GOVTrack: Towards Generative Open-Vocabulary Multi-Object Tracking

Zekun Qian[1] Ruize Han[2†] Zhixiang Wang[1] Liang Wan[1] Wei Feng[1]
[1]College of Intelligence and Computing, Tianjin University
[2]Faculty of Computer Science and Artificial Intelligence,
Shenzhen University of Advanced Technology
{clarkqian, zhixiang_wang, lwan, wfeng}@tju.edu.cn, hanruize@suat-sz.edu.cn

https://github.com/zekunqian/GOVTrackB

Abstract

We study a novel yet practical problem of generative open-vocabulary multi-object tracking (GOVMOT), which extends the MOT to localize, associate, and recognize generic-category objects from both seen (base) and unseen (novel) classes as a generative problem. This overcomes the limitations of previous open-set MOT problems, which either fail to classify novel classes or require a predefined list of category texts as prompts. To study this problem, the top priority is to build a benchmark. In this work, we build GOVTrackB, a large-scale and comprehensive benchmark, to provide a standard evaluation platform for the GOVMOT problem. Compared to previous datasets, GOVTrackB has more abundant and balanced base/novel classes, along with corresponding samples for evaluation with less bias. We also propose a new multi-granularity recognition metric to better evaluate the generative object recognition in GOVMOT. We further develop GOVTracker as a baseline method, featuring a consistency-aware focal loss that enhances object association by jointly modeling appearance and semantic consistency. Through extensive benchmark evaluations, we report and analyze the results of various state-of-the-art methods, which demonstrate the rationale of GOVMOT, as well as the usefulness and advantages of GOVTrackB.

1. Introduction

Multi-object tracking (MOT), which involves detecting and associating targets of interest in a video, is a classical and fundamental problem with many real-world applications, such as video surveillance, autonomous driving, *etc*. Recently, MOT has attracted broad attention with numerous algorithms and datasets [6, 40, 59, 63, 65]. For many years, MOT has mainly focused on tracking human targets, *e.g*., the datasets of MOT15 [27], MOT20 [11], DanceTrack [56].

Figure 1. Comparison of different types of open-set MOT.

Several works also focus on traffic scenes and aim to track vehicles, such as the well-known KITTI dataset [20].

In real-world scenes, the categories present in videos are diverse and far from being limited to humans and vehicles. TAO [10], as the first of its kind, constructs a large-scale benchmark to study the tracking of a diverse range of target categories, encompassing a total of 833 object classes. During the same period, GMOT-40 [3] builds a generic multi-object tracking benchmark with 10 object classes but a higher density of objects per frame. As the number of categories in the MOT task increases, the evaluation metrics have evolved from solely focusing on object localization and association to also incorporating class recognition. A new metric TETA (tracking-every-thing accuracy) has been proposed [30] to evaluate the generic MOT from these three aspects. More recently, open-world MOT (OWMOT) [36] is proposed to train a tracker using the samples from "base classes", and test it on videos containing objects from "novel classes". The tracker must recognize the base-class objects and *identify all other unseen classes as "unknown"*. Further, open-vocabulary MOT (OVMOT) [31] aims not only to distinguish novel-category objects, as OWMOT does, but also to classify each object, typically achieved through a pre-trained multi-modal model, *e.g*., CLIP [52].

Undoubtedly, the development of MOT from specific-category to (closed-set) generic-category and further to open-world/vocabulary (open-set) settings is becoming increasingly practical. However, existing open-set MOT tasks suffer

†Corresponding author.

from several limitations. As illustrated in Fig. 1, we compare existing open-set multi-object tracking tasks: OWMOT and OVMOT. OWMOT, as shown in Fig. 1(a), fails to provide specific classes for novel-class objects, instead labeling them generically as "unknown". While OVMOT, as shown in Fig. 1(b), can classify novel classes, it relies on a predefined class list. This is difficult to obtain in real applications, especially for novel classes, which are termed "novel" because the categories are previously unknown. This way, in this work, we propose a new problem called Generative Open-Vocabulary Multi-Object Tracking (GOVMOT), which treats the object recognition task as a generative problem, rather than as a classification problem in previous open-set MOT tasks. In Fig. 1(c), GOVMOT identifies novel classes and generates multi-granularity class labels without a predefined class list, addressing the limitations of prior open-set MOT. Moreover, the multi-granularity perception capability is also unavailable for both OWMOT and OVMOT.

To study GOVMOT, the top priority is to build a benchmark. Previous work OVTrack [31] directly constructs OVMOT evaluation sets by selecting TAO validation and test videos whose categories overlap with LVIS [21]. However, this simple category-intersection drastically shrinks the novel split: on the validation set only 35 novel classes remain with $\sim$3K annotations, and on the test set only 33 novel classes remain with $\sim$2K annotations. Such limited class and sample sizes make the resulting datasets unreliable for evaluating open-vocabulary performance. To address these limitations, we build a new and comprehensive evaluation benchmark, GOVTrackB, following the principles of category enrichment, sample enrichment, and semantic compatibility. Specifically, compared to previous datasets, GOVTrackB offers a more diverse and balanced distribution of base/novel classes, along with abundant videos for evaluation with less bias.

Beyond the benchmark, we develop GOVTracker as the first baseline for GOVMOT to address the scarcity of large-scale generic-object tracking annotations. GOVTracker employs a two-stage strategy: Stage 1 introduces a consistency-aware focal loss that dynamically adapts to sample difficulty by jointly modeling appearance and semantic consistency on synthetic image pairs, while Stage 2 refines the model via self-supervised learning on raw videos to capture temporal motion patterns.

In summary, the main contributions of this paper are as follows:

• We propose a new problem, generative open-vocabulary multi-object tracking (GOVMOT), which overcomes the limitations of previous MOT tasks, including the inability to classify novel classes and the requirement for a predefined class list. GOVMOT thus releases the potential of MOT for real-world applications.

• We build GOVTrackB, a large-scale and comprehensive benchmark, to provide a standard evaluation platform for the GOVMOT problem. We also propose a multi-granularity recognition metric to further improve the performance evaluation.

• We develop the first baseline method for GOVMOT with a two-stage association learning strategy. On GOVTrackB, we conduct benchmark evaluation experiments of our baseline and other state-of-the-art comparison methods, which demonstrate the rationale of GOVMOT and the usefulness and advantages of GOVTrackB.

