How do humans and machine learning models track multiple objects through occlusion?

Anonymous Author(s)
Affiliation
Address
email

Abstract

Interacting with a complex environment often requires us to track multiple task-relevant objects not all of which are continually visible. The cognitive literature has focused on tracking a subset of visible identical abstract objects (e.g. circles), isolating the tracking component from its context in real-world experience. In the real world, object tracking is harder in that objects may not be continually visible and easier in that objects differ in appearance and so their recognition can rely on both remembered position and current appearance. Here we introduce a generalized task that combines tracking and recognition of valued objects that move in complex trajectories and frequently disappear behind occluders. Humans and models (from the computer-vision literature on object tracking) performed tasks varying widely in terms of the number of objects to be tracked, the number of distractors, the presence of an occluder, and the appearance similarity between targets and distractors. We replicated results from the human literature, including a deterioration of tracking performance with the number and similarity of targets and distractors. In addition, we find that increasing levels of occlusion reduce performance. All models tested here behaved in qualitatively different ways from human observers, showing superhuman performance for large numbers of targets, and subhuman performance under conditions of occlusion. Our framework will enable future studies to connect the human behavioral and engineering literatures, so as to test image-computable multiple-object-tracking models as models of human performance and to investigate how tracking and recognition interact under natural conditions of dynamic motion and occlusion.

1 Introduction

Human vision parses the visual scene into objects. Task-relevant objects are tracked and represented in a persistent fashion that enables us to mind them even when they are out of our field of view or hidden behind an occluding object. This ability liberates our visual cognition from its tethers in the sensory input and enables us to see the world in terms of its physical constituent components, providing a basis for prediction and causal insight. (1; 2).

An influential class of behavioral tasks that has enabled cognitive scientists to probe this ability is multiple object tracking (MOT) (3; 4). MOT tasks are typically highly abstracted. Subjects are presented with a set of identical abstract objects (e.g. circles) in motion and cued to track a subset of them. At the end of each trial, the scene freezes and the subject indicates the cued objects. MOT is challenging for humans and computational models (5; 6). Humans can track up to three or four objects successfully even through brief occlusions (7; 3; 4; 8). Using abstract identical objects has enabled cognitive scientists to study the function of tracking in isolation from the broader process of

Submitted to 4th Workshop on Shared Visual Representations in Human and Machine Visual Intelligence (SVRHM) at NeurIPS 2022. Do not distribute.

real-world dynamic scene perception, where objects are diverse and tracking interacts with visual recognition to maintain persistent representations of the task-relevant objects. In the real world, object tracking is harder in that objects may be out of sight or hidden for longer periods. Real-world object tracking might also be easier in that diverse objects tend to be visually distinct enough to track them by their appearance. The visual system might therefore employ a mixture of strategies, tracking objects not only by maintaining and updating (even for currently occluded objects) their spatial positions and velocities but also by recognizing them by their appearance (9; 10; 11).

Computer vision, with its focus on engineering applications, has engaged the MOT challenge in real-world settings (12; 13; 14), building systems that can track objects in videos (e.g. pedestrians or cars in traffic). While tracking objects with robustness to periods of occlusion poses the same computational challenge to humans and machines, it is unclear to what extent current models employ computational strategies similar to those used by humans. Recent ML challenges involve a large number of pedestrians (15) (up to ~ 250 pedestrians at the same time) widely exceeding the tracking capabilities of humans. On the other hand, supervised training on real-world images might bias the tracker towards less general solutions that might not reach human performance when required to generalize to unexpected challenges. In real-world scenarios, long stretches of occlusion of an object are rare compared to periods of at least partial visibility. Most tracked objects are detected with high confidence throughout their life-cycle (i.e., low noise regime), and objects are visually distinct enough to distinguish them by their appearance. Supervised ML object trackers might therefore be biased towards strategies that favor recognition in the current frame (spatial integration of evidence) over maintaining internal belief states about positions and velocities (temporal integration of evidence, as in a Kalman or particle filter). We might therefore expect these models to fail when similarity of appearance and/or longer stretches of occlusion render the recognition-based strategy underconstrained.

We here compare object tracking performance of four state-of-the-art supervised multiple object trackers with the performance of humans on the same task. We introduce a task combining object recognition demands and tracking designed to take steps toward bridging the gap between the real-world complexity of machine learning tasks and the abstraction of cognitive tasks. We replicated results from the human literature: a deterioration of tracking performance with the number and similarity of targets and distractors and a relatively unimpaired tracking performance through brief occlusion. We find that models displayed qualitatively different behavior from humans with their performance independent of the number of objects, but deteriorating below human performance in the context of occlusions.

