TAG: Tracking at Any Granularity

Adam W. Harley* **Yang You** **Alex Sun** **Yang Zheng** **Nikhil Raghuraman**

Sheldon Liang **Wen-Hsuan Chu** **Suya You** **Achal Dave** **Pavel Tokmakov**

Rares Ambrus **Katerina Fragkiadaki** **Leonidas Guibas**

Abstract

We introduce Tracking at Any Granularity (TAG): a new task, model, and dataset for tracking arbitrary targets in videos. We seek a tracking method that treats points, parts, and objects as equally trackable target types, embracing the fact that the distinction between these granularities is ambiguous. We introduce a generic high-capacity transformer for the task, taking as input a video and a target prompt (indicating what to track, in the form of a click, box, or mask), and producing as output the target's segmentation on every frame. To train the model, we aggregate nearly all publicly-available tracking datasets that we are aware of, which currently totals 75, amounting to millions of clips with tracking annotations, including a long tail of rare subjects such as body keypoints on insects and microscopy data. Our model is competitive with state-of-the-art on standard benchmarks for point tracking, mask tracking, and box tracking, but more importantly, achieves zero-shot performance far superior to prior work, largely thanks to the data effort. We will publicly release our code, model, and aggregated dataset, to provide a foundation model for motion and video understanding, and facilitate future research in this direction.

1 Introduction

Large-scale models, powered by enormous compute and enormous data, have become the method of choice in a variety of computer vision domains, notably including classic tasks such as image classification (Dosovitskiy et al., 2021; Radford et al., 2021; Caron et al., 2021b), image segmentation (Kirillov et al., 2023), and action recognition (Feichtenhofer et al., 2022). The strong results from these "foundation models" (Bommasani et al., 2021) impel us to relinquish our hard-earned special-purpose techniques, and let data-driven optimization build new models from scratch. In this work, we take some first steps toward applying this method to a domain which is perhaps more heavily engineered than most: *visual tracking*.

Tracking is so fundamental to understanding the visual world, and its applications are so diverse, that the vision community has divided the problem into countless distinct sub-problems, and has devoted great energy to mastering them all, more or less independently. Existing "general-purpose" trackers first of all partition the world on the axis of *granularity*: point tracking methods and object tracking methods use entirely separate techniques, datasets, and metrics. Among object trackers, it is typical to organize targets into *categories*: different models are devoted to tracking people vs. cars vs. animals, and so on. Category-specific trackers often subdivide their subjects on granularity: humans, for example, are partitioned into faces, bodies, and hands, and these elements are subdivided further to enable the highest-precision applications, such as performance capture.

*aharley@cs.stanford.edu

Preprint. Under review.

This organization of the problem space has gradually led to extremely high-performing special-purpose methods, powered by carefully crafted techniques. Two techniques in particular appear to have stood the test of time, and are used across a wide spectrum of tracking problems: (1) cross correlation (Bolme et al., 2010), and (2) tracking by detection (Huang et al., 2008). Cross correlation simply refers to computing the cosine similarity between the feature of the target and the features of an incoming video frame, and using this as the primary cue for establishing matches across time. Tracking by detection relies on the assumption that on a new frame, the target's location can be reduced to a manageable number of candidate locations via the use of a detector, and from here the tracking problem is converted to the problem of associating detections across time. These well-reasoned techniques (often used together) offer a strict information bottleneck for learning-based methods, and thereby prevent overfitting: models that rely on these techniques are forced to solve tracking by *matching*.

Inspired by the success of large-scale models in other domains, our goal in this work is to build a *foundation model for visual tracking*. We propose a transformer architecture (Vaswani et al., 2017) that is prompted with a target to track in the form of a pixel, box or segmentation mask, and predicts the area the target occupies in the input set of frames, which we train supervised using an aggregate of diverse video datasets available. The architecture relies on spatio-temporal attention that jointly featurizes the target prompt at the first frame with the video frame set. Such joint featurization goes beyond cross-correlation and does not rely on linking detections, rather, learns the tracking objective end-to-end through large data training.

We test our method on a set of tracking benchmarks for points (Zheng et al., 2023; Pedersen et al., 2020), masks (Pont-Tuset et al., 2017; Tokmakov et al., 2023) and boxes (Nicora et al., 2020; Anjum and Gurari, 2020), and show we reach state-of-the-art performance in point and mask tracking and are close to the SOTA in segmentation tracking. Our code and data will be publicly available.

To summarize, our contributions are: (1) TAG, a foundation model for visual tracking that unifies point, box, and mask tracking into a single architecture, which relies on (2) jointly featurizing the target object with the video clip with a simple generic architecture, allowing end-to-end training, unlike prior work which relies on handcrafted techniques such as correlation and tracking by detection. Finally, (3) we aggregate existing datasets for disparate tracking tasks into the largest single collection of publicly available tracking data, covering a wide range of domains, which we use to train our model.