2. Related Work

From multiple object tracking (MOT) to open-set MOT. MOT is a classical problem, in which the dominant approach is the tracking-by-detection framework [1], which initially identifies objects in each frame and then associates them across frames using various cues such as object appearance features [4, 5, 18, 28, 41, 45, 53, 59], 2D motion features [15, 51, 54, 60, 66], or 3D motion features [22, 25, 37, 43, 55, 58].

To extend the object categories in the MOT task, the TAO benchmark [10] has been proposed, which handles the MOT under various object categories with a long-tail distribution. Several follow-up works are proposed to evaluate this benchmark including AOA [13], GTR [67], TET [30], QDTrack [18], *etc*. Open-set MOT has not been extensively explored. Some existing related works [9, 44] adopt the class-agnostic detectors with general trackers to implement open-world tracking. These methods focus solely on tracking salient objects in the scene without considering specific object categories. The recent TAO-OW [36] takes a step further, which divides all objects into known and unknown categories. Open-world MOT is achieved by tracking objects from both known and unknown categories. However, it still falls short in the recognition of specific object classes in unknown categories.

Further, OVTrack [31] incorporates open vocabulary into the tracking task as OVMOT, providing a baseline method and benchmark based on the TAO dataset. Subsequent studies have further improved OVMOT performance [32, 47–50]. However, a remaining problem is the requirement for the predefined category list during testing. Differently, GOVMOT does not require predefined category names as in the OVMOT task, and can generate multi-granularity target categories with a generative head.

MOT Benchmarks. Benchmarks have played a pivotal role in advancing the development of MOT. Early datasets like PETS2009 [17] focused on pedestrian tracking with limited video sequences. The MOT Challenge [11, 27] introduced more crowded scenes, significantly progressing the field. KITTI [20] and BDD100K [62], designed for autonomous driving, focus on tracking vehicles and pedestrians. Specialized datasets such as DanceTrack [56],

SportsMOT [8], and AnimalTrack [64] handle specific scenarios like dancing, sports, and wildlife. UAVDT [12] and VisDrone [68] support aerial tracking. Despite these advancements, many benchmarks still have limited object categories. Recent video datasets like GMOT-40 [3] and YT-VIS [61] aim to address specific tasks like one-shot MOT and video instance segmentation but still fall short in supporting a wide range of categories. A large-scale dataset TAO [10] annotates 833 categories, offering a broader platform for studying object tracking on long-tailed distributions.

Based on TAO, OVTrack [31] builds the OVTAO evaluation datasets. Since the current popular open-vocabulary related tasks commonly use the LVIS [21] dataset for base/novel category splits, OVTrack also follows this setting. However, the proportion of novel classes in OVTAO accounts for only 10% of the original novel classes in LVIS, totaling around 30 classes. The limited number of classes hinders the effective validation of the algorithm's performance on various open-vocabulary categories, making it unsuitable for the proposed GOVMOT problem. Therefore, there is an urgent need for a benchmark with rich categories and abundant videos to support GOVMOT. Thus, we propose a new benchmark GOVTrackB to effectively address these issues.

3. GOVTrack Benchmark

3.1. Problem Formulation

We first provide the problem formulation of GOVMOT. Given a video sequence with various objects, GOVMOT aims to simultaneously achieve the localization, association and recognition tasks, thus generating a bounding box $\mathbf{b} = [x, y, w, h]$, a continuous ID number d (along the video) and a category c for each target in the video. The annotated object categories that appear during training are defined as $\mathcal{C}^{\mathrm{b}}$, *i.e.*, the base class set. In testing, we aim to obtain the GOVMOT results, *i.e.*, the object category set $\mathcal{C}^{\mathrm{open}}$ is an open corpus. Obviously we have $\mathcal{C}^{\mathrm{b}} \subset \mathcal{C}^{\mathrm{open}}$, and we define the novel class set as $\mathcal{C}^{\mathrm{n}} = \mathcal{C}^{\mathrm{open}} \backslash \mathcal{C}^{\mathrm{b}}$. Note that, we treat the category recognition task as a generative task, with no need for the category list of $\mathcal{C}^{\mathrm{open}}$ as input during testing. Ideally, $\mathcal{C}^{\mathrm{open}}$ contains all the categories in the real world. In practice, for GOVMOT evaluation, we can limit $\mathcal{C}^{\mathrm{open}}$ to a large-scale thesaurus.

3.2. Principle of Benchmark Construction

To build the GOVMOT benchmark (GOVTrackB), we first establish the following principles:

P0: Principle of standardness. Following the base and novel class division mode proposed in LVIS;

P1: Category enrichment principle. Base/novel classes should be diverse and balanced;

P2: Sample enrichment principle. Evaluation videos/objects for all classes should be abundant;

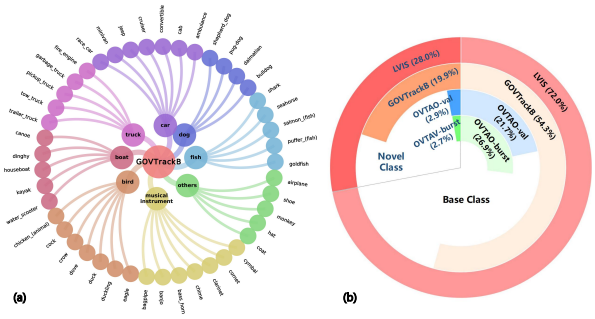

Figure 2. Statistics of the object categories in datasets.

P3: Semantic compatibility principle. The evaluation of object recognition should be compatible.

The first principle **P0** ensures the base and novel class division in our dataset is consistent with that in the widely used LVIS. This is because previous works, *e.g.*, many open-vocabulary detection methods [14, 24, 33, 39, 42], and the open-vocabulary tracker OVTrack all use LVIS as the training dataset. As a testing dataset, GOVTrackB, with the same base/novel class division is more convenient for evaluating the algorithms trained on LVIS.

