Matching Anything by Segmenting Anything

Siyuan Li[1] Lei Ke[1] Martin Danelljan[1] Luigi Piccinelli [1]
Mattia Segu[1] Luc Van Gool[1,2] Fisher Yu[1]

[1]ETH Zürich [2]INSAIT

Abstract

The robust association of the same objects across video frames in complex scenes is crucial for many applications, especially multiple object tracking (MOT). Current methods predominantly rely on labeled domain-specific video datasets, which limits the cross-domain generalization of learned similarity embeddings. We propose MASA, a novel method for robust instance association learning, capable of matching any objects within videos across diverse domains without tracking labels. Leveraging the rich object segmentation from the Segment Anything Model (SAM), MASA learns instance-level correspondence through exhaustive data transformations. We treat the SAM outputs as dense object region proposals and learn to match those regions from a vast image collection. We further design a universal MASA adapter which can work in tandem with foundational segmentation or detection models and enable them to track any detected objects. Those combinations present strong zero-shot tracking ability in complex domains. Extensive tests on multiple challenging MOT and MOTS benchmarks indicate that the proposed method, using only unlabeled static images, achieves even better performance than state-of-the-art methods trained with fully annotated in-domain video sequences, in zero-shot association. Our code is available at github.com/siyuanliii/masa.

1. Introduction

Multiple object tracking (MOT) is one of the fundamental problems in computer vision. It plays a pivotal role in numerous robotics systems such as autonomous driving. Tracking requires both detecting the objects of interest in videos and associating them across frames. While recent advancements in segmentation and detection foundation models [28, 30, 35, 63, 71] have demonstrated an exceptional ability to detect and segment any objects, associating those objects in videos remains challenging. Recent successful multiple object tracking approaches [31, 59] have emphasized the importance of learning discriminative instance embeddings for accurate association. Some [41] even argued that it is the only necessary tracking component besides detection.

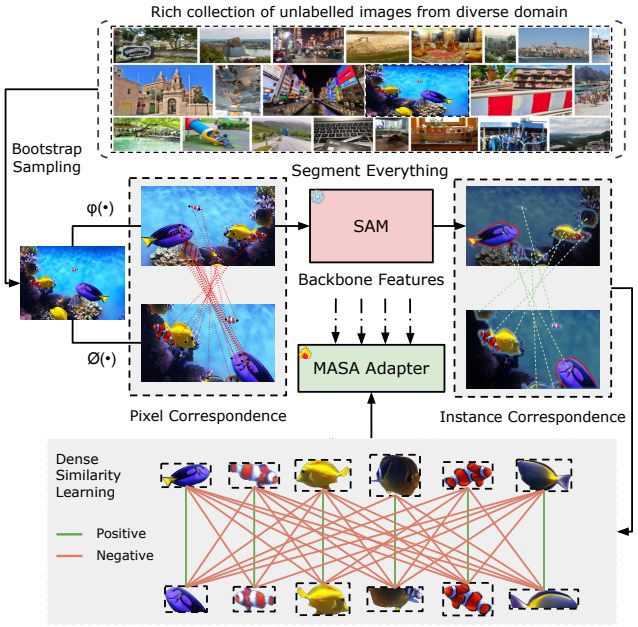


Figure 1. Given an unlabeled image from any domain, we apply strong augmentations, $\varphi(\cdot)$ and $\phi(\cdot)$, to the image, generating two different views with automatically established pixel correspondences. Then, we leverage the rich object-level information encoded by the foundation segmentation model SAM to transfer the pixel-level to dense instance-level correspondence. Such correspondences enable us to utilize a diverse collection of unlabelled images to train a universal tracking adapter atop any segmentation or detection foundation models e.g. SAM. This adapter empowers the foundational models to track any objects they have detected, and shows strong zero-shot tracking ability in complex domains.

However, learning effective object association usually requires a significant amount of annotated data. While collecting detection labels on a diverse set of images is laborious, obtaining tracking labels on videos is even more challenging. Consequently, current MOT datasets mostly focus on objects from a specific domain with a small number of

fixed categories or a limited number of labeled frames.

Training on those datasets limits the generalizability of tracking models to different domains and novel concepts. Although recent studies [30, 35, 71] have made successful attempts to address the model generalization issue for object detection and segmentation, the path to learning a universal association model for tracking any objects is still unclear.

Our goal is to develop a method capable of matching any objects or regions. We aim to integrate this generalizable tracking capability with any detection and segmentation methods to help them track any object they have detected. A primary challenge is acquiring matching supervision for general objects across diverse domains, without incurring substantial labelling costs.

To this end, we propose the **M**atching **A**nything by **S**egmenting **A**nything (MASA) pipeline to learn object-level associations from unlabeled images of any domain. Figure 1 presents an overview of our MASA pipeline. We leverage the rich object appearance and shape information encoded by the foundation segmentation SAM, combined with extensive data transformation, to establish strong instance correspondence.

Applying different geometric transformations to the same image gives automatic pixel-level correspondence in two views from the same image. SAM's segmentation ability allows for the automatic grouping of pixels from the same instance, facilitating the conversion of pixel-level to instance-level correspondence. This process creates a self-supervision signal for learning discriminative object representation, utilizing dense similarity learning between view pairs. Our training strategy enables us to use a rich collection of raw images from diverse domains, demonstrating that such automatic self-training on diverse raw images provides excellent zero-shot multiple object tracking performance, even surpassing models reliant on in-domain video annotations for association learning.

Beyond the self-training pipeline, we further build a universal tracking adapter — MASA adapter, to empower any existing open-world segmentation and detection foundation models such as SAM [30], Detic [71] and Grounding-DINO [35] for tracking any objects they have detected. To preserve their original segmentation and detection ability, we freeze their original backbone and add the MASA adapter on the top.

Moreover, we propose a multi-task training pipeline that jointly performs the distillation of SAM's detection knowledge and instance similarity learning. This approach allows us to learn the object's location, shape and appearance prior of SAM, and simulate real detection proposals during contrastive similarity learning. This pipeline further improves the generalization capabilities of our tracking features. Additionally, our learned detection head speeds up the original SAM dense uniform point proposals for segmenting everything by over tenfold, crucial for tracking applications.