2 Related Work

Foundation models for computer vision. In recent years, simple yet scalable models have demonstrated state-of-the-art performance across many areas, including computer vision (Bommasani et al., 2021). Methods such as CLIP (Radford et al., 2021), DINO (Caron et al., 2021a), Masked Autoencoding (He et al., 2022), and Stable Diffusion (Rombach et al., 2022) highlight the importance of large-scale training with simple training objectives. In object detection and segmentation, Kirillov et al. (2023), shows remarkable performance using a transformer-based model trained on a large annotated dataset, with minimal task-specific post-processing as used in prior work such as He et al. (2017). Goel et al. (2023) follows a similar paradigm for 3D human pose estimation and tracking, eschewing expensive post-processing steps used in prior work, such as Kolotouros et al. (2019). By contrast, nearly all tracking approaches focus on specific subtasks, with significant post-processing. In this work, we aim to build a scalable foundation model for visual tracking, by unifying the input and output space of different tracking subtasks, which in turn allows us to train a generic transformer-based model on a large number of tracking datasets.

Aggregated tracking datasets. To build diverse datasets for training and evaluation, a common tactic has been to *aggregate* disparate datasets with unique features, resulting in both increased size and diversity. The widely-used VOT dataset (Kristan et al. (2016)), for example, selectively incorporates sequences from other datasets to target diverse use cases, in addition to sequences curated by the benchmark organizers. Dave et al. (2020) gather sequences from multiple independent sources, ranging from autonomous driving to user-generated content. Our work is particularly inspired by Ranftl et al. (2020), which showed strong improvements in zero-shot depth estimation by introducing a training strategy that could leverage diverse datasets with incompatible annotations. Zhou et al. (2022) similarly aggregate datasets for object detection, building a unified model that can train on

disparate object vocabularies. Our approach follows this line of work, aggregating diverse datasets and presenting a unified approach that allows training on all of them.

Specialized tracking methods. The tracking literature is split into a number of subfields, each with a specific definition and a largely disjoint set of methods. We focus here on *user-initialized* tracking, where a point, box, or mask corresponding to the target to be tracked is provided. Unlike prior work, our work proposes a single trained model that performs well on each of these task definitions. We review the relevant task-specific methods here.

Point tracking methods (e.g. Harley et al. (2022); Doersch et al. (2022b)) often employ expensive cost volumes to distinguish different points from each other. Our approach diverges from this norm and directly estimates heatmaps locating the points in subsequent frames without computing cost volumes, allowing us to unify different tracking tasks under one model.

Box tracking methods methods initially (Bertinetto et al., 2016; Li et al., 2019) used feature matching for tracking, followed by an optional regression head for estimating scale. More recently, Yan et al. (2021) directly localize target objects using corner prediction heads to generate probability maps for the top-left and bottom-right corners of the target. Our approach is similar, but differs in that we simultaneously tackle other tracking tasks in other heatmap channels.

Mask tracking methods employ a segmentation-centric tracking process to propagate a mask over time (Perazzi et al., 2017; Li and Loy, 2018), optionally using a learned target detector (Caelles et al., 2017; Voigtlaender and Leibe, 2017). Some methods (e.g., Voigtlaender et al. (2020)) use a *box2seg* component that transforms the box output into a mask, while others (e.g., Wang et al. (2019)) generate a segmentation mask directly, guided by a segmentation loss. Our method follows the latter, featuring a specialized task head in our model to directly predict the mask. Compared to these specialized tracking architectures, our proposed model unifies these tasks into one common input and output format, providing a unified model that can track anything.

3 Tracking at Any Granularity (TAG)

3.1 Setup and overview

At its core, our TAG model is a high-capacity transformer, which takes a video as input, along with a prompt indicating what to track, and outputs tracking across all frames in the video simultaneously.

This is "offline" (as opposed to "online") tracking, in the sense that the model is processing a saved video rather than an incoming data stream. In practice, the model processes a window of frames from the video (rather than the entire video), which we sometimes refer to as a "clip". We later describe a simple method for joining the tracking information across clips.

Processing a clip of frames, rather than 1-2 frames at a time, is somewhat unusual in object trackers, but this choice follows a trend from point tracking (Harley et al., 2022). A multi-frame setup has two key benefits: (1) the model may more easily take advantage of priors that exist on the temporal axis, such as the continuity of the object's motion (and the camera's motion) over time, and (2) occlusions are less of a concern, because the object may return to view at some later timestep within the model's inference window.

Concretely, we are given as input a video of shape $T, H, W, 3$, along with a prompt of shape $H, W, 1$ (for a specific timestep), where T is the number of timesteps in the clip and H, W indicates the spatial resolution of each frame. Our method begins by compressing this to $T/8 \cdot H/32 \cdot W/32$ "tokens". We perform the bulk of our processing on these low-resolution tokens, to keep compute tractable. The final stage of the model is to upsample from this compact representation and output a spatiotemporal volume shaped T, H, W with K channels, where K is the number of heatmaps, corresponding to the different tracking granularities. The specific heatmap(s) taken as final output depends on the prompt type (e.g., if a point prompt is provided, the first heatmap is taken), but all of the heatmaps are produced in each forward pass. The model architecture is illustrated in Figure 1.

Outputting multiple granularities not only allows our model to be multi-purpose at test time, but allows it to accept supervision from arbitrary-granularity tracking annotations at training time.