Both **P1** and **P2** guarantee the richness of the dataset, which aims to increase the number of object categories and the amount of class-balanced samples. This is particularly significant for the generative open-vocabulary tracking task. As is well known, real-world data is often skewed and imbalanced. However, as an evaluation benchmark, we strive to maintain category balance to ensure that *abundant yet simplistic classes do not dominate the evaluation* in GOVTrackB. Next, we provide a detailed clarification of this point. As mentioned in related works [23, 38], it is essential to maintain class balance to prevent abundant yet simplistic classes from dominating the evaluation. Specifically, first, class imbalance can lead to distortions in evaluation metrics such as accuracy, precision, and F1 score, of which accuracy is the most commonly used metric in OVMOT/GOVMOT. Such distortions can lead to biased evaluation results, potentially producing unrealistically optimistic or pessimistic outcomes. For example, results may appear falsely optimistic when samples from easier classes dominate the dataset. Second, class imbalance can mislead comparison results by causing majority classes to overshadow minority classes, thereby creating an illusion of either strong or poor overall performance. Such distortions can significantly mislead average evaluation metrics in OVMOT/GOVMOT, making comparisons among methods unreliable.

The last principle **P3** aims to address the problem of semantic ambiguity, which stems from two aspects. The first aspect arises from the dataset annotation. Due to the large number of categories, the granularity of the category annotations in the dataset is misaligned, leading to inaccurate evaluations. For example, the granularity of classification

for some targets is only up to "bird", while for others, it is more specific, such as "goose" or "duck". This misalignment in labeling makes it difficult to compare recognition accuracy across algorithms during evaluation. The second aspect comes from the task GOVMOT. Different from the classification head in OVMOT, the proposed GOVMOT handles recognition as a generative problem. This may lead to semantic synonymy or subordination. For example, "cab" and "taxi" usually mean the same thing, which are both types of "car". Therefore, for the ground truth "cab", the prediction result of "taxi" or even "car" should be reconsidered and not simply taken as false. As shown in the example in Fig. 2(a), we divide the categories in LVIS into multiple levels and conduct multi-level evaluations to strive to achieve **P3**. Following **P0-P3**, we build GOVTrackB.

3.3. Dataset Collection and Annotation

Data collection. We observe that previous open-class works utilize existing datasets to build new benchmarks (*e.g.*, the first open-vocabulary tracking OVTrack [31] and the first open-world tracking OWTrack [36], both of which directly use the TAO [10] dataset for evaluation). However, the previous work OVTrack [31], which also follows **P0**, directly uses TAO's validation and test sets (annotations from BURST [2]) and only retains the classes that overlap with LVIS for data selection, forming the OVTrack testing datasets, i.e., OVTAO validation (OVTAO-val) and test (OVTAO-burst) sets. *This simple category intersection operation significantly reduces the number of classes*.

In this work, we first use TAO as the foundation for constructing GOVTrackB, which provides longer videos but has a limited number of categories (overlapped with LVIS). But different from [31], by investigating the recent video datasets, we also further select another large one with various object categories, *i.e.*, the LV-VIS [57] dataset, which can effectively compensate for the shortcomings of TAO, making it more suitable for addressing the GOVMOT task. Specifically, TAO is a generic-category object tracking dataset with a total of 833 classes and 2,907 videos. LV-VIS is a large-vocabulary video instance segmentation dataset with 1,196 classes and 4,828 videos. The accuracy of the annotations in both datasets is well-established. The TAO dataset ensures high-quality annotations through a hierarchical approach that enhances precision, with a re-annotation process that verifies a strong Intersection over Union (IoU) for accuracy. Additionally, the maintenance of the TAO dataset is supported by a blend of automation, manual review, and statistical analysis. Similarly, the LV-VIS dataset maintains accuracy through strict category reviews and cross-revision.

Data integration and manual verification. During the process of screening and integrating datasets, we discover inconsistencies between the category names in the TAO and LV-VIS datasets compared to those in LVIS. This inconsistency prevents us from directly using the original dataset's classification annotations. To meet principle **P0**, we manually unify category definitions across the different datasets. For example, several terms may share the same semantic meaning (*e.g.*, "police car" in LVIS and "police cruiser" in LV-VIS), a single vocabulary may have multiple meanings (*e.g.*, "bat", which can refer to both an animal and a piece of sports equipment), and the same object may represent vocabularies of varying granularities (*e.g.*, the "dog" category encompasses breeds such as "bulldog", "dalmatian", "pug" and "shepherd dog"). We manually review all categories in the original datasets and align them with the 1,203 categories defined in LVIS. Then we filter the test and validation sets of TAO and the training and validation sets of LV-VIS to select videos that meet principle **P0**. Importantly, due to inconsistencies in settings and annotation granularities between source datasets and LVIS, we re-annotated over 50K labels after adhering to **P0**, ensuring the final dataset strictly satisfies **P0** while maintaining the overall scale.

To satisfy **P1**, we use a greedy algorithm aiming to minimize the total number of videos while ensuring that each category contains at least two videos whenever possible, thus ensuring category diversity and balance. This results in a selection of 903 videos covering 892 categories (also contained in LVIS). To satisfy **P2**, we again use a greedy algorithm, aiming to allocate as many tracks as possible for each category while maintaining the similary total number of videos. This further results in 732 videos containing 4,766 tracks. In total, we collected 1,635 videos, of which 496 include novel category objects and 1,600 encompass base categories.

3.4. Dataset Statistics and Comparison

We then show the statistics of GOVTrackB and compare it with two existing OVTrack datasets, i.e., OVTAO-val [31] and OVTAO-burst [2]. GOVTrackB has the following typical advantages.

Various and balanced object categories. GOVTrackB contains a total of 892 available categories, composed of 653 base and 239 novel classes. We show some example classes in GOVTrackB in Fig. 2(a), the categories cover every aspect in the real-world applications, *e.g.*, various transportation, animals, and household items, *etc*. Note that, following the original class annotation in our basic datasets TAO and LV-VIS, GOVTrackB involves multi-granularity categories. For example, the fine-grained class "shepherd dog" and its general class "dog" are concomitant in GOVTrackB's category list. We leverage this subordinate relationship to design the new evaluation metric in the following.

As shown in Fig. 2(b), GOVTrackB includes 653 base and 239 novel classes, which account for 75.5% of the original LVIS base categories and 70.9% of the novel categories, respectively, effectively ensuring the category diversity. In

comparison, OVTAO-val and OVTAO-burst only contain 30.1% and 37.4% of the original LVIS base classes, respectively. Regarding the novel classes, the ratios are only about 10% (2.9%/2.7% *vs.* 28.0%). The various object categories make GOVTrackB more comprehensive in evaluating generative open-vocabulary object tracking performance.

