# Rethinking Generalization in Embodied Perception

Aljoša Ošep

July, 2023

"Alexa, please take this food delivery package to Smith Hall. Once you reach the front entrance, go upstairs, turn left. Go all the way down the hall through the door to Room 200 (across from the elevator) to drop off the package. If you cannot find 4802 Forbes Ave, drive down the small road between Hamburg Hall (4800 Forbes) and the Forbes bridge, and pull into the turnaround at the end." — *Me, hopefully soon!*

"*I'm sorry, I cannot assist with physical tasks or navigation in the real world. I'm here for any other questions!*" is a response of state-of-the-art language models to the inquiry above. Not surprisingly – to act in the real world, *embodied* agents require actuators that allow them to move and manipulate objects and visual sensors that act as the agents' eyes, capturing crucial visual information from the environment. The potential for embodied assistants to make this world a better place is vast, from automating transportation and aiding people with disabilities to environmental tasks like street and ocean cleaning. However, the current bottleneck is neither in hardware nor in the capability to process natural language [4]. The critical challenge lies in equipping these agents with the ability to *understand* the world through their sensors to enable *safe* and effective interactions with the world *and* to recognize limitations of their understanding whenever it is not possible to guarantee accurate interpretation of the visual input.

**Do not generalize. *Memorize*.** One cannot navigate this world based on the sensory input alone. Visuals, observed in real-time, do not directly convey where we are, what sequence of steps takes us to a desired end-point, or what happens if that chicken does, in fact, decide to cross the road. It is our past experience, in conjunction with readings we receive from the world, that enable us to accomplish our tasks, such as grasping an apple, predicting the future trajectory of a vehicle with a blinking turn signal, and understanding that we should *not* attempt to enter a Gravity-Defying Loop Junction based on our prior experience with roundabouts and cross-junctions, because we have never seen one and we are unlikely to navigate in there safely.[1]

**Tracking leads to memorization.** Contrary to the common pursuit of generalization for autonomous operation in unseen environments, we take the provocative stance that learning-enabled agents, when used for embodied control in safety-critical deployment, should *avoid generalization* and operate only in familiar scenarios. An embodied agent's interpretation of the sensory data should, therefore, be grounded in an internal memory of an agent, as opposed to attempting to "generalize by luck". This suggests that the burden of safety should be put on the thoroughness of the dataset used to train the model (*i.e.*, do not attempt to generalize your Ann Arbor-trained model to deploy an autonomous agent in London) and, crucially, the capability to effectively *memorize* the training data. However, our training data will not consist of a large corpus of text or image collections, but rather *hours, days, months* or even *years* of video, that will *continually* expand to widen the area of safe operation. How can we memorize critical aspects of such video to "ground" our observations to past experiences?

**Towards life-long tracking.** We posit that video tracking over *extremely* long periods of time may be the *missing key* in learning internal models of the (observed) world directly from the video. In Sec. 2, we discuss *why* tracking is a fundamental capability needed to *compress* raw video footage into a compact visual representation that can be used to understand and navigate in an environment captured in this footage and to recognize when our observations exit the "bounds" of the world covered in our training data. In Sec. 1, we discuss our past contributions in the context of video understanding in embodied perception.

[1] Yes, even if your language model has a pretty good idea of what to do in this situation.

# 1 Past and Related Work

This section briefly summarizes our past contributions to the field of embodied perception, where learning from video and, more broadly, multi-modal streams of sensory data plays a central role.