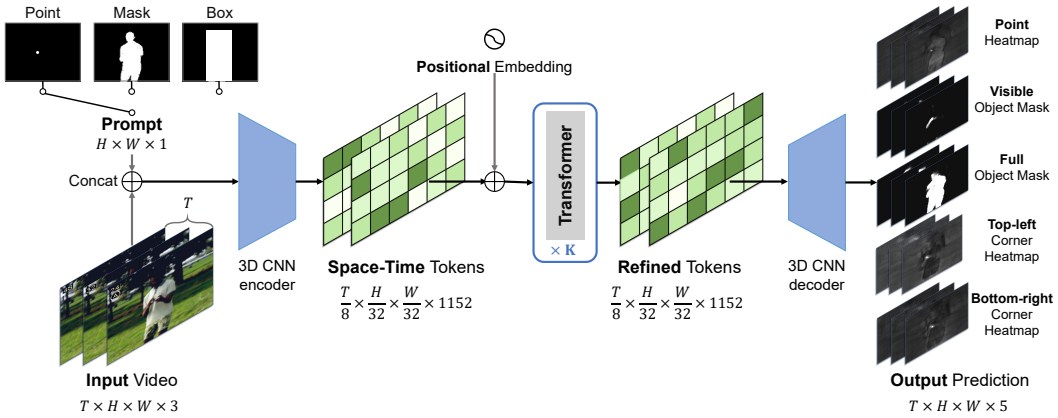

Figure 1: **Overview of our TAG model.** Our method can take any type of input prompt, e.g., point, mask, or bounding box, and then the prompt together with the input video are encoded into a highly-compressed set of space-time tokens. These tokens are then processed by a large transformer, and then decoded into multi-granularity tracking outputs. Depending on the prompt, different channels will be relevant, but the model outputs all channels for all timesteps in the forward pass.

3.2 Encoding

Our first stage encodes the input video, together with the prompt, into a manageable set of discrete tokens. We achieve this with a simple custom-designed convolutional network, based on the encoder from RAFT (Teed and Deng, 2020) and PIPs (Harley et al., 2022) but upgraded to 3D so as to perform a spacetime compression.

Our encoder network takes a 4-channel sequence of frames as input. The first three channels are RGB, and the fourth is a channel for the prompt, which on most frames is empty (zeros), but on one frame crucially indicates the "target" for tracking. This encoder has spatial stride 32, and temporal stride 8, and therefore reduces the input $T, H, W, 4$ to the shape $T/8, H/32, W/32, C$, where C is the embedding dimension, set here as $C = 1152$.[2]

Early experiments determined that this convolutional encoder is preferable over the "patchify" stem typically used in vision transformers, consistent with the findings of Xiao et al. (2021).

3.3 Processing

The main processing is carried out by a transformer (Vaswani et al., 2017). This stage aims to process the spatiotemporal information within and across tokens, so as to enable the end-task to be solved by the decoder (which is shallow by comparison).

In this stage, we reshape the encoder's output into a list of length $T/8 \cdot H/32 \cdot W/32$, and infuse these features with high-dimensional indicators of their spacetime coordinates, following common practice with sinusoidal embeddings. The transformer then proceeds with the standard "blocks" of layers: where each block consists of a LayerNorm, self-attention layer, residual connection, second LayerNorm, a fully connected layer, and a second residual connection. Twelve such blocks are applied.

Early experiments revealed that an MLP-Mixer (Tolstikhin et al., 2021) can be used as the processor instead of a transformer, with similar compute cost and similar accuracy, but this choice locks the model to a single sequence length (i.e., a fixed choice of $T/8 \cdot H/32 \cdot W/32$), whereas a transformer can be trained and tested with variable resolutions.

[2]We use $C = 1152$ so as to be easily divisible by 3, which is helpful when infusing space-time (3D) sinusoidal position embeddings.

3.4 Decoding

After processing, the information in the tokens is rapidly upsampled into the shape of a spacetime volume with dimensions similar to the initial volume, but with the spatial axes at half-resolution.

This is achieved by a convolutional network, based on the "Simple Feature Pyramid" from Li et al. (2022). This consists of four parallel upsampling branches, which upsample different amounts (1x, 2x, 4x, 8x), and a final head which merges these and upsamples again (2x). We additionally provide the final head with skip connections from the encoder, to facilitate the capture of high-frequency details. The upsampling is done with transposed 3D convolution layers. In total, the temporal dimension is upsampled 8x and the spatial dimensions are upsampled 16x, bringing the output to $T, H/2, W/2$, and finally we use a bilinear resize (parameter-free) to arrive at full resolution T, H, W.

Each of the decoder branches progressively lowers the channel dimension while upsampling, to help reduce memory and compute at the higher resolutions. The final layer of the decoder outputs $K = 5$ channels, and these are taken to represent multiple forms of tracking:

- **Point tracking**: This is represented as a heatmap, which ideally contains a strong peak at the point center.
- **Visible object tracking**: This is represented as a mask, delineating the region of the object that is visible in the current frame.
- **Full object tracking**: This is represented as a mask, delineating the region that the full object would occupy if it were not occluded.
- **Box tracking**: This is represented as two heatmaps, representing coordinates with peaks: the first indicates the top left corner of the box, and the second indicates the bottom right corner of the box.

These channels are supervised with ground truth, when available.