We evaluate MASA on multiple challenging benchmarks, including TAO MOT [16], Open-vocabulary MOT [32], MOT and MOTS on BDD100K [64], and UVO [49]. Extensive experiments indicate that compared with state-of-the-art object tracking approaches trained on thoroughly in-domain labeled videos, our method achieves on-par or even better association performance, using a single model with the same model parameters and testing in zero-shot association settings.

2. Related Work

2.1. Learning Instance-level Association

Learning robust instance-level correspondence is crucial to object tracking. Existing approaches can be divided into self-supervised [51] and supervised [8, 31, 39, 41, 50, 56, 58, 59, 65, 69] strategies. Specifically, as a representative self-supervised method, UniTrack [51] attempts to directly use off-the-shelf self-supervised representations [10, 57] for association. Despite competitive results on some benchmarks [40], these methods cannot fully exploit instance-level training data, limiting their performance in challenging scenarios. In contrast, supervised methods train discriminative instance embeddings on frame pairs, by contrastive learning. Although achieving superior performance on challenging benchmarks [16, 32, 36, 44, 64], these methods rely on tremendous in-domain labeled video data. Several methods [2, 19, 32, 70, 72] learn tracking signals from static images but still require substantial fine-grained instance annotations in specific domains or post-hoc test-time adaptation [47], limiting their ability for cross-domain generalization. To tackle these problems, we exploit the exhaustive object shape, and appearance information encoded by SAM to learn universal instance matching, purely from unlabeled images. Our learned representation shows exceptional zero-shot association ability across diverse domains.

2.2. Segment and Track Anything Models

Deva [13], TAM [60] and SAM-Track [14] integrate SAM [30] with video object segmentation (VOS) approaches (such as XMem [12] and DeAOT [62]) to enable an interactive pipeline for tracking any object, where SAM is mainly used for mask initialization/correction and XMem/DeAOT handle the tracking and prediction. SAM-PT [43] combines SAM with point-tracking methods such as [23, 25, 48] to perform tracking. However, all those approaches face limitations, such as poor mask propagation quality due to domain gaps and the inability to handle multiple diverse objects or rapid objects entry and exit, common in scenarios like autonomous driving. Our work focuses on a different direction. Instead of building an interactive tracking pipeline or using off-the-shelf VOS or point-

based trackers, we focus on learning universal association modules by leveraging SAM's rich instance segmentation knowledge.

3. Method

3.1. Preliminaries: SAM

SAM [30] is composed of three modules: **(a)** Image encoder: A heavy ViT-based backbone for feature extraction. **(b)** Prompt encoder: Modeling the positional information from the interactive points, box, or mask prompts. **(c)** Mask decoder: A transformer-based decoder takes both the extracted image embedding with the concatenated output and prompt tokens for final mask prediction. To generate all potential mask proposals, SAM adopts densely sampled regular grids as point anchors and generates mask predictions for each point prompt. The complete pipeline includes patch cropping with greedy box-based NMS, three-step filtering, and heavy post-processing on masks. For more details on SAM's everything mode, we refer readers to [30].

3.2. Matching Anything by Segmenting Anything

Our method consists of two key components. First, based on SAM, we develop a new pipeline: MASA (Section 3.2.1). With this pipeline, we construct exhaustive supervision for dense instance-level correspondence from a rich collection of unlabeled images. It enables us to learn strong discriminative instance representations to track any objects, without requiring any video annotations. Second, we introduce a universal MASA adapter (Section 3.2.2) to effectively transform the features from a frozen detection or segmentation backbone for learning generalizable instance appearance representations. As a byproduct, the distillation branch of the MASA adapter can also significantly improve the efficiency of segmenting everything. Besides, we also construct a unified model to jointly detect / segment and track anything (Section 3.2.3). Our complete training pipeline is shown in Figure 2.

3.2.1. MASA Pipeline

To learn instance-level correspondence, previous works [31, 41, 58, 59, 69] heavily relied on manually labeled in-domain video data. However, current video datasets [5, 40, 64] contain only a limited range of fixed categories. This limited diversity in datasets leads to learning appearance embeddings that are tailored to specific domains, posing challenges in their universal generalization.

UniTrack [51] demonstrates that universal appearance features can be learned through contrastive self-supervised learning techniques [7, 10, 57] from raw images or videos. These representations, harnessing the diversity of a large volume of unlabeled images, can generalize across different tracking domains. However, they often depend on clean, object-centered images, such as those in ImageNet [46], or videos like DAVIS17 [42], and focus on frame-level similarities. This focus causes them to fail in fully leveraging instance information, leading to difficulties in learning discriminative instance representations in complex domains with multiple instances, as demonstrated in Table 7.

To address these issues, we propose the MASA training pipeline. Our core idea is to increase diversity from two perspectives: training *image diversity* and *instance diversity*. As shown in Figure 1, we first construct a rich collection of raw images from diverse domains to prevent learning domain-specific features. These images also contain a rich number of instances in complex environments to enhance instance diversity. Given an image I, we simulate appearance changes in videos by adopting two different augmentations on the same image. By applying strong data augmentations $\varphi(I)$ and $\phi(I)$, we construct two different views V_1 and V_2 of I, thereby automatically obtaining pixel-level correspondence.

If the image is clean and contains only one instance, such as those in ImageNet, frame-level similarity can be applied as in [10, 57, 67]. However, with multiple instances, we need to further mine the instance information contained in such raw images. The foundational segmentation model SAM [30] offers us this capability. SAM automatically groups pixels belonging to the same instances and also provides the shape and boundary information of detected instances, valuable for learning discriminative features.