2 Methods

2.1 Models

We evaluate four state-of-the-art multiple-object tracking algorithms. Simple online and real-time tracking (**SORT**, (16)) uses a simple strategy for frame-by-frame tracking of multiple objects. Bounding boxes of object detections in each frame, provided e.g., by a Faster Region CNN (17)), are the input to the tracker. On the first frame, the tracker initializes a series of tracklets, one for each detection, representing the objects to be tracked. Each tracklet's state (i.e., its bounding box center, height, aspect ratio, and their velocities) is tracked by a Kalman filter. In a new frame, the matches between all predicted bounding-box positions and all new detections are computed as intersection-over-union (IoU) to form an assignment cost matrix. The detections of the new frame are then associated with existing tracklets by minimzing the overall assignment cost using the Hungarian algorithm. Albeit a fairly simple algorithm, SORT achieved high performance on the MOT challenge (12; 18) while enabling real-time application on standard hardware. SORT performs well when state uncertainty is low. However, during occlusions, SORT displays a high number of identity switches (i.e, multiple tracklets in turn track the same object over time). **DeepSORT** (19) is an extension of the SORT algorithm that also takes appearance information (an appearance descriptor from a CNN) into account during the association stage, thereby compensating for the high state uncertainty of the Kalman filter during occlusions. DeepSort substantially reduced the number of tracklet switches. Tracking-by-detection approaches typically only use high-confidence detections of the object detector, leading to false negative detections for partially occluded objects, which in turn can lead to track termination and subsequent tracklet switches once the object reappears. **ByteTrack** (20) employs a strategy that also associates low-confidence bounding-box detections, reducing the number of id

Figure 1: a) Sequence of events (objects enlarged for visualization). b) Object sizes at the scale of Experiment 1. (c) Experimental factors occlusion level and similarity.

switches compared to SORT and DeepSort. **OC-SORT** (21) introduced several improvements over SORT/DeepSORT to better handle longer stretches of occlusion. One of the drawbacks of SORT is that the Kalman filter, which tracks the position and velocity of the bounding box, accumulates errors during stretches of occlusion. When the object reappears from occlusion it gets re-associated with its (memorized) track using appearance-based models. The Kalman filter belief about the objects position and velocity then needs to be reconciled with the newly detected position of the object. OC-SORT addresses this problem by implementing several strategies that give detections a much larger weight compared to the Kalman state in the association stage.

2.2 Task

The details of the object-tracking task were the same for humans and models unless stated otherwise (Figure 1a). At the beginning of each trial N objects were presented. Phase 1 is the *cueing period* (Experiment 1: 1s, Experiment 2: 2.5s) in which the objects are stationary. For human subjects the target objects are highlighted in this period. Phase 2 is the *motion period* (five seconds here) in which the objects move according to random trajectories. We deliberately used a complex non-linear motion model rather than a linear motion model. In particular, angular motion direction perturbation were sampled at random time-points and the new direction was smoothly interpolated across multiple time-frames (intersection over union of an object's bounding boxes in adjacent time frames: 0.691, s.d.$= 0.024$ for Experiment 1 and 0.594, s.d.$= 0.032$ for Experiment 2). The objects then freeze for phase 3, the *response period*, in which the target objects are reported. Human subjects reported the targets by clicking on the objects. In occlusion trials, a large central opaque occluder covering either 20% or 40% of the field was superimposed to the scene. Stimuli were isolated objects extracted from the segmentation masks provided by the MS COCO challenge (18). For each of the 80 object categories, we extracted 50 different exemplars, excluding fragmented objects. Object starting positions were sampled such that (1) all objects were unoccluded at the start of the trial and (2) all objects ended up in positions such that no two were closer to each other than a threshold distance. Objects which stopped their movement behind the occluder were revealed by turning the occluder semi-transparent (for humans) or moving their depth plane in front of the occluder (for models, which were not assumed to be familiar with semi-transparency) .

Our goal was to compare how well humans and models track multiple objects through occlusion. Human short-term memory for dynamic bindings, e.g., the binding between a visual object and a label or id, is highly limited. Rather than assigning labels or ids to objects in the cueing period and asking participants to reproduce those labels/ids during the response period, we decided to denote a subset of $T < N$ objects as "targets" that need to be tracked throughout the motion period (3; 22). To obtain tracking performance, we employed different strategies for humans and models.