We also estimate visibility, as follows: for each timestep at the compressed resolution, we take an average of the spatial features, and then send this $T/8$ sequence of tokens through a convolutional 1D upsampler, arriving at T visibility scalars in the end.

3.5 Supervision

Different datasets provide varying amounts and varying types of ground truth. Our strategy is to use all available supervision. We simply convert every type of annotation into a binary cross-entropy loss defined on one or more heatmaps:

$$\mathcal{L}_{\text{ce}} = \mathcal{H}^* \log \mathcal{H} + (1 - \mathcal{H}^*) \log(1 - \mathcal{H}), \tag{1}$$

where $\mathcal{H}$ and $\mathcal{H}^*$ indicate estimated and ground truth heatmaps. We supervise the visibility head with another cross entropy loss of the same form.

Point annotations are converted into ground-truth heatmaps which have a single 1 at the point coordinate and zeros elsewhere. Box annotations are similarly converted, but we produce one heatmap for the top-left coordinate and another heatmap for the bottom-right coordinate of the box. Mask annotations, for the visible portion of the object and for the full object (which is rarely available), are treated as binary masks.

We additionally create weak supervision where possible, by making appropriate use of ignore masks. For example, we convert box annotations into partial labels for the "visible object" mask, by covering the box region with "ignore" and setting the remainder of the frame as "negative". We also propagate point and visible-object labels to the full-object channel, and create box corner labels from masks.

3.6 Iterative inference

Given the large spatial stride of our model, it is difficult for it to resolve fine-grained details, which is a necessary ability when tracking points. We resolve this issue during inference, by "zooming in" on the target.

Concretely, for point tracking specifically, we begin our tracking procedure by initializing a trajectory centered on the prompt coordinate (i.e., with zero velocity), and then perform a forward pass of the

Figure 2: Samples from the training dataset for TAG, along with ground truth annotation (green) and predictions from our model (red). We focused our effort on curating data from existing tracking datasets to ensure high diversity, as shown in the samples above, which include scenes commonly found in popular datasets (urban and natural environments) but also a long tail of challenging tasks for insect tracking and microscopy (left column). We target such diverse scenes to ensure our resulting model is useful for a range of applications, including for researchers and practitioners in other scientific fields.

model, and then *re-center* the crop at every timestep according to the estimated trajectory, and repeat inference. Inference typically stabilizes to a solution within 8 steps.

This achieves an iterative refinement procedure, quite similar to the method used in the current state-of-the-art in point tracking (Harley et al., 2022; Karaev et al., 2023), but our strategy does not necessitate the usage of refinement steps during training.

3.7 Implementation details

Our model is 285M parameters, the majority of which belong to the transformer at the center. We train it across 8 A100-80G GPUs. We find that with this compute, training is very slow: the model only plateaus in improvement after approximately *30 days* of continuous training, or about 5,760 A100 hours. As we are limited by compute, this is significantly less than other large-scale models like SegmentAnything (17,408 hours, Kirillov et al. (2023)) or CLIP (86,016 hours, Radford et al. (2021)), and we expect results would improve with additional compute. We vary sequence lengths and resolutions semi-randomly during training, and finish training with a high-resolution long-sequence fine-tuning stage (using $T = 64, H = 224, W = 256$). At inference time, the model runs at 224×256 resolution at 80 FPS. For object tracking datasets, we simply resize the video to this resolution. For point tracking datasets, we keep the original video resolution and take internal crops. To track across videos longer than 64 frames, we simply re-prompt the tracker using an estimate from the second half of the video, selecting this frame with the help of the visibility estimates.

4 Datasets

We provide a high-level overview of all the datasets considered in this work in Table 1 and Figure 2. We have parsed and aggregated nearly every tracking dataset that we are aware of, which currently totals 75. We separated the data into subsets for training, validation, and testing, using some datasets exclusively for testing so as to measure one-shot performance on unseen domains. We filtered the data to remove clips without motion. We additionally created heuristics to help reduce the amount of data with annotation errors, e.g., detecting "teleports" by setting thresholds on velocity and acceleration. In

Table 1: Datasets used in training, with annotation types and frame counts.