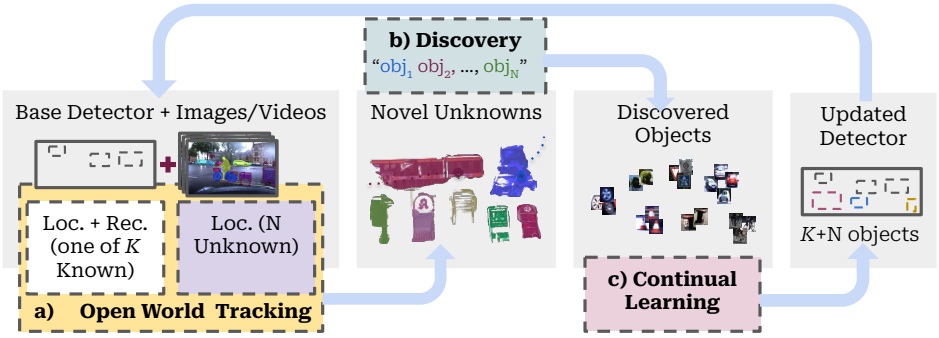


Figure 1: **Learning to detect objects from video.** Given *base* object detector, trained to localize $K$ object classes, we first localize both *known* and *unknown* (*i.e.*, those not recognized by the base detector) objects in video or images (**Open World Tracking**, tackled in [38, 40, 29]). Next, we learn to group *unknown* objects in the video (**Video Object Discovery**, tackled in [39, 12]). Finally, we re-train the base object detector to recognize discovered, *unknown* object classes (**Continual Learning**, as tackled in [36, 12]).

**What is around me?** This is the central question in embodied perception, commonly tackled via (supervised) object detection [54, 11, 47, 46]. Our past work investigates how we can learn to detect objects we observe in streams of sensory data without exhaustive human supervision (in Fig. 1, we outline the overall approach that we investigated in a series of publications [38, 39, 40, 29, 12]). In this context, object tracking emerged as a crucial component (Fig. 1a) that allowed us to compress hours of self-driving video into a compact set of (4D) video-object proposals, building upon our efforts towards *any* object tracking in stereo video [38, 40].

In our *ICRA'19* paper [39], we demonstrated we can *cluster* objects, localized in video, by transferring knowledge from a (labeled) image domain to the (unlabeled) video, and discover novel object classes (Fig. 1b). In [36] use clusters of discovered objects as pseudo-labels to train detectors for novel classes (Fig. 1c), where cluster assignments provide supervision for object categorization. We have consolidated this intuition in our recent *NeurIPS'22* paper [12], where we train object detectors in an end-to-end manner in unlabeled images (Fig. 1b–c). Our method consists of a two-stage detection network [47] that alternates between region proposal pseudo-labeling and updating the network weights given the pseudo-labels. After the initial supervised training phase (using labels for 80 classes [27]), our method learns to detect over 1200 classes, labeled in LVIS [16] dataset. This line of work follows our general philosophy of learning rich representations from hours of video recorded from an embodied agent perspective. The community followed this path in the context of learning to detect *moving* objects in RGB-D [17] and Lidar [68, 34] sequences. Several recent related efforts focus on the zero-shot (open-vocabulary) classification of detected objects by connecting visual features with language models [45].

**How does it move?** Beyond the role of Multi-Object Tracking (MOT, Fig. 2a) in learning to detect objects in video sequences (Fig. 1), object tracking is pivotal in real-time dynamic situational awareness for embodied navigation [14, 7, 67, 49, 5], as well as in studying animal behavior [41] and monitoring biological phenomena [1].

*3D localization matters.* In our *ICRA'17* publication [37], we proposed a Combined Image- and World-Space Tracker (CIWT) that lifts monocular object detections to 3D space to reason object trajectories jointly, in image domain and 3D space, as needed in embodied navigation. CIWT was the first 3D MOT entry on the popular KITTI Tracking Benchmark [14] and has sparked several follow-up methods in 3D MOT in Lidar [58, 59, 66] and video [31, 19]. Our recent *NeurIPS'22* paper [8] consolidates the importance of reasoning about object trajectories in 3D space and suggests that the key feature for improving long-term tracking performance (*i.e.*, in reducing long occlusion gaps) is to estimate multiple plausible long-term trajectory continuations.

*Tracking every point and pixel.* Our line of work in class-agnostic MOT (*ICRA'18* [38]), where

(a) MOT (b) MOTS (c) STEP (d) 4D Segmentation

Figure 2: Our progression from (2D) multi-object tracking (MOT) (2a) towards unifying object tracking and segmentation (MOTS) (2b) in video, (2c) and 4D segmentation in Lidar data (2d).