Since we construct the dataset by selecting images with multiple instances, SAM's exhaustive segmentation of the entire images automatically yields a dense and diverse collection of instance proposals Q. With pixel-level correspondences established, applying the same $\phi(\cdot)$ and $\varphi(\cdot)$ to Q transfers pixel-level correspondence to dense instance-level correspondence. This self-supervision signal enables us to use the contrastive learning formula from [29, 31, 41] to learn a discriminative contrastive embedding space:

$$\mathcal{L}_{\mathcal{C}} = -\sum_{q \in Q} \log \frac{e^{\frac{\mathrm{sim}(q,q^+)}{\tau}}}{e^{\frac{\mathrm{sim}(q,q^+)}{\tau}} + \sum_{q^- \in Q^-} e^{\frac{\mathrm{sim}(q,q^-)}{\tau}}},$$

Here, q^+ and q^- denote the positive and negative samples to q, respectively. Positive samples are the same instance proposals being applied different $\phi(\cdot)$ and $\varphi(\cdot)$. Negative samples are from different instances. Furthermore, $\mathrm{sim}(\cdot)$ denotes the cosine similarity and τ is a temperature parameter, set to 0.07 in our experiments. This contrastive learning formula pushes object embeddings belonging to the same instance closer while distancing embeddings from different instances. As demonstrated by existing works [9, 41], negative samples are crucial for learning discriminative representations. Under the contrastive learning paradigm, the dense proposals generated by SAM natu-

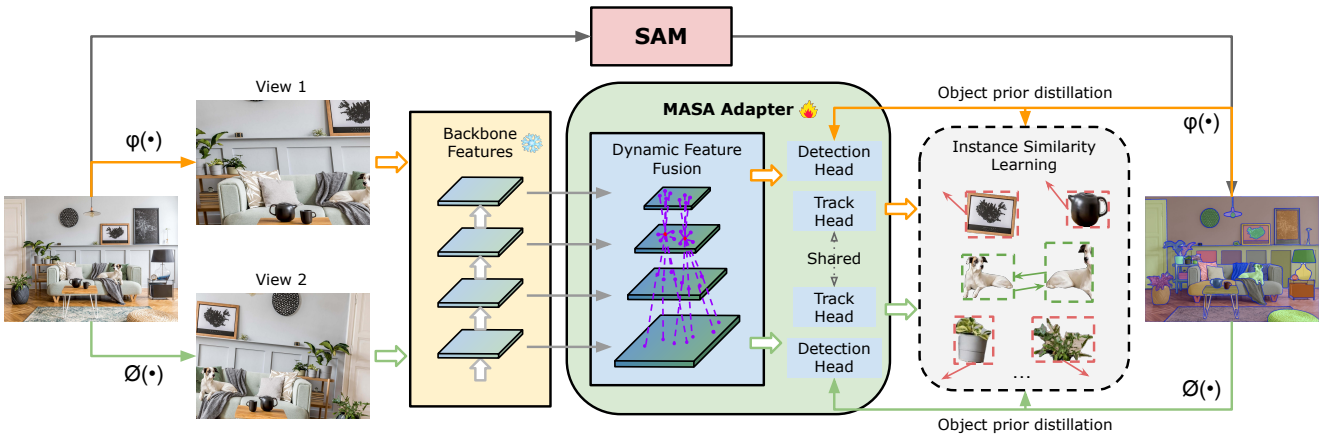


Figure 2. MASA training pipeline. Given an unlabeled image from any domain, SAM automatically generates exhaustive instance masks for it. Then we apply strong augmentations, $\phi(\cdot)$ and $\varphi(\cdot)$, to the original image and exhaustive instance segmentation, obtaining two different views as the inputs of our model. We train our MASA adapter by joint distillation of SAM's detection knowledge and instance similarity learning. Better view in color with zoom-in.

rally provide more negative samples, thus enhancing learning better instance representation for association.

3.2.2. MASA Adapter

We introduce the MASA adapter, designed to extend the open-world segmentation and detection models (such as SAM [30], Detic [71], and Grounding-DINO [35]) to track any detected objects. The MASA adapter operates in conjunction with frozen backbone features from these foundational models, ensuring their original detection and segmentation capabilities are preserved. However, as not all pretrained features are inherently discriminative for tracking, we first transform these frozen backbone features into new features more suitable for tracking.

Given the diversity in shapes and sizes of objects, we construct a multi-scale feature pyramid. For hierarchical backbones like the Swin Transformer [37] in Detic and Grounding DINO, we directly employ FPN [34]. For SAM, which utilizes a plain ViT [17] backbone, we use Transpose Convolution and MaxPooling to upsample and downsample the single-scale features of stride $16\times$ to produce hierarchical features with scale ratios of $\frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}$. To effectively learn discriminative features for different instances, it's essential that objects in one location are aware of the appearances of instances in other locations. Hence, we use deformable convolution to generate dynamic offsets and aggregate information across spatial locations and feature levels as [15]:

$$F(p) = \frac{1}{L}\sum_{j=1}^{L}\sum_{k=1}^{K} w_k \cdot F^j(p + p_k + \Delta p_k^j) \cdot \Delta m_k^j, \quad (1)$$

where L represents the feature level, K is the number of sampling locations for a convolutional kernel, w_k and p_k are the weight and predefined offset for the k-th location, respectively, and Δp_k^j and Δm_k^j are the learnable offset and modulation factor for the k-th location at the j-th feature level. For SAM-based models, we additionally use task-aware attention and scale-aware attention from Dyhead [15], since the detection performance is important for accurate auto mask generation as in Figure 3 (b). After acquiring the transformed feature map, we extract instance-level features by applying RoI-Align [24] to the visual features F, followed by processing with a lightweight track head comprising 4 convolutional layers and 1 fully connected layer to generate instance embeddings.

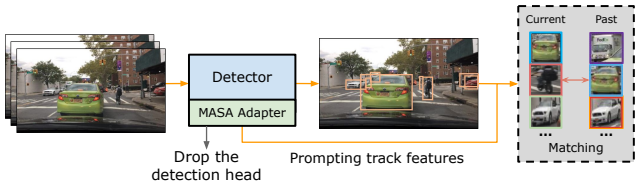
Additionally, we introduce an object prior distillation branch as an auxiliary task during training. This branch employs a standard RCNN [45] detection head to learn bounding boxes that tightly encompass SAM's mask predictions for each instance. It effectively learns exhaustive object location and shape knowledge from SAM and distils this information into the transformed feature representations. This design not only strengthens the features of the MASA adapter, resulting in improved association performance but also accelerates SAM's everything mode by directly providing the predicted box prompts.

The MASA adapter is optimized using a combination of detection and contrastive losses as defined in Section 3.2.1: $\mathcal{L} = \mathcal{L}_{\text{det}} + \mathcal{L}_C$. The detection loss is identical to that in [45].