Next, we consider the category balance of the dataset. As shown in Fig. 3, we calculate the normalized entropy of different units (object boxes, object tracks, videos) for the category set. Specifically, for N categories in the dataset, we compute the Shannon Entropy as $H(p) = -\sum_{i=1}^{N} p_i \log(p_i)$, where p_i denotes the probability of a unit belonging to category i, and the Maximum Entropy as $H_{\max} = \log(n)$. Then we get the Normalized Entropy as $\mathrm{NE} = \frac{H(p)}{H_{\max}}$, which can reflect the class balance in the dataset. We can see that, the class balance of the proposed GOVTrackB is higher than OVTAO-val and OVTAO-burst. We know that, in the real world, the object category distribution is long-tail but not balanced. However, as an evaluation benchmark, we strive to maintain the category balanced to *guarantee that the evaluation is not dominated by large-scale yet simple classes*.

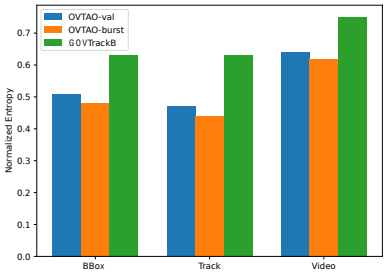

Figure 3. Statistics of normalized entropy.

Abundant samples for both base and novel classes. As shown in Fig. 4, we show the number of objects, tracks, and videos in OVTAO-val, OVTAO-burst, and GOVTrackB datasets. The statistics are split through the base and novel classes. We can see that, for the base class, the number of object boxes, tracks, and videos in GOVTrackB is greater than those of OVTAO-val and OVTAO-burst. Moreover, in terms of novel classes, we can see that the data amount of GOVTrackB is significantly larger than that of OVTAO-val and OVTAO-burst, with the increase ranging from 7.7 to 11.2 times.

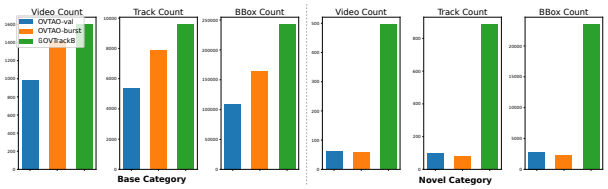

Figure 4. Statistics of the videos, track, and objects for base/novel classes in different datasets.

From the comparison, we can see that the proposed GOVTrackB is more in line with the above principles **P1** and **P2**. We further provide more statistics of GOVTrackB to show its data distribution and characteristics in Table 1 of the supplementary material. We can see that, regardless of the total number of videos, frames, tracks, and bounding boxes, GOVTrackB is larger than the other two datasets. The "Density" metric means the object count in each frame, where GOVTrackB exceeds the other two datasets significantly. Additionally, for the average number of tracks per video, and the average boxes per video or track, GOVTrackB is also larger. These fully reflect the enrichment of GOVTrackB.

3.5. Evaluation Metrics

Following [31], we use the open-category tracking metric namely tracking-every-thing accuracy (TETA) in [30] for evaluation. TETA is composed of three parts, i.e., object localization, association, and classification accuracies. First, the localization accuracy (LocA) is computed through the matching of the ground truth (GT) boxes with predicted boxes without considering class, as $\mathrm{LocA} = \frac{|\mathrm{TPL}|}{|\mathrm{TPL}|+|\mathrm{FPL}|+|\mathrm{FNL}|}$. Second, association accuracy (AssocA) is determined by matching the identities of associated GT instances with the predicted association, as $\mathrm{AssocA} = \frac{1}{|\mathrm{TPL}|}\sum_{b\in \mathrm{TPL}} \frac{|\mathrm{TPA}(b)|}{|\mathrm{TPA}(b)|+|\mathrm{FPA}(b)|+|\mathrm{FNA}(b)|}$. Finally, classification accuracy (ClsA) is calculated using all correctly localized instances, by comparing the predicted classes with the corresponding GT classes, as $\mathrm{ClsA} = \frac{|\mathrm{TPC}|}{|\mathrm{TPC}|+|\mathrm{FPC}|+|\mathrm{FNC}|}$. The TETA score is computed as the mean value of the above three scores as $\mathrm{TETA} = \frac{\mathrm{LocA}+\mathrm{ClsA}+\mathrm{AssocA}}{3}$.

In previous open-category tracking tasks [31, 36], object recognition is always taken as a classification problem using the above ClsA metric. We take the recognition as a generative task, which may generate multiple labels. This way, as shown in Fig. 5, we first use CLIP [52] to encode the predicted output (multiple generated object categories concatenated into a single prompt using ", ") and each base/novel category in LVIS. Next, we calculate the similarity between these encoded features to choose a high-similarity category label, *i.e.*, the matching class (a single class name in LVIS), which can be used to compute ClsA. Note that, the base and novel categories are only used for result evaluation, which is different from OVMOT that uses them to generate the prediction results.

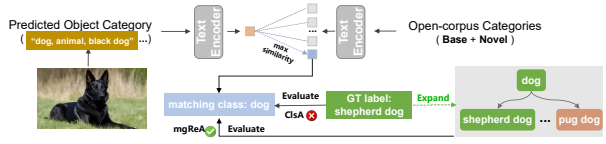

Figure 5. Illustration of the multi-granularity evaluation.

As discussed in **P3** in Section 3.2, the generative open-vocabulary tracking may introduce the semantic ambiguity problem. To address this problem, we design a novel multi-

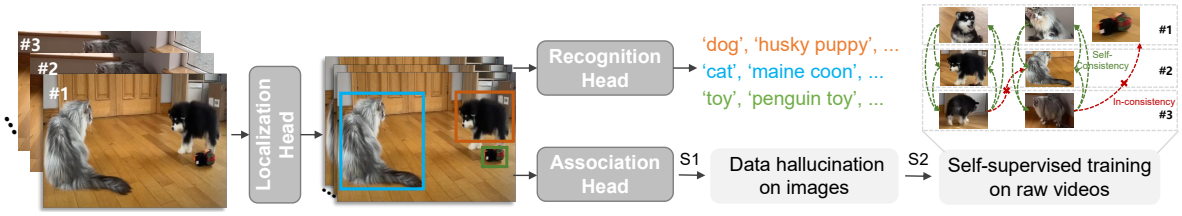

Figure 6. Pipeline of the proposed baseline method GOVTracker.

granularity recognition accuracy (mgReA). Specifically, considering the diversity of the generated vocabulary, we aggregate the categories in LVIS according to WordNet [16] into a hierarchy structure. We construct hierarchical relationships using WordNet's noun network. Since WordNet nodes have multiple paths causing relationship uncertainties, we employ a greedy algorithm to select parent nodes that maximize total child nodes, iteratively organizing all categories into clear hierarchies.