we aim at tracking *any* object, led us to conclusions that pixel-precise object localization (*i.e.*, segmentation) significantly aids with the disambiguation of object identities over time – the core challenge in MOT. We consolidated this intuition in the context of closed-set object tracking in our *CVPR'19* paper [55], where we introduce the first dataset and end-to-end network for Multi-Object Tracking and Segmentation (MOTS, Fig. 2b). MOTS has quickly become a vibrant field of research and has inspired the community to develop new methods [44, 63, 6] and diverse datasets (*e.g.*, video instance segmentation [64], BDD100k-MOTS [67]). We take this line of research a step further in STEP (Segmenting and Tracking Every Pixel, *NeurIPS'21* [57]), Fig. 2c, where we require assigning semantic classes and track identities to *all* pixels in a video, or *all* points in a Lidar sequence (*4D Panoptic Lidar Segmentation*, *CVPR'21* [3], Fig. 2d). This task was recently adopted by the large-scale NuScenes dataset [5, 13] and has been tackled in several exciting follow-up papers [24, 69, 33, 65].

*Tracking every object.* Our work on MOT, MOTS, and STEP (Fig. 2) tackle pixel-precise tracking of objects that correspond to a pre-defined (closed) vocabulary of *known* objects for which labeled data is available. Towards our long-term vision of learning to model new objects from video modality continually (Fig. 1), we need the ability to localize *any* object in videos. Our early efforts (*ICRA'18* [38], *ICRA'20* [40]) rely on early data-driven object proposal generation and segmentation methods [42, 43] in conjunction with depth and motion cues to localize objects in 4D space-time, derived from stereo video. Our recent *CVPR'22* publication consolidates this research direction as Open World Tracking (OWT), *i.e.*, *any* object tracking in the open world (Fig. 3), outlining benchmark methodology and baselines derived from our early work on stereo-based *any* object tracking [38, 40]. We are thrilled that our work has sparked the community's interest to suggest alternative evaluation metrics [25, 2], extended OWT with open-vocabulary classification [26, 35] and improved tracking performance using recent foundation models for class-agnostic segmentation [22].

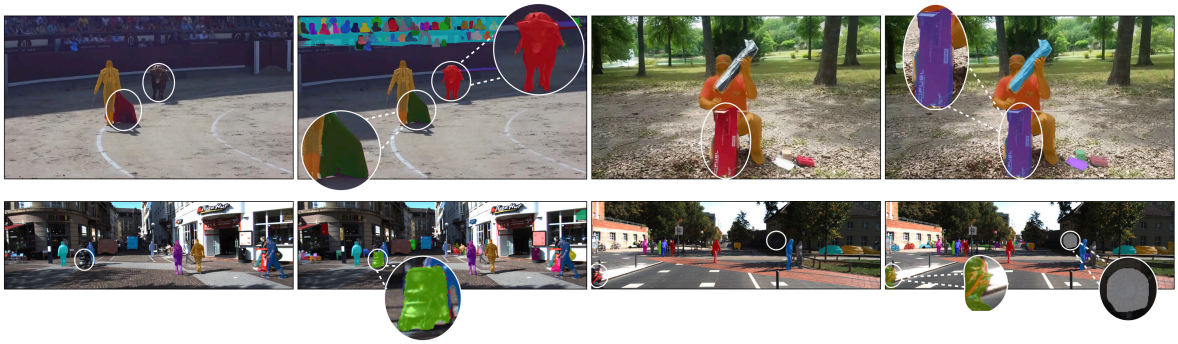

Figure 3: *Each pair left:* The standard approach to MOT(S) is to detect and track objects that correspond to specific, pre-defined semantic classes, such as cars and pedestrians [55]. *Each pair right:* In OWT, methods must track objects that may not labeled in the training set, freeing us from the tedious task of *having to label every possible object category in the world*.

**Where am I, and where am I going?** Localization and mapping are crucial for embodied navigation. Our *CVPR'22* paper [23] is the first investigation in natural language-based localization: given a 3D map of the environment our *Text2Pos* localizes an arbitrary location on a city scale based on a natural-language description of the visual surroundings. Building upon our dataset and baseline, recent community efforts [56, 62] significantly improved the localization accuracy. Our recent work [60], presented at *CVPR'23*, suggests that we can map our environment from monocular videos via data-driven, cross-modal retrieval.