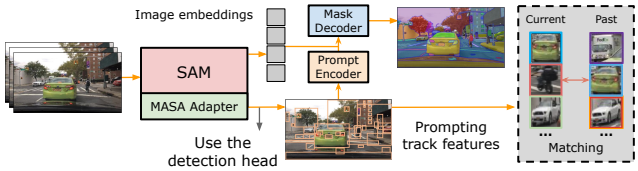
3.2.3. Inference

Figure 3 shows the test pipeline with our unified models.

Detect and Track Anything When we integrate the MASA adapter with object detectors, we remove the MASA detection head that was learned during training. The MASA adapter then solely serves as a tracker. The detectors predict the bounding boxes, and then they are utilized to prompt the MASA adapter, which retrieves corresponding tracking features for instance matching. We use a simple bi-softmax nearest neighbor search for accurate instance matching, as



(a) Detect and Track Anything: Inference with detection model such as Detic



(b) Segment and Track Anything: Fast inference with SAM

Figure 3. The inference pipeline of our unified methods.

Table 1. SOTA comparison on TAO TETA Benchmark [31]. † indicates using the same detection observations. Our zero-shot models can achieve better performance compared with the state-of-the-art fully supervised methods using the same detection observations. The performance can be even better when using original detections from our unified model.

Method	TETA	LocA	AssocA	ClsA
Fully-supervised, in-domain				
SORT [4]	24.9	48.1	14.3	12.1
Tracktor [3]	24.2	47.4	13.0	12.1
Tracktor++ [3]	28.0	49.0	22.8	12.1
DeepSORT [52]	26.0	48.4	17.5	12.1
AOA [18]	25.3	23.4	30.6	**21.9**
QDTrack [41]	30.0	50.5	27.4	12.1
TETer [31]†	34.6	**52.1**	36.7	15.0
Self-supervised, zero-shot				
Ours-Detic†	34.7	51.9	36.4	15.8
Ours-Grounding-DINO†	**34.9**	51.8	**37.6**	15.4
Ours-SAM-B†	34.5	51.8	36.6	15.1
Ours-SAM-H†	34.5	51.8	36.4	15.4
Ours-Detic	**46.3**	**65.8**	**44.1**	**28.9**

illustrated in Section J.4 of the Appendix.

Segment and Track Anything With SAM, we keep the detection head. We use it to predict all potential objects within a scene, forwarding box predictions as prompts to both the SAM mask decoder and the MASA adapter for segmenting and tracking everything. The predicted box prompts omit the need for the heavy post-processing illustrated in the original SAM's everything mode, therefore, significantly speeding up the auto mask generation of SAM.

Testing with Given Observations When detections are obtained from sources other than the one the MASA adapter is built upon, our MASA adapter serves as a tracking feature provider. We directly utilize the provided bounding boxes as prompts to extract tracking features from our MASA adapter through the ROI-Align [24] operation.

4. Experiments

We perform experiments on multiple challenging MOT/MOTS benchmarks with diverse domains.

Table 2. State-of-the-art comparison on Open-vocabulary MOT benchmark [32]. All methods are trained with base annotations.

Method	Base				Novel			
	TETA	LocA	AssocA	ClsA	TETA	LocA	AssocA	ClsA
DeepSORT [52]	28.4	52.5	15.6	17.0	24.5	49.2	15.3	9.0
Tracktor++ [3]	29.6	52.4	19.6	16.9	25.7	50.1	18.9	8.1
Bytetrack [68]	29.5	51.7	19.7	17.2	25.4	49.4	18.1	8.7
OC-SORT [6]	30.0	53.3	23.5	13.3	26.7	51.5	21.6	7.1
OVTrack [32]	36.3	53.9	36.3	18.7	32.0	51.4	33.2	11.4
Ours-Detic	**47.0**	**66.0**	**44.5**	**30.5**	**40.8**	**64.4**	**41.2**	**17.0**

Table 3. Comparsion on TAO Track mAP benchmark.† indicates using same detections with GTR. Note that GTR is an offline tracking method while ours is online.

Method	Track mAP50	Track mAP75	Track mAP
Fully-supervised, in-domain			
SORT-TAO [16]	13.2	-	-
QDTrack [41]	15.9	5.0	10.6
TAC [54]	17.7	5.8	7.3
BIV [53]	21.6	10.4	16.1
GTR† [72]	22.5	-	-
Self-supervised, zero-shot			
Ours-Detic†	22.0	12.2	17.1
Ours-Grounding-DINO†	22.8	12.3	17.6
Ours-SAM-B†	**23.9**	**13.0**	**18.4**
Ours-SAM-H†	22.9	12.1	17.5
Ours-Detic	**30.9**	**18.0**	**24.4**

Table 4. State-of-the-art comparison on BDD MOTS. † represents that we provide the same detection observations. AssocA, mIDF1, and IDF1 mainly evaluate the association quality. MASA achieves the best results on all metrics.

Method	mIDF1↑	AssocA↑	TETA↑	mMOTSA↑	mHOTA↑
Fully-supervised, in-domain					
MaskTrackRCNN [61]	26.2	-	-	12.3	-
STEm-Seg [1]	25.4	-	-	12.2	-
QDTrack-mots [41]	40.8	-	-	22.5	-
PCAN [26]	45.1	46.7	46.8	27.4	35.9
VMT [27]	45.7	47.3	47.1	28.7	36.6
Unicorn [58]	44.2	-	-	29.6	-
UNINEXT-H [59]†	48.5	53.2	53.6	35.7	40.6
Self-supervised, zero-shot					
Ours-Detic†	49.5	53.5	54.4	**36.4**	40.2
Ours-Grounding-DINO†	48.6	52.3	54.0	36.1	40.0
Ours-SAM-B†	49.2	53.9	**54.8**	35.2	40.7
Ours-SAM-H†	**49.7**	**54.5**	54.7	35.8	**40.8**

Table 5. State-of-the-art comparison on BDD MOT val set. † represents that we provide the same detection observations.