Name	Annotation type	Number of frames
Ant Det and Track (Wu et al., 2022)	Box	5334
AcinoSet (Joska et al., 2021)	Point	114856
Animaltrack (Zhang et al., 2022)	Box	19858
BDD100K (Yu et al., 2020)	Mask	44776
BEE23 (Cao et al., 2023)	Box	3562
BL30K (Cheng et al., 2021)	Mask	2400000
BURST (Athar et al., 2023)	Mask	2099437
COCO (Lin et al., 2014)	Mask	123287
DanceTrack (Sun et al., 2022)	Point	67304
FBMS (Brox and Malik, 2010; Ochs et al., 2013)	Mask	13860
Fishschooling (Lauer et al., 2022)	Point	70
FlyingThings (Mayer et al., 2016)	Point & Mask	53520
GMOT-40 (Bai et al., 2021)	Mask	9643
GOT10k (Huang et al., 2019)	Box	1447200
HOB (Kuipers et al., 2020)	Box	55388
HOOT (Sahin and Itti, 2023)	Box	435790
Horse-10 (Mathis et al., 2021)	Point	8814
Interhand2.6M (Moon et al., 2020)	Point	2590347
KITTI (Geiger et al., 2013)	Box	19103
KubricPoints (Greff et al., 2022)	Point	32256
KubricContainers (Van Hoorick et al., 2023)	Vis. & Full Mask	972
KubricRandom (Van Hoorick et al., 2023)	Vis. & Full Mask	136800
LASOT (Fan et al., 2019)	Box	3517342
LV-VIS (Wang et al., 2023)	Mask	89203
MOSE (Ding et al., 2023)	Mask	11054
MOT17 (Milan et al., 2016)	Box	33705
MOTS (Voigtlaender et al., 2019)	Mask	5906
Marmoset (Lauer et al., 2022)	Point	5316
OTB (Wu et al., 2013)	Box	75135
OVIS (Qi et al., 2022)	Mask	62641
OpenMonkeyStudio (Bala et al., 2020)	Point	194518
Parenting-mice (Lauer et al., 2022)	Point	379
PointOdyssey (Zheng et al., 2023)	Point & Mask	508142
PoseTrack21 (Doering et al., 2022)	Point	162064
SAIL-VOS (Hu et al., 2019)	Vis. & Full Mask	113490
SportsMot (Cui et al., 2023)	Box	150379
SurgicalHands (Louis et al., 2023)	Point	5134
TEyeD (Fuhl et al., 2021)	Point+Mask	562139
TLP (Moudgil and Gandhi, 2019)	Box	653484
TempleColor-128 (Liang et al., 2015)	Box	55652
TrackingNet (Muller et al., 2018)	Box	7089102
Tri-mice (Lauer et al., 2022)	Point	112
UAV123 (Mueller et al., 2016)	Box	113476
UBody (Lin et al., 2023)	Point	1065925
VSB100 (Galasso et al., 2013)	Mask	923
VideoCube (Hu et al., 2022)	Box	1029712
Visdrone-SOT2019 (Du et al., 2019)	Box	109909
YouTube-VOS (Yang et al., 2022)	Mask	108298
Total		24891859

total this process yielded approximately 4 million unique 32-frame clips for training (with 1 tracking annotation each), and approximately 2 million higher-resolution 64-frame clips.

5 Experiments

We test our model in point, mask and box tracking across diverse held-out benchmarks.

Baselines. We compare against the following state-of-the-art point, mask and box trackers:

Table 2: Comparison against state-of-the-art in point tracking, across multiple datasets. Columns report δ_{avg}, an accuracy metric (higher is better; see text for details).

Dataset	Method	
	CoTracker (Karaev et al., 2023)	Ours
TapVID-Davis (Doersch et al., 2022a)	75.7	**75.8**
PointOdyssey-Test (Zheng et al., 2023)	46.4	**51.0**
ZebraFish (Pedersen et al., 2020)	41.6	**77.3**
CroHD (Sundararaman et al., 2021)	55.5	**57.1**
NewZealand-Wildlife (LILA)	50.1	**52.3**

Table 3: Ablation study on "iterative zooming", across multiple point tracking datasets. Columns report δ_{avg}, an accuracy metric (higher is better; see text for details).

	Method	
Dataset	Ours	Ours w/o zooming
TapVID-Davis (Doersch et al., 2022a)	**75.8**	58.4
PointOdyssey-Test (Zheng et al., 2023)	**51.0**	42.6
ZebraFish (Pedersen et al., 2020)	**77.3**	61.5
CroHD (Sundararaman et al., 2021)	**57.1**	46.2
NewZealand-Wildlife (LILA)	**52.3**	42.3

(1) CoTracker (Karaev et al., 2023), which is a state of the art point tracking method that jointly infers multiple point trajectories through iterative inference. At each iteration, it computes cross correlation features maps around the estimated point locations and a transformer attends across different tracks to update their location and feature estimates.

(2) Cutie (Cheng et al., 2023), a state-of-the-art approach for mask tracking, employs a query-based object transformer to maintain both a bottom-up dense memory, as well as a top-down memory of objects to distinguish distractor objects.

(3) MixFormer (Cui et al., 2022) introduces a mixed attention module (MAM) to allow joint extraction of features for search regions and matching to the target region. The method only performs box tracking, and achieves state-of-the-art results on box tracking benchmarks.

Evaluation metrics. We report the standard evaluation metrics for each type of tracking output. Point tracking uses δ_{avg}, which is the average of $[\delta_1, \delta_2, \delta_4, \delta_8, \delta_{16}]$, where each δ_k captures the proportion of timesteps where the estimated track is within k pixels of the ground truth. Mask tracking uses J and F, which are the Jaccard score and F-measure, where the Jaccard score measures the intersection over union (IoU) of the masks, and the F measure highlights accuracy near boundaries. Object tracking uses two precision metrics, P for precision itself and P_{norm} which is normalized using the box dimensions, as well as AUC which captures the area under the precision/recall curve.

5.1 Point tracking

We show quantitative results for point tracking in Table 2. Despite being trained for point, mask, and box tracking jointly, our model outperforms the specialized point-tracking state-of-the-art in all benchmarks. While large-scale training is a key contributor to our results, we additionally find that one component of our architecture, iterative zooming, is critical for fine-grained tracking, as seen in Table 3.