## 2 Research Philosophy and Future Work

We will work towards *safety*-centered neural architectures for visual dynamic scene understanding. From our point of view, autonomous agents should not attempt to generalize and *hallucinate* how a never-observed-before object will move, what action to take in an unfamiliar junction, or which path to take in an entirely unknown environment. Instead, we posit that an embodied agent should operate within the bounds of the visual world, captured explicitly in our training data. Our strategy is influenced by the concept of "Explanation by Example" [28], a principle based on the idea that observed instances should be relatable to prior observations, as used, for example, in the legal profession to justify an action based on prior *legal precedent*. This principle guides our approach of leveraging past examples to interpret new, unseen data.

**The data.** Our research endeavors necessitate training data in the form of unlabeled video sequences that capture glimpses of a persistent, dynamic environment. Our "Explanation by Example" based approach necessitates models that compress hours, days, months, or, eventually, years of such video into a compact internal representation that embodied agents can query and use to *ground* online observations to past visual experience. Toward this goal, we plan to utilize real-world footage and virtual worlds. For synthesizing data, we will build upon our prior work [10], however, by contrast, we will focus on modeling a persistent virtual world and synthesize videos from different characters' perspectives over extended time periods. In the context of real-world footage, we will utilize video such as *full*-length movies and TV shows (by contrast to prior works that focus on understanding short movie clips [48, 20, 51, 53, 15]), as well video recordings captured from a mobile platform, such as OxfordRobotCar dataset [32]. OxfordRobotCar contains over $20h$ of unlabeled driving video recorded Oxford over the span of two years.

**Attending your memory.** State-of-the-art computer vision models [9, 30, 18] utilize self-attention mechanism [52] to capture global relationships among visual elements in the data, which limits their applicability to short temporal windows due to the quadratic complexity of the attention operation w.r.t. number of visual tokens. By contrast, our research goals necessitate attending to *extremely* long sequences of visual memory and, therefore, rethinking existing paradigms tailored towards learning representations from single images. A promising direction was recently outlined in the field of natural language modeling that suggests retrieving a fixed set of the $k$ most relevant queries [61] from offline storage to effectively "expand" the memory. Such an approach may already be applicable for grounding image recognition models to external memory, where $N$ visual tokens extracted from training examples would act as a database index. Pilot experiments suggest similarity scores between such an index and visual inputs can be computed in milliseconds for $N \approx 100M$ via fast nearest-neighbor libraries such as FAISS [21]. This suggests that brute-force inference may be possible for moderately large training sets (such as the $400M$ examples used to train vision-language foundation models such as CLIP [45]). However, such exhaustive evaluation may still be slow for a dataset with truly massive $N$. For example, a training dataset of $1K$ minute-long videos sampled at $10$ frames/sec composed of $1024 \times 1024$ images split into patches of size $32 \times 32$ pixels would generate $614B$ tokens.

**Video object tokenization.** A possible path forward would be to condense extensive video collections into a *visual memory* explicitly in an offline phase. Such an approach would involve (i) initially segmenting videos into object tracks using foundation models for image segmentation [22] in conjunction with our prior work on *any* object tracking [29], followed by (ii) encoding these segments as visual tokens across the entire dataset. Representing each video sequence as a set of instance-level tokens reduces the number of tokens by $1000x$ (by assuming any one-minute-long video contains at most $1000$ objects). Ultimately, we will store these exemplars as key-value pairs, which can be retrieved by the model online. The safeguarding model can then attend over the (compressed) datastore to compute the final representation, which will be used to ascertain if the observed instance can be grounded to our past observations, condensed in the VM.

**Track to memorize.** We argued that tracking may be central for the construction and *compression* of visual memory from raw video. However, ideally, memory construction should not be manual, but implicit, and learned in an end-to-end manner. We argue, that such visual memory may automatically emerge as a by-product of ***life-long tracking (LLT)***, which we pose as a task of *tracking the identities of all pixels at all times* over a large time periods.

**Challenges and emerging properties.** LLT is a challenging problem, as the set of (dense) plausible correspondences grows exponentially with the length of the video (*i.e.*, the number of video frames), while establishing dense correspondences is challenging due to the self-similarity of structures and poorly-textured regions. We hypothesize that due to the aforementioned challenges, along with limited computational and memory resources, we can expect the following capabilities to emerge in the process.