Method	mIDF1↑	IDF1↑	TETA↑	AssocA↑	mMOTA↑
Fully-supervised, in-domain					
QDTrack [41]	50.8	71.5	47.8	48.5	36.6
TETer [31]	53.3	71.1	50.8	**52.9**	39.1
MOTR [65]	54.0	65.8	-	-	32.3
Unicorn [58]	54.0	71.3	-	-	41.2
UNINEXT-H [59]	**56.7**	69.9	-	-	44.2
ByteTrack [68]†	54.8	70.4	**55.7**	51.5	**45.5**
Self-supervised, zero-shot					
Ours-Detic†	55.8	71.3	54.4	**52.9**	44.6
Ours-Grounding-DINO†	55.6	**71.7**	54.5	52.7	44.5
Ours-SAM-B†	55.6	71.6	54.0	52.6	44.1
Ours-SAM-H†	55.3	**71.7**	54.2	51.9	44.5

4.1. Experimental Setup

TAO MOT TAO dataset [16] is designed to track a diverse range of objects, encompassing over 800 categories, making it the most diverse MOT dataset with the largest class collection to date. It contains 500, 988, and 1,419 videos annotated at 1 FPS in the train, validation, and test sets, respectively. We report performances on the validation set. TAO comprises several benchmarks, each highlighting different characteristics and requirements. The TAO TETA benchmark [31] emphasizes association by rewarding trackers that produce clean trajectories with no overlaps. Conversely, the TAO Track mAP benchmark [16] values particularly the classification of trajectories, and does not heavily penalize overlapping trajectories. The open-vocabulary MOT benchmark [32] requires trackers to avoid training with annotations from novel classes, focusing on the generalization ability to track novel categories.

BDD100K MOT [64] requires trackers to track common objects in autonomous driving scenarios. The dataset is annotated at 5 FPS with 200 videos in the validation set.

BDD100K MOTS Different from BDD100K MOT, BDD100K MOTS [64] requires trackers to track and segment objects simultaneously, evaluating tracking performance on masks. There are 154 videos for training, 32 videos for validation, and 37 videos for testing.

UVO [49] is a challenging benchmark for open-world instance segmentation in videos. Compared with previous video-level object segmentation datasets [61], it annotates much more diverse instances. UVO has two evaluation tracks, an image track, and a video track. We evaluate all methods on the UVOv0.5 validation set.

Evaluation Metrics As analyzed in previous works [31], traditional tracking metrics like mMOTA [64], and track mAP [16] can be misleading, particularly in long-tail scenarios, due to their high sensitivity to classification. To address this issue, [31] introduced TETA, a new tracking metric that decomposes into three separate components: AssocA, LocA, and ClsA, reflecting the accuracy of association, localization, and classification, respectively. In standard MOT benchmarks, to ensure a fair comparison of trackers' association abilities, we adopt the same detection observations used by leading state-of-the-art trackers. Therefore, our focus is primarily on **association-related** metrics like **AssocA**, **mIDF1**, and **IDF1**. Additionally, when evaluating our unified models, we consider the full spectrum of metrics to capture their comprehensive capabilities. Particularly for open-world segmentation on UVO, our emphasis is on AR100 and Track AR100 metrics in the image and video levels. This is due to the fact that SAM often segments every part of an object, whereas UVO lacks such detailed annotations, making traditional AP evaluations less accurate.

Training Data SA-1B [30] consists of 11M diverse, high-resolution images, containing diverse scenarios with multiple object interactions in complex environments. We subsample the SA-1B raw images to construct a training set of 500K images, SA-1B-500K.

Implementation Details For our models, we utilize the official weights of SAM [30], Detic, and Grounding-DINO, ensuring that all components of these models remain frozen during the training phase. Specifically, we employ SAM with both ViT-Base and ViT-Huge backbones, and Detic and Grounding-DINO are used with the SwinB backbone. We train the models with bootstrapping sampling for 200,000 images per epoch, with a batch size of 128. We use SGD with an initial learning rate of 0.04, coupled with a step policy for learning rate decay. Momentum and weight decay parameters are set to 0.9 and 1e-4. Our training spans 12 epochs, with the learning rate being reduced at the 8th and 11th epochs. For data augmentation, we use random affine, MixUp [66], and Large-scale Jittering [20], in addition to standard practices like flipping, color jittering, and random cropping. More details are provided in Section J of the Appendix.

4.2. State-of-the-Art Comparison

We evaluate our methods in two ways. Firstly, to accurately assess the association ability of our method, we always provide the same detection observations as current state-of-the-art methods in standard MOT benchmarks. Secondly, to evaluate the integrated abilities of our unified models, we follow this protocol: for SAM-based models, we evaluate on the open-world video segmentation dataset UVO. For the detectors-based models, we evaluated on the Open-vocabulary MOT benchmark [32]. We also report the scores on TAO TETA and TAO TrackmAP benchmarks. Note that we perform zero-shot association tests for all our variants, and use the same weights across all benchmarks.

TAO TETA We use the same observations as TETer-SwinT [31]. As shown in Table 1, our method with Grounding-DINO's backbone performs the best, in the zero-shot setting, without training on any in-domain labeled videos, on both AssocA and TETA. We also test our unified Detic model which jointly outputs the detection and tracking results. It outperforms all other methods significantly and achieves the new state-of-the-art. It demonstrates our method can couple well with current detection foundation models and transfer their strong detection ability into tracking.

Open-vocabulary MOT Similar to the open-vocabulary object detection task [21], open-vocabulary MOT [32] stipulates that methods should only use the frequent and common classes annotations from LVIS [22] for training, treating the rare classes as novel. We evaluated our unified 'detect and track anything' model Detic, which was trained exclusively with base class annotations. Table 2 shows

our unified Detic model outperforms existing models on all metrics across both base and novel splits, and it achieves this significant lead despite our tracker being trained solely with out-of-domain, unlabeled images.

TAO Track mAP We use the same observations as GTR [72]. As shown in Table 3, our method with SAM-B performs the best (Track mAP50 of 23.9) given the same detections. Most of our models outperform the current state-of-the-art GTR, which is an offline method that utilizes future information for association. In contrast, our methods conduct tracking in an online fashion and test in a zero-shot setting. Our unified Detic model again, achieves the new state-of-the-art by outperforming GTR by a large margin.