2.2.1 Human behavior

For humans, target objects were visually cued during the cueing period via a glowing outline and by overlaying an icon of a coin on top of each target object. During the response period, human participants clicked on T objects, which they believed to be the initially cued target objects, after which the identity of the selected objects was revealed by an animation that displayed and "collected" the coins of the target objects. There was no time-limit for human responses.

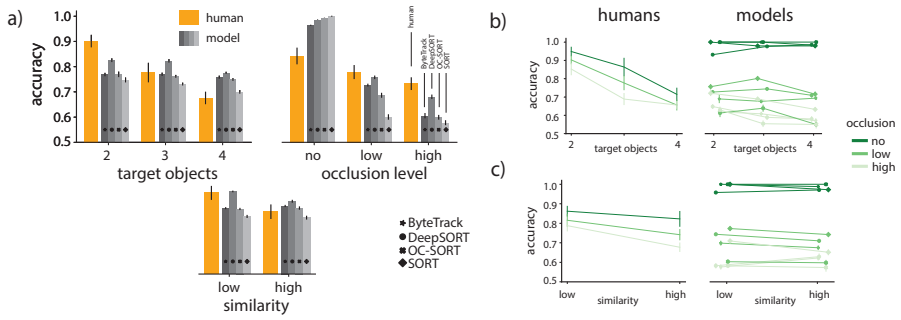

Figure 2: Results of Experiment 1. Response accuracy main effects (a) and selected interactions (b) for the factors of number of tracked objects, occlusion level, and object similarity. Error bar indicates standard error of the mean across participants (humans) or across repeated evaluation runs (models).

2.2.2 Model behavior

A major factor in object tracking performance is the quality of the object detector. We here focus on the tracking task rather than the ability to detect objects. Therefore, instead of relying on each model's detector, we provide models with ground truth bounding boxes for unoccluded objects in the current frame. This is akin to the public detections variant of the MOT challenge (12). Note that all models still used image input for appearance-based associations. We also provide the bounding box of the occluder as a detection, indicating to models that missing detections might be due to object-by-object occlusion rather than signal dropout.

To obtain responses from models, we analysed whether the model consistently tracked target objects. First, we associated the ground-truth object positions with the model's tracklet positions (i.e., the model's belief about object positions) using the Hungarian algorithm. Second, we identified the track id that the model assigned to the cued (target) objects on the first frame. A target object was considered to be successfully tracked if its associated track id was identical for the first frame and the last frame of the motion period. Model "responses" were all successfully tracked objects and a random choice of the remaining objects in case less than T objects were successfully tracked (i.e., guessing).

2.3 Experiments

In Experiment 1, we presented 108 sequences with the following experimental factors: We varied the number of objects (4, 6, or 8). Half of the objects were targets in each case. A third of the trials contained no occluder, the remaining trials contained a rectangular occluder that extended over the full vertical (horizontal) length of the screen and either covered 20% or 40% of the horizontal (vertical) length (Figure 1c). Objects within a trial could either all come from the same category (high similarity) or from different categories (low similarity) (Figure 1c). The factors, namely number of objects, occlusion level, occluder orientation (horizontal or vertical), and similarity were counter-balanced across trials. The maximum bounding box side length of an object (height or width) was 10% of the screen (Figure 1b).

Experiment 2 was identical to Experiment 1, with the following exceptions: First, there were four set size conditions: 6 objects - 2 targets, 8 objects - 4 targets, 12 objects - 2 targets, 14 objects - 4 targets. We also reduced the size of the objects to a maximum bounding box side length of 7%. Because of the higher visual load, we extended the cueing period to 2.5s to provide human observers with a sufficient amount of time to identify the cued objects.