*Pixel and semantic grouping (*i.e.*, segmentation&categorization).* The number of correspondences in a video containing $T$ frames grows exponentially, in the order of $O(w^T \times h^T)$, where $w \times h$ denotes the image resolution. Even when considering only short-term tracking (*e.g.*, two consecutive frames), there are $O(w^2 \times h^2)$ possible correspondences that can be reduced significantly by quantizing consecutive images to $N$ and $M$ regions. While this simplifies the correspondence search to $N \times M$ plausible associations across two consecutive frames, it can be further reduced if we constrain the region-wise correspondence search to semantically related concepts, determined based on appearance similarity. *Image segmentation and appearance-based object categorization may be an emergent phenomenon of LLT.*

*Visual memory.* While image quantization and categorization may be pivotal in short-term tracking given constrained resources, the problem of combinatorial complexity w.r.t. number of frames $T$ persists. We hypothesize that successfully addressing LLT neccissitates summarizing past visual observations in visual memory. Short-term memory may help us to keep track of identities to bridge occlusions, while re-identifications of pixels, objects, or regions after substantial temporal gaps necessitate long-term memory. As this memory is bounded, models will require a "forgetting" mechanism, implying that obtaining perfect accuracy in *life-long* tracking would require infinite memory. Therefore, the forgetting mechanism must operate in such a way as to maximize the final model performance under memory constraints.

*Place recognition, localization, and mapping [50].* We do not expect all objects at all places at all times – locations of objects and regions are commonly geographically constrained. Recognizing a *particular* constellations of objects and regions (place recognition) should further simplify the problem of LLT. We do not assume that such a memory will store metric-precise 3D maps of our surroundings, but rather into an *internal* representation that enables agents to navigate and perform tasks in an environment captured in video sequences. *Place recognition and mapping may emerge from memory organization in the process of LLT.*

**Vision for the future.** Our ultimate goal is to create a paradigm shift in how embodied agents perceive and interact with their environment. We envision a future where these agents, equipped with advanced visual perception capabilities, can safely and efficiently navigate and operate in complex, dynamic environments. Our research plan is based on the philosophy that sensory interpretations and understanding of *how* an object can be used to accomplish a task or may interfere with our trajectory must be grounded in past experiences that will be condensed in visual memory. We will work on a family of safety-centric perception models that summarize their past sensory experiences in visual memory, constructed *explicitly*, or may emerge implicitly in the process of *life-long* tracking. Embodied agents should always relate to these past experiences when interpreting sensory inputs and not attempt to act when encountering unfamiliar situations, as this could compromise their, and, importantly, *our* safety.

# References


[1] S. Anjum and D. Gurari. CTMC: Cell tracking with mitosis detection dataset challenge. In *CVPR Workshops*, 2020.

[2] A. Athar, J. Luiten, P. Voigtlaender, T. Khurana, A. Dave, B. Leibe, and D. Ramanan. Burst: A benchmark for unifying object recognition, segmentation and tracking in video. In *IEEE Winter Conf. on Applications of Computer Vision*, 2023.

[3] M. Aygün, A. Ošep, M. Weber, M. Maximov, C. Stachniss, J. Behley, and L. Leal-Taixé. 4d panoptic lidar segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark,

C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[5] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[6] A. Choudhuri, G. Chowdhary, and A. G. Schwing. Assignment-space-based multi-object tracking and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[7] P. Dendorfer, A. Ošep, A. Milan, K. Schindler, D. Cremers, I. Reid, and S. R. L. Leal-Taixé. Motchallenge: A benchmark for single-camera multiple target tracking. *International Journal of Computer Vision (IJCV)*, 2020.

[8] P. Dendorfer, V. Yugay, A. Ošep, and L. Leal-Taixé. Quo vadis: Is trajectory forecasting the key towards long-term multi-object tracking? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[10] M. Fabbri, G. Brasó, G. Maugeri, O. Cetintas, R. Gasparini, A. Ošep, S. Calderara, L. Leal-Taixé, and R. Cucchiara. MOTSynth: How can synthetic data help pedestrian detection and tracking? In *IEEE International Conference on Computer Vision*, 2021.

[11] P. Felzenszwalb, D. Mcallester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[12] V. Fomenko, I. Elezi, D. Ramanan, L. Leal-Taixé, and A. Ošep. Learning to discover and detect objects. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[13] W. K. Fong, R. Mohan, J. V. Hurtado, L. Zhou, H. Caesar, O. Beijbom, and A. Valada. Panoptic nuscenes: A large-scale benchmark for lidar panoptic segmentation and tracking. *arXiv preprint arXiv:2109.03805*, 2021.