BDD100K MOTS We use the same observations as the state-of-the-art method, UNINEXT-H [59] and perform zero-shot association test on BDD100K MOTS benchmark. As shown in Table 4, our method achieves the best association performance (mIDF1 of 49.7 and AssocA of 54.5) among all approaches. This demonstrates the superiority of the instance embeddings learned by our method.

BDD100K MOT As shown in Table 5, given the same observations as ByteTrack [68], our method achieves the best IDF1 of 71.7 and AssocA 52.9. Compared with state-of-the-art ByteTrack [68], our method also achieves better association performance, being about 1.4% higher on both IDF1 and AssocA, without using any BDD images for training. ByteTrack additionally selects low-confidence boxes and adds them to the tracklets, resulting in a better mMOTA score which prioritises detection performance [38].

UVO VIS We perform zero-shot tests for our unified 'segment and track anything' model based on SAM. We directly use the box prompts from the MASA detection head for faster segmenting everything. As shown in Figure 4a, our method achieves the best performance on both image and video tracks, outperforming its counterparts by a large margin. Besides, we also compare our method with SAM's default auto mask segmentation. As shown in Figure 4b, as the inference time increases, AR100 of our method grows much faster than SAM due to the distillate detection branch. The upper bound AR100 of our method with ViT-Base backbone even surpasses SAM by 10%. Besides, when achieving the same AR100, our method is about $10\times$ faster than SAM. This stems from the fact that our method learns a strong object prior to capturing potential objects with a small number of sparse proposals. However, to segment everything, SAM has to sample about $1k$ points evenly, which is inflexible and inefficient, while also relying on hand-crafted complex post-processing methods.

Compare with VOS Methods We evaluated the VOS-based method Deva [13], which integrates XMem [12] for tracking multiple objects and SAM-PT [43], which uses point-tracking. To ensure a fair comparison, we provide the same observations on BDD MOTS, TAO TETA and UVO

(a) Quantitative results on UVO [49] dataset.

Track	Method	AR100
image	Mask R-CNN [24]	41.3
	Ours-SAM-B	43.7
	Ours-SAM-H	**50.8**
video	*Fully-supervised, in-domain*	
	MaskTrack R-CNN [61]	17.2
	QDTrack [41]	20.9
	SeqFormer [55]	21.3
	Zero-shot test	
	Ours-SAM-B	24.9
	Ours-SAM-H	**28.4**

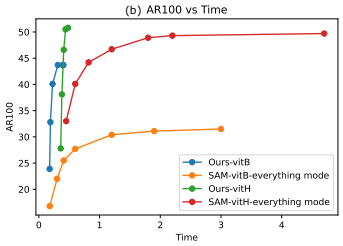


Figure 4. Comparison on the UVO [49] dataset. (a) We evaluate class-agnostic object detection and video object tracking results with our MASA. Both object localization and association achieve promising performance compared with previous in-domain training methods. (b) We compare the inference time (s) with the original SAM by sampling different numbers of prompt points. Our detection head learns to localize all the potential objects effectively.

Table 6. Compare with VOS Methods. † represents that we provide the same detection observation as inputs.

Method	BDD MOTS AssocA	BDD MOTS TETA	TAO AssocA	UVO AR100
SAM-PT [43]	-	-	-	31.8
Deva† [13]	46.8	32.1	22.4	36
Ours-SAM-H†	**54.5**	**54.7**	**36.4**	**37.5**

benchmarks. For UVO, we use SAM's auto-mask generation to generate masks first, then we resolve the overlapping masks following the heuristic in Deva [13] and use Deva to generate per-frame observations. Table 6 shows that our method outperforms Deva across all benchmarks. Notably, on the autonomous driving BDD100K benchmark, where objects frequently enter and exit the scene, VOS-based methods like Deva are prone to a significant increase in false positives. This is reflected in the TETA scores, where such errors are heavily penalized. Additionally, Deva struggles with overlapping predictions, a common issue with current detection models. We provide a more in-depth analysis in Section H of the Appendix.

Compare with Self-supervised Methods We further compare our approach with self-supervised methods aimed at learning universal appearance features from raw images or videos. To ensure a fair comparison, we train all methods using a mix of BDD and COCO raw images. Specifically, for VFS, we utilize raw videos from BDD. We employ a ResNet-50 model for VFS [57] and MoCov2 [10], and a ViT-B model for DINO [7], following the association tracking strategy outlined in UniTrack [51]. Additionally, we ensure that detection observations are identical across all models. Table 7 demonstrates that our methods significantly outperform other self-supervised approaches. This advantage stems from the fact that traditional self-supervised learning primarily focuses on frame-level similarities, which limits their effectiveness in leveraging instance information and causes struggles when training with images containing multiple objects. Further analysis of this is provided in Section G of the Appendix.

Table 7. Compare with self-supervised based methods. All methods use the same BDD and COCO raw images for training and the same detections for testing.

Method	Video	BDD MOT AssocA	BDD MOT mIDF1	TAO AssocA	BDD MOTS AssocA	BDD MOTS mIDF1
Train on BDD & COCO						
VFS [57]	✓	29.2	35.0	19.1	30.7	30.1
MoCov2 [10]	✗	42.7	46.7	30.7	51	45.3
DINO [7]	✗	23.1	16.8	12.9	20.2	22.2
Ours-SAM-B	✗	**51.9**	**54.9**	**35.8**	**53.7**	**49.1**

Table 8. Effect of training strategies and model architectures. The performance is evaluated on BDD MOT [64] dataset.

MASA training	Dynamic feature fusion	Object prior distillation	AssocA	mIDF1
			32.9	37.3
✓			48.5	51.7
✓	✓		50.1	53.3
✓	✓	✓	51.9	54.9

Table 9. Ablation study on different augmentations strategies, proposal quality and quantity.

(a) The effect of proposal quality.

Proposals	BDD AssocA	TAO AssocA
Mask2Former	46.4	29.8
SAM	50.9	34.1

(b) The effect of proposal number.

Instance Number	BDD AssocA	TAO AssocA
64	47.1	31.5
128	50.9	34.6
256	51.9	35.8

(c) The effect of data augmentation.