5.2 Mask tracking

We show quantitative results comparing to the state-of-the-art for mask tracking in Table 4. We evaluate on both the J and F metrics across four datasets. We find that our method underperforms compared to the task-specific method, possibly due to over-compression at the bottleneck, or the model being under-capacity. Unlike for box and point tracking, we are unable to evaluate Cutie on

Table 4: Comparison against state-of-the-art in mask tracking, across multiple datasets. Columns report Jaccard scores (J) and F-measure (F) (higher is better; see text for details).

Dataset	Method			
	Cutie Cheng et al. (2023)		Ours	
	J	F	J	F
DAVIS (Pont-Tuset et al., 2017)	**81.3**	**89.1**	76.0	80.5
VOST (Tokmakov et al., 2023)	**42.8**	**52.7**	31.3	46.3
UVO (Wang et al., 2021)	**78.4**	**86.0**	70.5	70.8
YCBInEOAT (Wen et al., 2020)	**65.7**	**80.5**	52.9	58.8

Table 5: Comparison against state-of-the-art in box tracking, across multiple datasets. Columns report precision scores (P, P_{norm}) and area under the curve (AUC) (higher is better; see text for details).

Dataset	Method					
	MixFormer Cui et al. (2022)			Ours		
	P	P_{norm}	AUC	P	P_{norm}	AUC
LaTOT (Zhu et al., 2023)	60.7	**41.6**	**36.4**	**89.0**	32.0	35.3
MOCA (Nicora et al., 2020)	79.9	**98.1**	**85.7**	**83.0**	76.7	73.4
CTMC (Anjum and Gurari, 2020)	49.2	45.5	39.9	**67.4**	**46.2**	**52.4**
TOTB (Fan et al., 2021)	90.8	**90.4**	**82.8**	**93.4**	79.2	74.4
Zebrafish (Pedersen et al., 2020)	62.0	53.1	47.8	**94.3**	**53.6**	**58.2**

out-of-domain natural science datasets, because those datasets are only labeled with point or box annotations.

5.3 Box tracking

Finally, we evaluate box tracking in Table 5 on 5 diverse datasets, reporting precision (P, P_{norm}) and AUC (area under the curve). Again, although our method contains no specialized machinery for box tracking and can perform point, mask, and box tracking jointly, we match or outperform the state-of-the-art in specialized box tracking approaches.

6 Conclusion and Limitations

We proposed a model and a training strategy that allows training on datasets for disparate tracking tasks. Our single model achieves state-of-the-art performance on a number of tracking tasks without any additional post-processing, outperforming prior methods which are tailored for specific tasks. Our work shows the potential of solving tracking with a general transformer model trained with a large amount of data. As with all tracking approaches, our method can be used for positive and negative purposes, but we focus here on foundational model capabilities. We highlight a few key limitations for future work.

Scaling to more datasets. While we have collected a large set of point, mask, and box tracking datasets, this was a largely manual process that we were unable to extend to a number of important tracking datasets. We plan to open-source our implementation to enable the community to build on the datasets we've collected, but also to allow others to contribute implementations for datasets we have not yet implemented. We expect this common repository of datasets will lead to faster, more general advances in object tracking.

Dense correspondences. Our method achieves state-of-the-art results on point tracking by a wide margin. However, similar to other point-tracking approaches, extracting dense, fine-grained temporal correspondences under our model would require running inference on every pixel separately.

Unifying detection. Our approach unifies disjoint tasks in user-initialized tracking, where the point, mask, or box of interest is provided as a prior. However, we do not tackle multi-object tracking, which

requires additionally detecting all objects from a given list of categories, as we do not address object detection as part of our pipeline. A useful extension of this work would be to address object detection and multi-object tracking with our method, which may further improve results on user-initialized tracking as well.

Acknowledgements

This work was supported by the Toyota Research Institute under the University 2.0 program, ARL grant W911NF-21-2-0104, and a Vannevar Bush Faculty Fellowship.

References

Samreen Anjum and Danna Gurari. Ctmc: Cell tracking with mitosis detection dataset challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 982–983, 2020.

Ali Athar, Jonathon Luiten, Paul Voigtlaender, Tarasha Khurana, Achal Dave, Bastian Leibe, and Deva Ramanan. Burst: A benchmark for unifying object recognition, segmentation and tracking in video. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1674–1683, 2023.

Hexin Bai, Wensheng Cheng, Peng Chu, Juehuan Liu, Kai Zhang, and Haibin Ling. Gmot-40: A benchmark for generic multiple object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6719–6728, 2021.

Praneet C Bala, Benjamin R Eisenreich, Seng Bum Michael Yoo, Benjamin Y Hayden, Hyun Soo Park, and Jan Zimmermann. Automated markerless pose estimation in freely moving macaques with openmonkeystudio. *Nature communications*, 11(1):4560, 2020.

Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*, pages 850–865. Springer, 2016.

David S Bolme, J Ross Beveridge, Bruce A Draper, and Yui Man Lui. Visual object tracking using adaptive correlation filters. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 2544–2550. IEEE, 2010.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Thomas Brox and Jitendra Malik. Object segmentation by long term analysis of point trajectories. In *European conference on computer vision*, pages 282–295. Springer, 2010.

Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 221–230, 2017.