[14] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[15] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[16] A. Gupta, P. Dollar, and R. Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[17] A. W. Harley, Y. Zuo, J. Wen, A. Mangal, S. Potdar, R. Chaudhry, and K. Fragkiadaki. Track, check, repeat: An em approach to unsupervised tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[18] A. Hassani, S. Walton, J. Li, S. Li, and H. Shi. Neighborhood attention transformer. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023.

[19] H.-N. Hu, Q.-Z. Cai, D. Wang, J. Lin, M. Sun, P. Krahenbuhl, T. Darrell, and F. Yu. Joint monocular 3d vehicle detection and tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[20] Q. Huang, Y. Xiong, A. Rao, J. Wang, and D. Lin. Movienet: A holistic dataset for movie understanding. In *European Conference on Computer Vision*, 2020.

[21] J. Johnson, M. Douze, and H. Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

[22] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In *IEEE International Conference on Computer Vision*, 2023.

[23] M. Kolmet, Q. Zhou, A. Ošep, and L. Leal-Taixe. Text2pos: Text-to-point-cloud cross-modal localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[24] L. Kreuzberg, I. E. Zulfikar, S. Mahadevan, F. Engelmann, and B. Leibe. 4d-stop: Panoptic segmentation of 4d lidar using spatio-temporal object proposal generation and aggregation. In *ECCV AVVision Workshop*, 2022.

[25] S. Li, M. Danelljan, H. Ding, T. E. Huang, and F. Yu. Tracking every thing in the wild. In *European Conference on Computer Vision*, 2022.

[26] S. Li, T. Fischer, L. Ke, H. Ding, M. Danelljan, and F. Yu. Ovtrack: Open-vocabulary multiple object tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023.

[27] 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 *European Conference on Computer Vision*, 2014.

[28] Z. C. Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. In *ICML Workshop on Human Interpretability in Machine Learning*, 2016.

[29] Y. Liu, I. E. Zulfikar, J. Luiten, A. Dave, D. Ramanan, B. Leibe, A. Ošep, and L. Leal-Taixé. Opening up open-world tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[30] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE International Conference on Computer Vision*, 2021.

[31] J. Luiten, T. Fischer, and B. Leibe. Track to reconstruct and reconstruct to track. *IEEE Robotics and Automation Letters*, 5(2):1803–1810, 2020.

[32] W. Maddern, G. Pascoe, C. Linegar, and P. Newman. 1 year, 1000km: The Oxford RobotCar dataset. *International Journal of Robotics Research*, 36(1):3–15, 2017.

[33] R. Marcuzzi, L. Nunes, L. Wiesmann, E. Marks, J. Behley, and C. Stachniss. Mask4d: End-to-end mask-based 4d panoptic segmentation for lidar sequences. *IEEE Robotics and Automation Letters*, 2023.

[34] M. Najibi, J. Ji, Y. Zhou, C. R. Qi, X. Yan, S. Ettinger, and D. Anguelov. Motion inspired unsupervised perception and prediction in autonomous driving. In *European Conference on Computer Vision*, 2022.

[35] P. Nguyen, K. G. Quach, K. Kitani, and K. Luu. Type-to-track: Retrieve any object via prompt-based tracking. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[36] A. Ošep, P. Voigtlaender, J. Luiten, S. Breuers, and B. Leibe. Towards large-scale video video object mining. *European Conference on Computer Vision Workshop on Interactive and Adaptive Learning in an Open World*, 2018.

[37] A. Ošep, W. Mehner, M. Mathias, and B. Leibe. Combined image- and world-space tracking in traffic scenes. In *International Conference on Robotics and Automation (ICRA)*, 2017.

[38] A. Ošep, W. Mehner, P. Voigtlaender, and B. Leibe. Track, then decide: Category-agnostic vision-based multi-object tracking. In *International Conference on Robotics and Automation (ICRA)*, 2018.

[39] A. Ošep, P. Voigtlaender, J. Luiten, S. Breuers, and B. Leibe. Large-scale object mining for object discovery from unlabeled video. In *International Conference on Robotics and Automation (ICRA)*, 2019.