3 Results

In Experiment 1, we replicated the set-size effect in humans. Namely, tracking more targets led to lower response accuracy (main effect set-size $F(2, 18) = 52.5856$, $p < 0.001$). Models, in contrast,

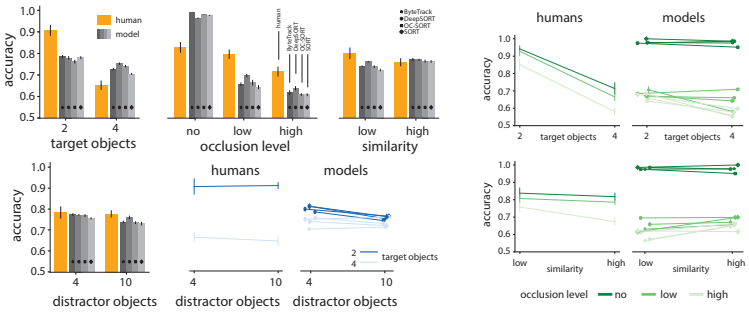

Figure 3: Results of Experiment 2.

showed a much smaller, almost absent set-size effect (Figure 2a) ($F(2, 6) = 9.4250$, $p = 0.0141$)[1]. Human observers and models both showed lower performance with higher occlusion levels (Figure 2a, humans: $F(2, 18) = 24.8148$, $p < 0.001$, models: $F(2, 6) = 87.2780$, $p < 0.001$). Models showed super-human performance during no occlusion trials, even at a small set size of four (i.e., two target objects) for which humans performed best. Strikingly, humans were better able to track objects trough occlusion than any of the SOTA object trackers.

How do humans track multiple objects, in particular for complex non-linear object motion? A first indication can be gleaned from the effect of object similarity. In particular, increased similarity between objects reduced tracking performance for humans and model (humans: $F(1, 9) = 31.2805$, $p < 0.001$, models: $F(1, 3) = 0.5374$, $p = 0.5166$). This suggests that humans performance at least partially depended on appearance information.

In order to replicate and extend our findings, we conducted a second experiment in which we increased the total number of objects. First, we replicated all key findings from experiment 1 (Figure 3a). Second, from the results in experiment 1 alone it was unclear, whether the set-size effect in humans was caused by the tracking demands or by the visual load of all objects (targets and distractors). Figure 3b shows that performance was virtually unaffected by the number of distractors in humans (main effect number of distractor objects: $F(1, 9) = 0.0601$, $p = 0.8119$) and models ($F(1, 3) = 25.4859$, $p = 0.015$). A larger number of distractor objects is irrelevant to tracking performance in humans as well as models.

4 Discussion and Conclusion

We compared state of the art multiple object tracking models and humans in their ability to track multiple objects through occlusion. We find that all models behaved in qualitatively different ways from human observers. First, without occlusion, models do not display the same capacity limitations as humans, showing superhuman performance for large numbers of targets. However, models show subhuman performance when faced with longer stretches of occlusion. These findings raise several interesting questions and future research directions suggesting interesting combinations of normative and approximate approaches.

We decoupled the tracking performance of the evaluated models from their ability to detect objects reliably by providing the same defections to all tracking models. Most models still contained components (e.g., the deep re-identification components) pre-trained on other datasets. Our stimuli (crops from MS Coco) were chosen to fall within the training dataset distribution. However, fine-tuning models on our dataset might yield slightly more favorable performance for models.

The experimental framework introduced here promises to connect the engineering and the cognitive science of MOT and to add elements missing from both literatures: namely longer occlusions and complex motion trajectories as well as a continuum of controlled and naturalistic task variants that probe how the tracking and recognition cooperate to meet the challenge of object tracking.

[1]While showing some differences in their performance, object tracking models perform comparably similar. This is why we refrained from interpreting small differences between models. Instead, we treat the models as a sample from the distribution of supervised trackers trained on ML benchmarks.

References

[1] B. Peters and N. Kriegeskorte, "Capturing the objects of vision with neural networks," *Nature Human Behaviour*, vol. 5, no. 9, pp. 1127–1144, Sep. 2021, number: 9 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41562-021-01194-6

[2] K. Greff, S. van Steenkiste, and J. Schmidhuber, "On the Binding Problem in Artificial Neural Networks," *arXiv:2012.05208 [cs]*, Dec. 2020, arXiv: 2012.05208. [Online]. Available: http://arxiv.org/abs/2012.05208

[3] Z. W. Pylyshyn and R. W. Storm, "Tracking multiple independent targets: Evidence for a parallel tracking mechanism," *Spatial Vision*, vol. 3, no. 3, pp. 179–197, 1988. [Online]. Available: https://brill.com/view/journals/sv/3/3/article-p179_3.xml

[4] B. J. Scholl and Z. W. Pylyshyn, "Tracking Multiple Items Through Occlusion: Clues to Visual Objecthood," *Cognitive Psychology*, vol. 38, no. 2, pp. 259–290, 1999, iSBN: 0010-0285 (Print).