Xiaoyan Cao, Yiyao Zheng, Yao Yao, Huapeng Qin, Xiaoyu Cao, and Shihui Guo. Topic: A parallel association paradigm for multi-object tracking under complex motions and diverse scenes. *arXiv preprint arXiv:2308.11157*, 2023.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021a.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021b.

Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Modular interactive video object segmentation: Interaction-to-mask, propagation and difference-aware fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5559–5568, 2021.

Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the object back into video object segmentation. In *arXiv*, 2023.

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.

Yutao Cui, Chenkai Zeng, Xiaoyu Zhao, Yichun Yang, Gangshan Wu, and Limin Wang. Sportsmot: A large multi-object tracking dataset in multiple sports scenes. *arXiv preprint arXiv:2304.05170*, 2023.

Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. Tao: A large-scale benchmark for tracking any object. In *ECCV*, 2020.

Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip HS Torr, and Song Bai. Mose: A new dataset for video object segmentation in complex scenes. *arXiv preprint arXiv:2302.01872*, 2023.

Andreas Doering, Di Chen, Shanshan Zhang, Bernt Schiele, and Juergen Gall. PoseTrack21: A dataset for person search, multi-object tracking and multi-person pose tracking. In *CVPR*, 2022.

Carl Doersch, Ankush Gupta, Larisa Markeeva, Adria Recasens Continente, Kucas Smaira, Yusuf Aytar, Joao Carreira, Andrew Zisserman, and Yi Yang. Tap-vid: A benchmark for tracking any point in a video. In *NeurIPS Datasets Track*, 2022a.

Carl Doersch, Ankush Gupta, Larisa Markeeva, Adrià Recasens, Lucas Smaira, Yusuf Aytar, João Carreira, Andrew Zisserman, and Yi Yang. Tap-vid: A benchmark for tracking any point in a video. *arXiv preprint arXiv:2211.03726*, 2022b.

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. In *ICLR*, 2021.

Dawei Du, Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Ling, Qinghua Hu, Jiayu Zheng, Tao Peng, Xinyao Wang, Yue Zhang, et al. Visdrone-sot2019: The vision meets drone single object tracking challenge results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

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.

Heng Fan, Halady Akhilesha Miththanthaya, Siranjiv Ramana Rajan, Xiaoqiong Liu, Zhilin Zou, Yuewei Lin, Haibin Ling, et al. Transparent object tracking benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10734–10743, 2021.

Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. Masked autoencoders as spatiotemporal learners. *Advances in neural information processing systems*, 35:35946–35958, 2022.

Wolfgang Fuhl, Gjergji Kasneci, and Enkelejda Kasneci. Teyed: Over 20 million real-world eye images with pupil, eyelid, and iris 2d and 3d segmentations, 2d and 3d landmarks, 3d eyeball, gaze vector, and eye movement types. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 367–375. IEEE, 2021.

Fabio Galasso, Naveen Shankar Nagaraja, Tatiana Jimenez Cardenas, Thomas Brox, and Bernt Schiele. A unified video segmentation benchmark: Annotation, metrics and analysis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3527–3534, 2013.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013.

Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. *arXiv preprint arXiv:2305.20091*, 2023.

Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3749–3761, 2022.

Adam W Harley, Zhaoyuan Fang, and Katerina Fragkiadaki. Particle video revisited: Tracking through occlusions using point trajectories. In *ECCV*, 2022.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

Shiyu Hu, Xin Zhao, Lianghua Huang, and Kaiqi Huang. Global instance tracking: Locating target more like humans. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):576–592, 2022.

Y.-T. Hu, H.-S. Chen, K. Hui, J.-B. Huang, and A. G. Schwing. SAIL-VOS: Semantic Amodal Instance Level Video Object Segmentation – A Synthetic Dataset and Baselines. In *Proc. CVPR*, 2019.

Chang Huang, Bo Wu, and Ramakant Nevatia. Robust object tracking by hierarchical association of detection responses. In *ECCV*. 2008.

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.

Daniel Joska, Liam Clark, Naoya Muramatsu, Ricardo Jericevich, Fred Nicolls, Alexander Mathis, Mackenzie W Mathis, and Amir Patel. Acinoset: a 3d pose estimation dataset and baseline models for cheetahs in the wild. In *2021 IEEE international conference on robotics and automation (ICRA)*, pages 13901–13908. IEEE, 2021.

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. *arXiv preprint arXiv:2307.07635*, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023.

Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, 2019.

Matej Kristan, Aleš Leonardis, Jiri Matas, Michael Felsberg, Roman Pflugfelder, Luka Čehovin, Tomas Vojir, Gustav Häger, Alan Lukežič, and Gustavo Fernandez. The visual object tracking vot2016 challenge results. Springer, 2016.

Thijs P Kuipers, Devanshu Arya, and Deepak K Gupta. Hard occlusions in visual object tracking. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 299–314. Springer, 2020.

Jessy Lauer, Mu Zhou, Shaokai Ye, William Menegas, Steffen Schneider, Tanmay Nath, Mohammed Mostafizur Rahman, Valentina Di Santo, Daniel Soberanes, Guoping Feng, et al. Multi-animal pose estimation, identification and tracking with deeplabcut. *Nature Methods*, 19(4):496–504, 2022.

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.