#	Affine	Mixup	LSJ	BDD MOT mIDF1	BDD MOT AssocA	TAO AssocA
1				48.2	43.9	28.5
2	✓			53.0	49.1	33.3
3		✓		52.8	49.9	31.5
4			✓	52.9	49.3	32.3
5	✓	✓	✓	54.9	51.9	35.8

4.3. Ablation Study and Analysis

To reduce the training costs, we bootstrap fewer raw images (40K) for training for the ablation experiments. Unless specified we train the model with an image collection containing 70k raw images from [64] and 110k images from [33] training set respectively. We employ the Ours-SAM-B model and test on BDD MOT and TAO TETA benchmarks.

Effect of Training Strategies and Model Architectures Table 8 illustrates that directly using the off-the-shelf SAM features (row 1) for association yields poor results. The primary reason is that SAM's original features are optimized for segmentation, not for instance-level discrimination. However, integrating our MASA training approach and adding a lightweight track head significantly enhances performance, yielding improvements of 15.6% in AssocA and 14.4% in mIDF1 on BDD MOT. This underscores the efficacy of our training strategy. Incorporating a dynamic feature fusion block further enhances performance by 1.6%. Additionally, joint training with the object prior distillation branch leads to an increase of 1.8% in AssocA and 1.6% in mIDF1, showing the effect of these architectural designs.

Effect of Proposal Diversity We evaluate different proposal generation mechanisms in association learning. We use only raw images from the training set of the BDD detection task for training. By substituting SAM in our MASA pipeline with Mask2former-SwinL [11], pre-trained on COCO. As shown in Table 9a, we found that the model trained with SAM's proposals significantly enhanced both in-domain performance on BDD and zero-shot tracking on TAO. This underscores the importance of SAM's dense, diverse object proposals for superior contrastive similarity learning.

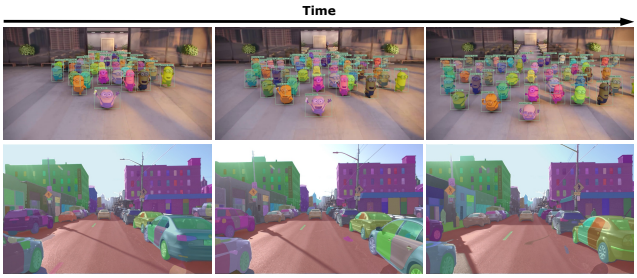


Figure 5. Qualitative results of our unified models using Ours-Grounding-DINO (top) and Ours-SAM-H (bottom). We use SAM-H to generate masks given the detected boxes.

Effect of Proposal Quantity Investigating the impact of SAM's proposal quantity on learning, we experimented with different upper bounds of 64, 128, and 256 proposals per batch. Table 9b shows consistent improvements in AssocA on BDD and TAO with increasing proposal numbers, indicating that a rich collection of instances fosters more discriminative tracking features.

Effect of Data Augmentations As shown in Table 9c, the combination of random affine, Mixup [66] and LSJ [20] gives the best performance. Method 1 represents basic data augmentation including flipping, resizing, color jitter and random cropping. If there is no strong augmentation (method 1), its mIDF1 on BDD MOT drops by 6.7%, being much worse than that with method 5. These results illustrate the necessity of strong augmentations in training only on static images.

Qualitative Results In Figure 5, we present the qualitative results of our unified methods, Grounding-DINO and SAM-H. Our methods accurately detect, segment, and track multiple objects and even their parts across diverse domains. This includes animated movie scenes featuring many similar-looking characters and driving scenes within complex environments.

5. Conclusion

We present MASA, a novel method that exploits the extensive instance-level shape and appearance information from SAM to learn generalizable instance associations from unlabeled images. MASA demonstrates exceptional zero-shot association performance across various benchmarks, eliminating the need for expensive domain-specific labels. Moreover, our universal MASA adapter can be added to any existing detection and segmentation models, enabling them to efficiently track any objects across diverse domains.

References

[1] Ali Athar, Sabarinath Mahadevan, Aljosa Osep, Laura Leal-Taixé, and Bastian Leibe. STEm-Seg: Spatio-temporal embeddings for instance segmentation in videos. In *ECCV*, 2020. 5

[2] Ali Athar, Jonathon Luiten, Alexander Hermans, Deva Ramanan, and Bastian Leibe. Hodor: High-level object descriptors for object re-segmentation in video learned from static images. In *CVPR*, 2022. 2

[3] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *ICCV*, 2019. 5

[4] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *ICIP*, 2016. 5

[5] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 3

[6] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9686–9696, 2023. 5

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 3, 7, 8

[8] Orcun Cetintas, Guillem Brasó, and Laura Leal-Taixé. Unifying short and long-term tracking with graph hierarchies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22877–22887, 2023. 2

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR, 2020. 3

[10] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2, 3, 7, 8

[11] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 8

[12] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *ECCV*, 2022. 2, 7

[13] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. Tracking anything with decoupled video segmentation. In *ICCV*, 2023. 2, 7

[14] Yangming Cheng, Liulei Li, Yuanyou Xu, Xiaodi Li, Zongxin Yang, Wenguan Wang, and Yi Yang. Segment and track anything. *arXiv preprint arXiv:2305.06558*, 2023. 2

[15] Xiyang Dai, Yinpeng Chen, Bin Xiao, Dongdong Chen, Mengchen Liu, Lu Yuan, and Lei Zhang. Dynamic head: Unifying object detection heads with attentions. In *CVPR*, 2021. 4

[16] Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. TAO: A large-scale benchmark for tracking any object. In *ECCV*, 2020. 2, 5, 6

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[18] Fei Du, Bo Xu, Jiasheng Tang, Yuqi Zhang, Fan Wang, and Hao Li. 1st place solution to eccv-tao-2020: Detect and represent any object for tracking. *arXiv preprint arXiv:2101.08040*, 2021. 5

[19] Yang Fu, Sifei Liu, Umar Iqbal, Shalini De Mello, Humphrey Shi, and Jan Kautz. Learning to track instances without video annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8680–8689, 2021. 2

[20] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *CVPR*, 2021. 6, 8

[21] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 6

[22] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 6

[23] Adam W Harley, Zhaoyuan Fang, and Katerina Fragkiadaki. Particle video revisited: Tracking through occlusions using point trajectories. In *European Conference on Computer Vision*, pages 59–75. Springer, 2022. 2