As shown in Fig. 5, when computing mgReA, if the ground-truth category label belongs to any category within this aggregated multi-granularity class hierarchy, it is considered an expanded successful recognition. A simple example is that, for the ground-truth label "shepherd dog", we expand it to "dog". For the matching class (prediction) of "dog", ClsA will judge it as a false result, but mgReA takes it as true.

This metric provides a more intuitive and compatible evaluation, since we do not need very fine-grained classifications in many cases. Based on mgReA, we define a new comprehensive metric called tracking&recognizing-every-thing accuracy (TRETA) as $\mathrm{TRETA} = \frac{\mathrm{LocA}+\mathrm{mgReA}+\mathrm{AssocA}}{3}$ for the GOVMOT problem. In short, mgReA is **feasible** (fits multi-granular generative outputs and labels), **discriminative** (separates hierarchy confusion from true errors), and **universal** (works for both GOV and OV, improving error diagnosis).

4. A Baseline Method: GOVTracker

4.1. Method Pipeline

Figure 6 illustrates the framework of our pipeline, which consists of three main components: a localization head, a recognition head, and an association head.

Localization Head. As shown in Fig. 6, similar to most tracking-by-detection based MOT approaches, we first need to obtain object bounding boxes for each frame. Since our focus is on generative open-vocabulary object tracking, we aim to localize *generic-class objects*. We employ the well-known detector Deformable DETR [69] as the basic network of the salient localization head with a binary cross-entropy loss.

Recognition Head. The recognition head is used to generate the category name of the object. It mainly consists of a generative language model, for which we use FlanT5-base [7] and initialize it with its pre-trained weights. The visual features of the candidate objects obtained from Deformable DETR are mapped to the input space of the generative model through a projection network. Then the generative model is trained following the manner and loss function in [34], using the VG [26] and GRIT [46] image-text pairs as training data. The beam size of the language model is set to 2, meaning that we generate two category nouns for each object.

Association Head. The association head learns to match objects across frames, which is crucial for multi-object tracking. Given the scarcity of large-scale generic-object video datasets with tracking annotations [31], we propose a two-stage training strategy: (1) Stage 1 trains on synthetic image pairs using a novel consistency-aware focal loss (Section 4.2), and (2) Stage 2 refines the model on raw videos via self-supervised learning (Section 4.3). This design leverages both static image data and temporal video dynamics for robust association learning.

4.2. Stage 1: Image Pair Training with Consistency-aware Focal Loss

Image Pair Generation. In the first training stage, we learn the association model using static images. Following [31], we apply the data hallucination strategy to generate pairwise images for training. Specifically, given an image of base categories in LVIS [21], we apply the diffusion model to generate its adjoint image with the same object categories but different styles, and format an image pair $(I^{\text{key}}, I^{\text{ref}})$.

Motivation for Consistency-aware Focal Loss. Previous association learning methods [18, 31] typically compute a cross-frame similarity matrix $\mathbf{S}$ from association features F, then align it with the ground-truth assignment matrix $\mathbf{M}_{\text{gt}}$ using cross-entropy loss. However, this treats all samples equally, ignoring their varying difficulty levels.

While focal loss [35] $L_{\text{focal}}(p) = -\alpha(1-p)^{\gamma}\log(p)$ can emphasize hard samples through the modulating factor $(1-p)^{\gamma}$, it has two limitations: (1) the fixed γ lacks adaptability to diverse sample difficulties in open-vocabulary scenarios; (2) $(1-p)^{\gamma}$ only considers pairwise similarity, ignoring the global consistency across all objects.

To address these issues, we propose to dynamically compute γ based on appearance and semantic consistency across

all scene objects, enabling adaptive focus on truly difficult samples.

Appearance and Semantic Consistency Factors. To measure global consistency across all objects in two frames, we first compute the appearance similarity matrix $\mathbf{S} = F^{\text{key}} \cdot (F^{\text{ref}})^\top$, where F^{key} and F^{ref} are association features from the key and reference frames. We then apply softmax normalization: $f(\mathbf{S}(r,c)) = \frac{\exp(\tau \mathbf{S}(r,c))}{\sum_{c'=1}^{C} \exp(\tau \mathbf{S}(r,c'))}$, where τ is a temperature parameter. The cross-frame self-consistency matrix is computed as $\mathbf{Y} = f(\mathbf{S}) \cdot f(\mathbf{S}^\top)$. Ideally, $\mathbf{Y}$ should approximate the identity matrix $\mathbf{I}$ when same-object similarities are higher than cross-object similarities. We thus define the appearance consistency factor as: $\gamma_{\text{app}} = \tanh\left(|\text{diag}(\mathbf{Y}) - \text{diag}(\mathbf{I})|\right)$, where $\text{diag}(\cdot)$ extracts diagonal elements and $\tanh(\cdot)$ constrains values to [0, 1]. When objects are consistently matched across frames, γ_{app} is small (easy samples); otherwise it is large (hard samples). Similarly, applying the same procedure to semantic features from the recognition head yields the semantic consistency factor γ_{sem}.

Integrating the consistency factors, we formulate the consistency-aware focal loss as: $\mathcal{L}_c = -\alpha \cdot (1-p)^\gamma \log(p)$ where the dynamic modulating factor is $\gamma = 1 - (\beta_1 \gamma_{\text{app}} + \beta_2 \gamma_{\text{sem}})$, with $\beta_1 + \beta_2 = 1$ as hyperparameters. This design allows γ to adapt to sample difficulty: when consistency is low (hard samples), γ increases to emphasize them; when consistency is high (easy samples), γ decreases to down-weight them.

4.3. Stage 2: Self-supervised Video Training

While Stage 1 learns robust association features from synthetic image pairs, it lacks exposure to real temporal dynamics in videos. Therefore, in Stage 2, we refine the association model using raw videos without tracking annotations. Following [19], we employ a self-supervised cycle-consistency strategy: given a reference object in frame t, we find its most similar object in frame t', and then verify that this object's most similar match in frame t is the original reference object. This cycle-consistency principle is enforced through a self-supervised consistency loss [19]: $\mathcal{L}_{\text{self-sup}}$. This stage complements Stage 1 by capturing temporal motion patterns and appearance variations in real videos, bridging the gap between synthetic image pairs and actual tracking scenarios. For more implementation details, please refer to Section 2 of the supplementary material.

