, 2019. 3

[43] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8971–8980, 2018. 3

[44] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. In *CVPR*, 2021. 4

[45] Pengpeng Liang, Erik Blasch, and Haibin Ling. Encoding color information for visual tracking: Algorithms and benchmark. *IEEE transactions on image processing*, 24(12):5630–5644, 2015. 3

[46] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 7

[47] Miao Liu, Siyu Tang, Yin Li, and James M. Rehg. Forecasting human object interaction: Joint prediction of motor attention and egocentric activity. *ArXiv*, abs/1911.10967, 2019. 4

[48] Cewu Lu, Renjie Liao, and Jiaya Jia. Personal object discovery in first-person videos. *IEEE Transactions on Image Processing*, 24(12):5789–5799, 2015. 4

[49] Zheng Lu and Kristen Grauman. Story-driven summarization for egocentric video. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2714–2721, 2013. 4

[50] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixe, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision (IJCV)*, 2020. 3

[51] Alan Lukežič, Luka Čehovin Zajc, Tomáš Vojíř, Jiří Matas, and Matej Kristan. Now you see me: evaluating performance in long-term visual tracking. *arXiv preprint arXiv:1804.07056*, 2018. 2, 7

[52] Christoph Mayer, Martin Danelljan, Goutam Bhat, Matthieu Paul, Danda Pani Paudel, Fisher Yu, and Luc Van Gool. Transforming model prediction for tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8731–8740, 2022. 3, 7

[53] Christoph Mayer, Martin Danelljan, Danda Pani Paudel, and Luc Van Gool. Learning target candidate association to keep track of what not to track. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13444–13454, 2021. 3

[54] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016. 3

[55] Abhinav Moudgil and Vineet Gandhi. Long-term visual object tracking benchmark. In *asian conference on computer vision*, pages 629–645. Springer, 2018. 2

[56] Matthias Mueller, Neil Smith, and Bernard Ghanem. A benchmark and simulator for uav tracking. In *European conference on computer vision*, pages 445–461. Springer, 2016. 3

[57] Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European conference on computer vision (ECCV)*, pages 300–317, 2018. 1, 3, 4

[58] S. Oh, J. Lee, N. Xu, and S. Kim. Video object segmentation using space-time memory networks. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9225–9234, Los Alamitos, CA, USA, nov 2019. IEEE Computer Society. 4

[59] Kenji Okuma, Ali Taleghani, De Freitas, J.J. Little, and David Lowe. A boosted particle filter: Multitarget detection and tracking. In *European Conference on Computer Vision*, 2004. 7

[60] Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2847–2854, 2012. 4

[61] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017. 4

[62] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 4

[63] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9339–9347, 2019. 1

[64] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9869–9878, 2020. 2

[65] Maximilian Speicher, Brian D Hall, and Michael Nebeling. What is mixed reality? In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–15, 2019. 1

[66] Yu-Chuan Su and Kristen Grauman. Detecting engagement in egocentric video. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 454–471, Cham, 2016. Springer International Publishing. 4

[67] Ran Tao, Efstratios Gavves, and Arnold WM Smeulders. Tracking for half an hour. *arXiv preprint arXiv:1711.10217*, 2017. 7

[68] Jack Valmadre, Luca Bertinetto, Joao F Henriques, Ran Tao, Andrea Vedaldi, Arnold WM Smeulders, Philip HS Torr, and Efstratios Gavves. Long-term tracking in the wild: A benchmark. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–685, 2018. 2, 3, 4

[69] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[70] Paul Voigtlaender, Jonathon Luiten, Philip HS Torr, and Bastian Leibe. Siam r-cnn: Visual tracking by re-detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6578–6588, 2020. 2, 3, 4, 7

[71] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. Fast online object tracking and segmentation: A unifying approach. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 1328–1338, 2019. 3

[72] Qianqian Wang, Yen-Yu Chang, Ruojin Cai, Zhengqi Li, Bharath Hariharan, Aleksander Holynski, and Noah Snavely. Tracking everything everywhere all at once. *arXiv preprint arXiv:2306.05422*, 2023. 9

[73] Weiyao Wang, Matt Feiszli, Heng Wang, Jitendra Malik, and Du Tran. Open-world instance segmentation: Exploiting pseudo ground truth from learned pairwise affinity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4422–4432, 2022. 7

[74] Weiyao Wang, Matt Feiszli, Heng Wang, and Du Tran. Unidentified video objects: A benchmark for dense, open-world segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10776–10785, 2021. 4

[75] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12692–12702, 2020. 4

[76] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2411–2418, 2013. 1, 3

[77] Fei Xie, Chunyu Wang, Guangting Wang, Yue Cao, Wankou Yang, and Wenjun Zeng. Correlation-aware deep tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8751–8760, 2022. 3

[78] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 603–619, 2018. 4

[79] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10448–10457, 2021. 3, 6, 7, 8

[80] Yipin Zhou and Tamara L. Berg. Temporal perception and prediction in ego-centric video. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4498–4506, 2015. 4

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Please see Supplementary section Limitations.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Please see Supplementary section Potential negative societal impacts. The societal impact of this work is consistent with that of the Ego4D.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Please see Supplementary Appendix A.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] Training/finetuning models is computationally costly to do multiple runs, similar to others in this field who do not report error bars. For all other baselines, we use their official checkpoints for inference only, so their results are deterministic.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 5.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] Our annotations are on top of Ego4D [28], which we discuss at length.
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Please see Supplementary Appendix A.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See Section 3.1; the videos in Ego4D [28] were collected by consenting participants.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See Section 3.1; by annotating Ego4D [28], we inherit its de-identification and screening for offensive content.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] We include detailed instructions in the supplementary. For proprietary reasons, we are not able to provide screenshots of the annotation tools.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] There are no human studies done in this paper. However, we build on Ego4D [28], which was collected with IRB approval.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] The participants were employed by a third-party vendor and are compensated based on the agreement with their employer.