[24] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 4, 5, 7

[25] 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. 2

[26] Lei Ke, Xia Li, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Prototypical cross-attention networks for multiple object tracking and segmentation. *NeurIPS*, 2021. 5

[27] Lei Ke, Henghui Ding, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Video mask transfiner for high-quality video instance segmentation. *ECCV*, 2022. 5

[28] Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. In *NeurIPS*, 2023. 1

[29] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *NeurIPS*, 33, 2020. 3

[30] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and

Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 1, 2, 3, 4, 6

[31] Siyuan Li, Martin Danelljan, Henghui Ding, Thomas E Huang, and Fisher Yu. Tracking every thing in the wild. In *ECCV*. Springer, 2022. 1, 2, 3, 5, 6

[32] Siyuan Li, Tobias Fischer, Lei Ke, Henghui Ding, Martin Danelljan, and Fisher Yu. Ovtrack: Open-vocabulary multiple object tracking. In *CVPR*, 2023. 2, 5, 6

[33] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 8

[34] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 4

[35] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 1, 2, 4

[36] Yang Liu, Idil Esen Zulfikar, Jonathon Luiten, Achal Dave, Deva Ramanan, Bastian Leibe, Aljoša Ošep, and Laura Leal-Taixé. Opening up open world tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19045–19055, 2022. 2

[37] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 4

[38] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. HOTA: A higher order metric for evaluating multi-object tracking. *IJCV*, 2021. 7

[39] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8844–8854, 2022. 2

[40] 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. 2, 3

[41] Jiangmiao Pang, Linlu Qiu, Xia Li, Haofeng Chen, Qi Li, Trevor Darrell, and Fisher Yu. Quasi-dense similarity learning for multiple object tracking. In *CVPR*, 2021. 1, 2, 3, 5, 7

[42] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 Davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 3

[43] Frano Rajič, Lei Ke, Yu-Wing Tai, Chi-Keung Tang, Martin Danelljan, and Fisher Yu. Segment anything meets point tracking. *arXiv preprint arXiv:2307.01197*, 2023. 2, 7

[44] Esteban Real, Jonathon Shlens, Stefano Mazzocchi, Xin Pan, and Vincent Vanhoucke. Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video. In *CVPR*, 2017. 2

[45] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R–CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 4

[46] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, and Michael Bernstein. ImageNet Large scale visual recognition challenge. *IJCV*, 2015. 3

[47] Mattia Segu, Bernt Schiele, and Fisher Yu. Darth: Holistic test-time adaptation for multiple object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9717–9727, 2023. 2

[48] Qianqian Wang, Yen-Yu Chang, Ruojin Cai, Zhengqi Li, Bharath Hariharan, Aleksander Holynski, and Noah Snavely. Tracking everything everywhere all at once. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19795–19806, 2023. 2

[49] Weiyao Wang, Matt Feiszli, Heng Wang, and Du Tran. Unidentified video objects: A benchmark for dense, open-world segmentation. In *ICCV*, 2021. 2, 6, 7

[50] Zhongdao Wang, Liang Zheng, Yixuan Liu, Yali Li, and Shengjin Wang. Towards real-time multi-object tracking. In *ECCV*, 2020. 2

[51] Zhongdao Wang, Hengshuang Zhao, Ya-Li Li, Shengjin Wang, Philip Torr, and Luca Bertinetto. Do different tracking tasks require different appearance models? *NeurIPS*, 2021. 2, 3, 7

[52] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *ICIP*, 2017. 5

[53] Sanghyun Woo, Kwanyong Park, Seoung Wug Oh, In So Kweon, and Joon-Young Lee. Bridging images and videos: A simple learning framework for large vocabulary video object detection. In *European Conference on Computer Vision*, pages 238–258. Springer, 2022. 5

[54] Sanghyun Woo, Kwanyong Park, Seoung Wug Oh, In So Kweon, and Joon-Young Lee. Tracking by associating clips. In *European Conference on Computer Vision*, pages 129–145. Springer, 2022. 5

[55] J Wu, Y Jiang, S Bai, W Zhang, and X Bai. Seqformer: Sequential transformer for video instance segmentation. *ECCV*, 2021. 7

[56] Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Yuille, and Xiang Bai. In defense of online models for video instance segmentation. *ECCV*, 2022. 2

[57] Jiarui Xu and Xiaolong Wang. Rethinking self-supervised correspondence learning: A video frame-level similarity perspective. In *ICCV*, 2021. 2, 3, 7, 8

[58] Bin Yan, Yi Jiang, Peize Sun, Dong Wang, Zehuan Yuan, Ping Luo, and Huchuan Lu. Towards grand unification of object tracking. In *ECCV*, 2022. 2, 3, 5

[59] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Zehuan Yuan, Ping Luo, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *CVPR*, 2023. 1, 2, 3, 5, 7

[60] Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track anything: Segment anything meets videos. *arXiv preprint arXiv:2304.11968*, 2023. 2

[61] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *ICCV*, 2019. 5, 6, 7

[62] Zongxin Yang and Yi Yang. Decoupling features in hierarchical propagation for video object segmentation. In *NeurIPS*, 2022. 2

[63] Mingqiao Ye, Lei Ke, Siyuan Li, Yu-Wing Tai, Chi-Keung Tang, Martin Danelljan, and Fisher Yu. Cascade-detr: Delving into high-quality universal object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6704–6714, 2023. 1

[64] 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 *CVPR*, 2020. 2, 3, 6, 8

[65] Fangao Zeng, Bin Dong, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. MOTR: End-to-end multiple-object tracking with transformer. *arXiv preprint arXiv:2105.03247*, 2021. 2, 5

[66] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 6, 8

[67] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. DINO: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 3

[68] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *ECCV*, 2022. 5, 7

[69] Yuang Zhang, Tiancai Wang, and Xiangyu Zhang. MOTRv2: Bootstrapping end-to-end multi-object tracking by pretrained object detectors. *arXiv preprint arXiv:2211.09791*, 2022. 2, 3

[70] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *ECCV*, 2020. 2

[71] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022. 1, 2, 4

[72] Xingyi Zhou, Tianwei Yin, Vladlen Koltun, and Philipp Krähenbühl. Global tracking transformers. In *CVPR*, 2022. 2, 5, 7