5. Experimental Results

5.1. Comparison Methods

As a new problem, there is no approach that can directly handle the GOVMOT. We try to include as many approaches as possible with necessary modifications for comparison on the proposed GOVTrackB. ❶ MOT algorithms: *i.e.*, QDTrack [18] and TETer [30], which are trained on both base and novel categories in LVIS [21] and TAO [10] training sets by a closed-set training approach. ❷ Two public OV-MOT algorithms: *i.e.*, OVTrack [31] and OVTR [29], by additionally giving the base and novel class list during testing as the OV setting of it. ❸ Combining the open-vocabulary detection (OVD) with the object tracking method for association: We select three state-of-the-art OVD algorithms, *i.e.*, VLDet [33], CoDet [39], MM-OVOD [24], and three tracking methods, including the appearance-based tracking method, namely DiffuTrack in OVTrack [31], motion-based tracking in ByteTrack [65] and OC-SORT [6]. Note that, these methods also need the base and novel class lists for testing. ❹ Open-ended generative object detection method GenerateU [34] as the detector, combined with the above tracking modules. GenerateU is similar to GOVTrack without the class lists when testing.

5.2. Benchmark Results

Comparison among state-of-the-art methods. As shown in Table 1, classical MOT algorithms QDTrack [18] and TETer [30] show poor overall performance when facing 892 categories, achieving the lowest TETA scores. Open-vocabulary methods OVTrack [31] and OVTR [29] achieve reasonable results using predefined class lists. The other combination-based OVMOT approaches achieve performance comparable to OVTrack, yet remain inferior to the more recent OVTR method.

GOV setting results. We present comprehensive results under the generative open-vocabulary (GOV) setting in Table 1, where methods use no predefined class lists during testing. We evaluate GenerateU [34] combined with three tracking strategies (DiffuTrack, ByteTrack, OC-SORT) and our proposed GOVTracker. In this challenging GOV setting, GOVTracker significantly outperforms all alternatives across all metrics. Most notably, GOVTracker achieves 38.4% TETA for base classes and 38.3% for novel classes, substantially surpassing GenerateU combinations (best: 31.0% and 29.5%). The superior AssocA performance (54.3% base, 62.8% novel) demonstrates GOVTracker's exceptional temporal association capabilities. Remarkably, even compared to OV methods that use class lists, GOVTracker maintains competitive or superior performance without any predefined vocabulary, *proving GOVMOT's practical viability and future potential.*

Results of new metrics. As shown in Table 1, our proposed mgReA consistently aligns with ClsA while providing superior discriminative ability. For example, QDTrack and TETer have nearly identical ClsA scores for novel classes (0.1% vs. 0.1%), making them indistinguishable. However, our mgReA metric effectively differentiates their performance (7.6% vs. 2.5%), demonstrating mgReA's superior discriminative ability when ClsA fails. The margin between

Table 1. Comparison results on the proposed GOVTrackB (%).

Methods	Train Data		Open Type	Base Class						Novel Class					
	Base	Novel	OV/OC	TETA	LocA	AssocA	ClsA	mgReA	TRETA	TETA	LocA	AssocA	ClsA	mgReA	TRETA
QDTrack [18]	✓	✓	-	26.6	32.2	38.8	8.8	13.7	28.2	28.1	37.8	46.4	0.1	7.6	30.6
TETer [30]	✓	✓	-	26.5	36.7	41.5	1.2	3.4	27.2	31.4	45.4	48.7	0.1	2.5	32.2
OVTrack [31]	✓	†	OV	34.6	37.8	44.3	21.7	28.9	37.0	32.8	44.1	50.6	3.6	12.3	35.7
OVTR [29]	✓	†	OV	36.5	41.5	51.3	16.7	22.5	38.4	36.2	**48.2**	57.5	2.9	9.4	38.4
VLDet [33]															
+ DiffuTrack [31]	✓	†	OV	32.9	36.1	45.2	17.5	25.4	35.6	32.9	40.9	49.5	8.2	15.1	35.2
+ ByteTrack [65]	✓	†	OV	29.3	32.0	41.6	14.4	19.8	31.1	29.5	34.5	48.0	6.0	12.0	31.5
+ OC-SORT [6]	✓	†	OV	26.1	29.2	36.1	13.1	18.1	27.8	27.0	34.1	40.5	6.5	12.5	29.0
CoDet [39]															
+ DiffuTrack [31]	✓	†	OV	35.1	36.7	46.3	**22.4**	**29.3**	37.4	33.0	39.2	48.0	**11.8**	**18.4**	35.2
+ ByteTrack [65]	✓	†	OV	31.4	33.3	42.8	18.1	23.9	33.3	31.5	37.2	47.3	10.2	16.5	33.7
+ OC-SORT [6]	✓	†	OV	28.7	31.0	37.5	17.5	22.9	30.5	28.5	35.3	40.2	9.9	15.8	30.4
MM-OVOD [24]															
+ DiffuTrack [31]	✓	†	OV	35.4	38.6	47.5	20.0	26.3	37.5	32.6	42.5	51.1	4.3	6.8	33.5
+ ByteTrack [65]	✓	†	OV	31.4	33.0	43.2	18.0	23.8	33.3	29.7	36.8	47.7	4.6	11.0	31.8
+ OC-SORT [6]	✓	†	OV	27.3	29.3	36.8	15.8	20.9	29.0	25.0	32.0	38.6	4.5	10.2	26.9
GenerateU [34]															
+ DiffuTrack [31]	✓	✗	GOV	31.0	37.5	40.3	15.2	21.0	32.9	29.5	43.4	43.5	1.6	9.3	32.1
+ ByteTrack [65]	✓	✗	GOV	28.2	30.7	38.7	15.2	20.7	30.0	27.1	36.8	42.8	1.7	9.8	29.8
+ OC-SORT [6]	✓	✗	GOV	27.0	28.8	37.4	14.8	20.2	28.8	25.1	32.7	40.7	1.8	10.1	27.8
GOVTracker	✓	✗	GOV	**38.4**	**43.0**	**54.3**	17.9	25.7	**41.0**	**38.3**	47.3	**62.8**	4.9	16.7	**42.3**

'✓' denotes using the corresponding videos of base/novel class, '†' denotes only using the class list but not the videos for testing, and '✗' means using nothing about the novel class.