[5] D. Linsley, G. Malik, J. Kim, L. N. Govindarajan, E. Mingolla, and T. Serre, "Tracking Without Re-recognition in Humans and Machines," *arXiv:2105.13351 [cs]*, Jun. 2021, arXiv: 2105.13351. [Online]. Available: http://arxiv.org/abs/2105.13351

[6] E. Vul, G. Alvarez, J. B. Tenenbaum, and M. J. Black, "Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model," in *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, Eds. Curran Associates, Inc., 2009, pp. 1955–1963. [Online]. Available: http://papers.nips.cc/paper/3828-explaining-human-multiple-object-tracking-as-resource-constrained-approximate-inference-in-a-dynamic-probabilis pdf

[7] J. Intriligator and P. Cavanagh, "The spatial resolution of visual attention," *Cognitive Psychology*, vol. 43, no. 3, pp. 171–216, 2001, iSBN: 0010-0285. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0010028501907558

[8] S. Yantis, "Multielement visual tracking: Attention and perceptual organization," *Cognitive Psychology*, vol. 24, no. 3, pp. 295–340, Jul. 1992. [Online]. Available: http://www.sciencedirect.com/science/article/pii/001002859290010Y

[9] F. Papenmeier, H. S. Meyerhoff, G. Jahn, and M. Huff, "Tracking by location and features: Object correspondence across spatiotemporal discontinuities during multiple object tracking," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 40, no. 1, pp. 159–171, 2014.

[10] K. Zhou, H. Luo, T. Zhou, Y. Zhuo, and L. Chen, "Topological change disturbs object continuity in attentive tracking," *Proceedings of the National Academy of Sciences*, vol. 107, no. 50, pp. 21 920–21 924, Dec. 2010, publisher: National Academy of Sciences Section: Biological Sciences. [Online]. Available: https://www.pnas.org/content/107/50/21920

[11] J. Li, L. Oksama, and J. Hyönä, "Model of Multiple Identity Tracking (MOMIT) 2.0: Resolving the serial vs. parallel controversy in tracking," *Cognition*, vol. 182, pp. 260–274, Jan. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0010027718302749

[12] A. Milan, L. Leal-Taixe, I. Reid, S. Roth, and K. Schindler, "MOT16: A Benchmark for Multi-Object Tracking," *arXiv:1603.00831 [cs]*, May 2016, arXiv: 1603.00831. [Online]. Available: http://arxiv.org/abs/1603.00831

[13] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[14] M. Müller, A. Bibi, S. Giancola, S. Alsubaihi, and B. Ghanem, "TrackingNet: A Large-Scale Dataset and Benchmark for Object Tracking in the Wild," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, vol. 11205, pp. 310–327, series Title: Lecture Notes in Computer Science. [Online]. Available: http://link.springer.com/10.1007/978-3-030-01246-5_19

[15] P. Dendorfer, H. Rezatofighi, A. Milan, J. Shi, D. Cremers, I. Reid, S. Roth, K. Schindler, and L. Leal-Taixé, "MOT20: A benchmark for multi object tracking in crowded scenes," Mar. 2020, arXiv:2003.09003 [cs]. [Online]. Available: http://arxiv.org/abs/2003.09003

[16] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 3464–3468.

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 91–99. [Online]. Available: http://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks.pdf

[18] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *Computer Vision – ECCV 2014*, ser. Lecture Notes in Computer Science, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 740–755.

[19] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE international conference on image processing (ICIP)*. IEEE, 2017, pp. 3645–3649.

[20] Y. Zhang, P. Sun, Y. Jiang, D. Yu, Z. Yuan, P. Luo, W. Liu, and X. Wang, "ByteTrack: Multi-Object Tracking by Associating Every Detection Box," in *arXiv preprint arXiv:2110.06864*, 2021.

[21] J. Cao, X. Weng, R. Khirodkar, J. Pang, and K. Kitani, "Observation-Centric SORT: Rethinking SORT for Robust Multi-Object Tracking," *arXiv preprint arXiv:2203.14360*, 2022.