[40] A. Ošep, P. Voigtlaender, M. Weber, J. Luiten, and B. Leibe. 4d generic video object proposals. In *International Conference on Robotics and Automation (ICRA)*, 2020.

[41] M. Pedersen, J. B. Haurum, S. H. Bengtson, and T. B. Moeslund. 3D-ZeF: A 3D zebrafish tracking benchmark dataset. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[42] P. Pinheiro, R. Collobert, and P. Dollár. Learning to segment object candidates. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[43] P. Pinheiro, T. Lin, R. Collobert, and P. Dollár. Learning to refine object segments. In *European Conference on Computer Vision*, 2016.

[44] L. Porzi, S. R. Bulo, A. Colovic, and P. Kontschieder. Seamless scene segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[45] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. *International Conference on Learning Representations (ICLR)*, 2021.

[46] J. Redmon and A. Farhadi. Yolo9000: better, faster, stronger. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[47] 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 (NIPS)*, 2015.

[48] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele. A dataset for movie description. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[49] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[50] S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*. The MIT Press, 2005.

[51] A. Torabi, C. J. Pal, H. Larochelle, and A. Courville. Using descriptive video services to create a large data source for video annotation research. *arXiv preprint arXiv:1503.01070*, 2015.

[52] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[53] P. Vicol, M. Tapaswi, L. Castrejon, and S. Fidler. Moviegraphs: Towards understanding human-centric situations from videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[54] P. Viola, M. Jones, et al. Robust real-time object detection. *International Journal of Computer Vision (IJCV)*, 4(34-47):4, 2001.

[55] P. Voigtlaender, M. Krause, A. Ošep, J. Luiten, B. Sekar, A. Geiger, and B. Leibe. MOTS: Multi-object tracking and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[56] G. Wang, H. Fan, and M. Kankanhalli. Text to point cloud localization with relation-enhanced transformer. In *Association for the Advancement of Artificial Intelligence*, 2023.

[57] M. Weber, J. Xie, M. Collins, Y. Zhu, P. Voigtlaender, H. Adam, B. Green, A. Geiger, B. Leibe, D. Cremers, , A. Ošep, L. Leal-Taixé, and L.-C. Chen. Step: Segmenting and tracking every pixel. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[58] X. Weng, J. Wang, D. Held, and K. Kitani. 3D Multi-Object Tracking: A Baseline and New Evaluation Metrics. *IEEE Intelligent Robots and Systems (IROS)*, 2020.

[59] X. Weng, Y. Wang, Y. Man, and K. Kitani. Gnn3dmot: Graph neural network for 3d multi-object tracking with multi-feature learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[60] X. Wu, K. Lau, F. Ferroni, A. Ošep, and D. Ramanan. Pix2map: Cross-modal retrieval for inferring street maps from images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023.

[61] Y. Wu, M. N. Rabe, D. Hutchins, and C. Szegedy. Memorizing transformers. In *International Conference on Learning Representations (ICLR)*, 2022.

[62] Y. Xia, L. Shi, Z. Ding, J. F. Henriques, and D. Cremers. Text2loc: 3d point cloud localization from natural language. *arXiv preprint arXiv:2311.15977*, 2023.

[63] Z. Xu, W. Zhang, X. Tan, W. Yang, H. Huang, S. Wen, E. Ding, and L. Huang. Segment as points for efficient online multi-object tracking and segmentation. In *European Conference on Computer Vision*, 2020.

[64] L. Yang, Y. Fan, and N. Xu. Video instance segmentation. In *IEEE International Conference on Computer Vision*, 2019.

[65] K. Yilmaz, J. Schult, A. Nekrasov, and B. Leibe. Mask4d: Mask transformer for 4d panoptic segmentation. *arXiv preprint arXiv:2309.16133*, 2023.

[66] T. Yin, X. Zhou, and P. Krahenbuhl. Center-based 3d object detection and tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[67] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[68] L. Zhang, A. J. Yang, Y. Xiong, S. Casas, B. Yang, M. Ren, and R. Urtasun. Towards unsupervised object detection from lidar point clouds. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023.

[69] M. Zhu, S. Han, H. Cai, S. Borse, M. Ghaffari, and F. Porikli. 4d panoptic segmentation as invariant and equivariant field prediction. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023.