Table 2. Ablation study on the proposed two-stage association learning (%).

Stage 1 Loss	Consistency Factors		Stage 2	Base Class						Novel Class					
	γ_{app}	γ_{sem}		TETA	LocA	AssocA	ClsA	mgReA	TRETA	TETA	LocA	AssocA	ClsA	mgReA	TRETA
CE	-	-	-	34.8	41.7	46.9	15.7	21.4	36.7	33.0	46.8	50.9	3.2	12.3	36.7
Focal	-	-	-	35.3	42.6	47.2	16.2	21.6	37.1	34.3	46.6	52.5	3.9	13.9	37.7
Consistency-aware Focal	✓	-	-	37.0	42.7	51.5	16.9	23.1	39.2	36.4	47.3	57.3	4.5	15.2	39.9
	✓	✓	-	37.8	42.3	53.5	17.6	25.1	40.5	37.4	47.1	60.4	4.8	16.8	41.4
	✓	✓	✓	38.4	43.0	54.3	17.9	25.7	41.0	38.3	47.3	62.8	4.9	16.7	42.3

mgReA scores is generally larger than ClsA scores, enabling better differentiation among methods. When ClsA scores are similar, TETA becomes dominated by LocA and AssocA, making TRETA with mgReA more discriminative for evaluation. Additional validation experiments are provided in Section 1.2 of the supplementary material.

5.3. Experimental Analysis

Ablation study on two-stage association learning. Table 2 validates each component of our two-stage strategy. Using standard cross-entropy (CE) loss achieves 34.8% TETA on base classes, while replacing it with standard focal loss yields only marginal improvement (35.3% TETA), indicating that a fixed γ provides limited benefits for open-vocabulary tracking.

Our consistency-aware focal loss with dynamic factors brings substantial gains (rows 3-5): Incorporating appearance consistency γ_{app} alone improves TETA to 37.0%/36.4% (base/novel), with AssocA increasing from 47.2% to 51.5%, demonstrating that dynamically adjusting sample difficulty enables more effective emphasis on hard samples. Further adding semantic consistency γ_{sem} achieves 37.8%/37.4% TETA with 53.5% AssocA, validating that multi-modal consistency factors synergistically enhance association learning. Finally, Stage 2 self-supervised video training reaches 38.4%/38.3% TETA with 54.3%/62.8% AssocA, confirming our two-stage design effectively combines supervised image pair training and self-supervised video refinement to capture both static appearance and temporal dynamics.

Further Analysis. We also provide a comprehensive analysis of the dataset, a series of visual analyse, insights and limitations in the supplementary material.

6. Conclusion

In this work, we have proposed a novel yet practical problem GOVMOT. We build a large-scale and comprehensive benchmark, GOVTrackB, to provide the standard evaluation platform for this problem. Compared to similar competitor datasets, GOVTrackB has the advantage of containing more diverse and balanced object categories, along with significantly more testing samples for both base and novel classes, particularly the novel classes. Besides the dataset, we also design a new multi-granularity recognition metric and a simple yet effective baseline method with a two-stage association learning strategy. Extensive benchmark evaluations on numerous state-of-the-art methods have demonstrated the rationale of the proposed GOVMOT problem, and the usefulness of the GOVTrackB benchmark.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62402490 and 62572349.

References

[1] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. People-tracking-by-detection and people-detection-by-tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.

[2] 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 (WACV)*, 2023.

[3] 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 (CVPR)*, pages 6719–6728, 2021.

[4] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 941–951, 2019.

[5] Jiarui Cai, Mingze Xu, Wei Li, Yuanjun Xiong, Wei Xia, Zhuowen Tu, and Stefano Soatto. MeMOT: Multi-object tracking with memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8090–8100, 2022.

[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 (CVPR)*, pages 9686–9696, 2023.

[7] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research (JMLR)*, 25(70):1–53, 2024.

[8] Yutao Cui, Chenkai Zeng, Xiaoyu Zhao, Yichun Yang, Gangshan Wu, and Limin Wang. SportsMOT: A large multi-object tracking dataset in multiple sports scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9921–9931, 2023.

[9] Achal Dave, Pavel Tokmakov, and Deva Ramanan. Towards segmenting anything that moves. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2019.

[10] Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. TAO: A large-scale benchmark for tracking any object. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 436–454, 2020.

[11] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. MOT20: A benchmark for multi-object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*, 2020.

[12] Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 370–386, 2018.

[13] 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.

[14] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14084–14093, 2022.

[15] Yunhao Du, Zhicheng Zhao, Yang Song, Yanyun Zhao, Fei Su, Tao Gong, and Hongying Meng. StrongSORT: Make DeepSORT great again. *IEEE Transactions on Multimedia (TMM)*, 2023.

[16] Christiane Fellbaum. WordNet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer, 2010.

[17] James Ferryman and Ali Shahrokni. PETS2009: Dataset and challenge. In *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, pages 1–6, 2009.

[18] Tobias Fischer, Thomas E Huang, Jiangmiao Pang, Linlu Qiu, Haofeng Chen, Trevor Darrell, and Fisher Yu. QDTrack: Quasi-dense similarity learning for appearance-only multiple object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2023.

[19] Yiyang Gan, Ruize Han, Liqiang Yin, Wei Feng, and Song Wang. Self-supervised multi-view multi-human association and tracking. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pages 282–290, 2021.

[20] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361, 2012.

[21] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[22] Kuan-Chih Huang, Ming-Hsuan Yang, and Yi-Hsuan Tsai. Delving into motion-aware matching for monocular 3D object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6909–6918, 2023.

[23] László A Jeni, Jeffrey F Cohn, and Fernando De La Torre. Facing imbalanced data–recommendations for the use of performance metrics. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 245–251. IEEE, 2013.

[24] Prannay Kaul, Weidi Xie, and Andrew Zisserman. Multi-modal classifiers for open-vocabulary object detection. In

Proceedings of the International Conference on Machine Learning (ICML), pages 15946–15969, 2023.