Xiaoxiao Li and Chen Change Loy. Video object segmentation with joint re-identification and attention-aware mask propagation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 90–105, 2018.

Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European Conference on Computer Vision*, pages 280–296. Springer, 2022.

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.

LILA. New zealand wildlife thermal imaging. `https://lila.science/datasets/new-zealand-wildlife-thermal-imaging/`. Accessed: 22-May-2024.

Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. *CVPR*, 2023.

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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

Nathan Louis, Luowei Zhou, Steven J Yule, Roger D Dias, Milisa Manojlovich, Francis D Pagani, Donald S Likosky, and Jason J Corso. Temporally guided articulated hand pose tracking in surgical videos. *International Journal of Computer Assisted Radiology and Surgery*, 18(1):117–125, 2023.

Alexander Mathis, Thomas Biasi, Steffen Schneider, Mert Yuksekgonul, Byron Rogers, Matthias Bethge, and Mackenzie W. Mathis. Pretraining boosts out-of-domain robustness for pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1859–1868, 2021.

N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, 2016.

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.

Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 548–564. Springer, 2020.

Abhinav Moudgil and Vineet Gandhi. Long-term visual object tracking benchmark. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part II 14*, pages 629–645. Springer, 2019.

Matthias Mueller, Neil Smith, and Bernard Ghanem. A benchmark and simulator for uav tracking. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 445–461. Springer, 2016.

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.

Elena Nicora, Gaurvi Goyal, Nicoletta Noceti, Alessia Vignolo, Alessandra Sciutti, and Francesca Odone. The moca dataset, kinematic and multi-view visual streams of fine-grained cooking actions. *Scientific Data*, 7(1): 432, 2020.

Peter Ochs, Jitendra Malik, and Thomas Brox. Segmentation of moving objects by long term video analysis. *IEEE transactions on pattern analysis and machine intelligence*, 36(6):1187–1200, 2013.

Malte Pedersen, Joakim Bruslund Haurum, Stefan Hein Bengtson, and Thomas B Moeslund. 3d-zef: A 3d zebrafish tracking benchmark dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2426–2436, 2020.

Federico Perazzi, Anna Khoreva, Rodrigo Benenson, Bernt Schiele, and Alexander Sorkine-Hornung. Learning video object segmentation from static images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2663–2672, 2017.

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.

Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip Torr, and Song Bai. Occluded video instance segmentation: A benchmark. *International Journal of Computer Vision*, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE TPAMI*, 44(3):1623–1637, 2020.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Gozde Sahin and Laurent Itti. Hoot: Heavy occlusions in object tracking benchmark. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4830–4839, 2023.

Peize Sun, Jinkun Cao, Yi Jiang, Zehuan Yuan, Song Bai, Kris Kitani, and Ping Luo. Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Ramana Sundararaman, Cedric De Almeida Braga, Eric Marchand, and Julien Pettre. Tracking pedestrian heads in dense crowd. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3865–3875, 2021.

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, pages 402–419. Springer, 2020.

Pavel Tokmakov, Jie Li, and Adrien Gaidon. Breaking the "object" in video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22836–22845, 2023.

Ilya O. Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. MLP-mixer: An all-mlp architecture for vision. *ArXiv*, abs/2105.01601, 2021.

Basile Van Hoorick, Pavel Tokmakov, Simon Stent, Jie Li, and Carl Vondrick. Tracking through containers and occluders in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13802–13812, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Paul Voigtlaender and Bastian Leibe. Online adaptation of convolutional neural networks for video object segmentation. *arXiv preprint arXiv:1706.09364*, 2017.

Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. Mots: Multi-object tracking and segmentation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 7942–7951, 2019.

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.

Haochen Wang, Cilin Yan, Shuai Wang, Xiaolong Jiang, Xu Tang, Yao Hu, Weidi Xie, and Efstratios Gavves. Towards open-vocabulary video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4057–4066, 2023.

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.

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.

Bowen Wen, Chaitanya Mitash, Baozhang Ren, and Kostas E. Bekris. se(3)-tracknet: Data-driven 6d pose tracking by calibrating image residuals in synthetic domains. *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020.

Meihong Wu, Xiaoyan Cao, Xiaoyu Cao, and Shihui Guo. A dataset of ant colonies motion trajectories in indoor and outdoor scenes for social cluster behavior study. *arXiv preprint arXiv:2204.04380*, 2022.

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.

Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross Girshick. Early convolutions help transformers see better. *Advances in neural information processing systems*, 34:30392–30400, 2021.

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.

Linjie Yang, Yuchen Fan, and Ning Xu. The 4th large-scale video object segmentation challenge - video object segmentation track, 2022.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020.

Libo Zhang, Junyuan Gao, Zhen Xiao, and Heng Fan. Animaltrack: A large-scale benchmark for multi-animal tracking in the wild. *arXiv preprint arXiv:2205.00158*, 2022.

Yang Zheng, Adam W. Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J. Guibas. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In *ICCV*, 2023.

Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Simple multi-dataset detection. In *CVPR*, 2022.

Yabin Zhu, Chenglong Li, Yao Liu, Xiao Wang, Jin Tang, Bin Luo, and Zhixiang Huang. Tiny object tracking: A large-scale dataset and a baseline. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.