[22] C. Merkel, J.-M. Hopf, and M. A. Schoenfeld, "How to Perceive Object Permanence in Our Visual Environment: The Multiple Object Tracking Paradigm," pp. 1–20, 2019. [Online]. Available: https://link.springer.com/protocol/10.1007/7657_2019_28

[23] M. Contributors, "MMTracking: OpenMMLab video perception toolbox and benchmark," 2020. [Online]. Available: https://github.com/open-mmlab/mmtracking

[24] B. Peters, M. Retchin, and N. Kriegeskorte, "Flying Objects: Challenging humans and machines in dynamic object vision," in *2022 Conference on Cognitive Computational Neuroscience*. San Francisco: Cognitive Computational Neuroscience, 2022. [Online]. Available: https://2022.ccneuro.org/view_paper.php?PaperNum=1301

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] The limitations of our approach are described in the discussion section (4)
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] We do not forsee any negative societal impacts of our work.
   (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...

(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)? [No] We provided detailed descriptions allowing for replication of our results. Data and analysis code will be made freely available upon publication of the final, full-length report of this study. The task will be made freely available as a toolbox together with a separate final, full-length publication.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A] Pre-trained models were used without fine-tuning. Pre-trained weights and training details are available with the papers and repositories that we cite.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We do report error bars. Error bars for human observers represent the standard error of the mean (SEM) across participants. Error bars for models represent the SEM over repeated runs of the same model.

(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)? [N/A] As we use models only in inference mode and do not train models, the total amount of compute is negligible and all analyses can be performed in a couple of hours on a standard desktop machine.

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]

(b) Did you mention the license of the assets? [N/A] Assets and their license information are publicly available via the cited asset links.

(c) Did you include any new assets either in the supplemental material or as a URL? [No] Data and analysis code will be made freely available upon publication of the final, full-length report of this study. The task will be made freely available as a toolbox together with a separate final, full-length publication describing the toolbox.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] Informed consent was obtained from human participants following the procedure approved by the appropriate IRB (see B.1 for details).

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] We state that there is no identifiable information in the data collected in the appendix (B.1). The study involves no materials with possibly 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 provide the screenshot of the instruction given to participants in the appendix (Figure 4)

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [Yes] There were no potential participant risks. The IRB approval will be included in the non-anonymized version of the manuscript.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes] We provide details on compensation in the Appendix (B.2)

A Details on the models

We evaluated tracking performance for SORT (16), DeepSORT (19), ByteTrack (20), and OC-SORT (21) using pre-trained weights via the MMTracking toolbox (24). We evaluated tracking performance for all models using videos generated by the Unity framework (30 fps) (23).

B Data crowdsourcing

B.1 Recruitment, procedure, and data

Participants were recruited via Prolific (ten new participants for each experiment) and the experiment was delivered via a browser app implemented in Unity (23). The study was advertised as a brief

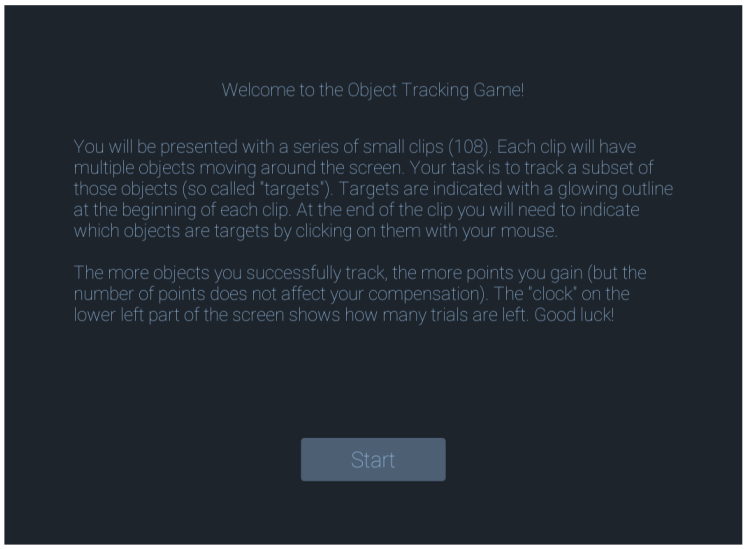

Figure 4: **Instruction screen**

online study on Prolific (duration: 30 minutes). Upon acceptance of the experiment, participants gave informed consent (screenshot will be included in the non-anonymized version of the manuscript) and completed a brief demographic questionnaire in accordance following the procedure approved by the appropriate local IRB. The experiment then started by presenting an instruction screen (see Figure 4). After completing the experiment, participant were given a completion code with which they could verify experiment completion with Prolific and obtain their compensation. The collected data and responses on the task are non-identifiable in accordance with the IRB.

B.2 Human participant compensation

The estimated hourly wage paid to the participants was $12. We spent a total amount of $192 for participant compensation.