[25] Jan Krejčí, Oliver Kost, Ondřej Straka, and Jindřich Duník. Pedestrian tracking with monocular camera using unconstrained 3D motion model. In *2024 27th International Conference on Information Fusion (FUSION)*, pages 1–8, 2024.

[26] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision (IJCV)*, 123:32–73, 2017.

[27] Laura Leal-Taixé, Anton Milan, Ian Reid, Stefan Roth, and Konrad Schindler. MOTChallenge 2015: Towards a benchmark for multi-target tracking. *arXiv preprint arXiv:1504.01942*, 2015.

[28] Laura Leal-Taixé, Cristian Canton-Ferrer, and Konrad Schindler. Learning by tracking: Siamese CNN for robust target association. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 33–40, 2016.

[29] Jinyang Li, En Yu, Sijia Chen, and Wenbing Tao. OVTR: End-to-end open-vocabulary multiple object tracking with transformer. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

[30] Siyuan Li, Martin Danelljan, Henghui Ding, Thomas E. Huang, and Fisher Yu. Tracking every thing in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[31] Siyuan Li, Tobias Fischer, Lei Ke, Henghui Ding, Martin Danelljan, and Fisher Yu. OVTrack: Open-vocabulary multiple object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5567–5577, 2023.

[32] Haiji Liang and Ruize Han. OVT-B: A new large-scale benchmark for open-vocabulary multi-object tracking. *Advances in Neural Information Processing Systems (NeurIPS)*, 37: 14849–14863, 2024.

[33] Chuang Lin, Peize Sun, Yi Jiang, Ping Luo, Lizhen Qu, Gholamreza Haffari, Zehuan Yuan, and Jianfei Cai. Learning object-language alignments for open-vocabulary object detection. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

[34] Chuang Lin, Yi Jiang, Lizhen Qu, Zehuan Yuan, and Jianfei Cai. Generative region-language pretraining for open-ended object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[35] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[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 (CVPR)*, pages 19045–19055, 2022.

[37] Jonathon Luiten, Tobias Fischer, and Bastian Leibe. Track to reconstruct and reconstruct to track. *IEEE Robotics and Automation Letters*, 5(2):1803–1810, 2020.

[38] Amalia Luque, Alejandro Carrasco, Alejandro Martín, and Ana de Las Heras. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, 91:216–231, 2019.

[39] Chuofan Ma, Yi Jiang, Xin Wen, Zehuan Yuan, and Xiaojuan Qi. CoDet: Co-occurrence guided region-word alignment for open-vocabulary object detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[40] 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 (CVPR)*, pages 8844–8854, 2022.

[41] Anton Milan, Seyed Hamid Rezatofighi, Anthony Dick, Konrad Schindler, and Ian Reid. Online multi-target tracking using recurrent neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), 2017.

[42] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[43] Aljoša Ošep, Wolfgang Mehner, Paul Voigtlaender, and Bastian Leibe. Track, then decide: Category-agnostic vision-based multi-object tracking. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3494–3501, 2018.

[44] Aljoša Ošep, Paul Voigtlaender, Mark Weber, Jonathon Luiten, and Bastian Leibe. 4d generic video object proposals. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10031–10037, 2020.

[45] Jiangmiao Pang, Linlu Qiu, Xia Li, Haofeng Chen, Qi Li, Trevor Darrell, and Fisher Yu. Quasi-dense similarity learning for multiple object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 164–173, 2021.

[46] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. KOSMOS-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

[47] Zekun Qian, Ruize Han, Junhui Hou, Linqi Song, and Wei Feng. VOVTrack: Exploring the potentiality in raw videos for open-vocabulary multi-object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7472–7482, 2025.

[48] Zekun Qian, Ruize Han, Zhixiang Wang, Junhui Hou, and Wei Feng. COVTrack: Continuous open-vocabulary tracking via adaptive multi-cue fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10054–10063, 2025.

[49] Zekun Qian, Ruize Han, Zhixiang Wang, Junhui Hou, and Wei Feng. DOVTrack: Data-efficient open-vocabulary tracking. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

[50] Zekun Qian, Wei Feng, Ruize Han, and Junhui Hou. COVTrack++: Learning open-vocabulary multi-object tracking

from continuous videos via a synergistic paradigm. *arXiv preprint arXiv:2603.24016*, 2026.

[51] Zheng Qin, Sanping Zhou, Le Wang, Jinghai Duan, Gang Hua, and Wei Tang. MotionTrack: Learning robust short-term and long-term motions for multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17939–17948, 2023.

[52] 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 *Proceedings of the International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021.

[53] Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Tracking the untrackable: Learning to track multiple cues with long-term dependencies. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 300–311, 2017.

[54] Fatemeh Saleh, Sadegh Aliakbarian, Hamid Rezatofighi, Mathieu Salzmann, and Stephen Gould. Probabilistic tracklet scoring and inpainting for multiple object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14329–14339, 2021.

[55] Sarthak Sharma, Junaid Ahmed Ansari, J Krishna Murthy, and K Madhava Krishna. Beyond pixels: Leveraging geometry and shape cues for online multi-object tracking. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3508–3515, 2018.

[56] 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.

[57] 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 (ICCV)*, pages 4057–4066, 2023.

[58] Li Wang, Xinyu Zhang, Wenyuan Qin, Xiaoyu Li, Jinghan Gao, Lei Yang, Zhiwei Li, Jun Li, Lei Zhu, Hong Wang, et al. CAMO-MOT: Combined appearance-motion optimization for 3D multi-object tracking with camera-LiDAR fusion. *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[59] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3645–3649, 2017.

[60] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 466–481, 2018.

[61] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5188–5197, 2019.

[62] 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 (CVPR)*, pages 2636–2645, 2020.

[63] Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. MOTR: End-to-end multiple-object tracking with transformer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 659–675, 2022.

[64] Libo Zhang, Junyuan Gao, Zhen Xiao, and Heng Fan. AnimalTrack: A benchmark for multi-animal tracking in the wild. *International Journal of Computer Vision (IJCV)*, 131 (2):496–513, 2023.

[65] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. ByteTrack: Multi-object tracking by associating every detection box. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[66] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 474–490, 2020.

[67] Xingyi Zhou, Tianwei Yin, Vladlen Koltun, and Philipp Krähenbühl. Global tracking transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8771–8780, 2022.

[68] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(11):7380–7399, 2022.

[69] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations (ICLR)*, 2021.