

# COVTrack: Continuous Open-Vocabulary Tracking via Adaptive Multi-Cue Fusion

Zekun Qian<sup>1,3</sup>, Ruize Han<sup>2†</sup>, Zhixiang Wang<sup>1</sup>, Junhui Hou<sup>3</sup>, Wei Feng<sup>1</sup>

<sup>1</sup>College of Intelligence and Computing, Tianjin University

<sup>2</sup>Shenzhen University of Advanced Technology

<sup>3</sup>City University of Hong Kong

{clarkqian, zhixiang.wang, wfeng}@tju.edu.cn, hanruize@suat-sz.edu.cn, jh.hou@cityu.edu.hk

<https://github.com/zekunqian/COVTrack>

## Abstract

*Open-Vocabulary Multi-Object Tracking (OVMOT) aims to detect and track diverse object categories in videos, including both seen (base) and unseen (novel) categories. Current methods rely on appearance features from generated image pairs or utilize the discontinuous annotations of the video dataset (TAO) for training, primarily due to the lack of available continuous annotated video datasets for OVMOT. This limitation affects their effectiveness, since continuous target trajectories are necessary for robust tracker learning. In this work, we propose the C-TAO dataset, which provides a continuous version of TAO, thereby constructing the first continuous annotated training dataset for OVMOT. This addresses the previous limitations in training data availability. Additionally, we introduce COVTrack, a unified framework that effectively integrates motion and semantic features with appearance features, in which the multi-cue feature aggregation strategy dynamically aggregates and balances these features, based on the confidence estimation from both intra-frame and inter-frame contexts. Our proposed framework significantly improves OVMOT performance, establishing COVTrack as a state-of-the-art solution on OVMOT benchmarks.*

## 1. Introduction

Open-Vocabulary Multi-Object Tracking (OVMOT) is a burgeoning problem in the computer vision area, which aims to localize, classify and track various categories of objects in real-world videos [22]. This problem has significant practical applications, *e.g.*, Internet video understanding, autonomous driving [15].

To develop an effective model for OVMOT, the quality of training datasets is crucial. However, due to the multi-task nature of this problem, collecting and annotating train-

ing videos is demanding. First, the collected videos must encompass a sufficiently large variety of object categories and diverse motion patterns. Second, comprehensive annotations are required, including object bounding boxes, as well as identities (for trajectory) and categories. However, existing large-scale datasets that provide a wide variety of object categories are predominantly image-based, *e.g.*, COCO [26], LVIS [14]. In contrast, video datasets often suffer from limited category diversity, focusing primarily on a small number of classes, such as humans, *e.g.*, MOT20 [7], DanceTrack [39], vehicles [12]. Neither of them can meet the requirements. To address this gap, TAO [6] is a recent dataset proposed for the multi-category object tracking problem, which contains 2,907 videos covering 833 categories. However, the annotation of TAO is not continuous, with the annotation frame rate being only 1 fps, resulting in critical issues: *not enough available frames to train a robust open-vocabulary tracker* and *the lack of continuous temporal information* needed to train the tracker effectively. As a result, existing methods for OVMOT, such as OVTrack [22], rely on large image-based datasets, *i.e.*, LVIS [14], to generate image pairs for training. This forces them to rely solely on appearance features for tracking, resulting in suboptimal performance. Some later works, such as SLAck [24], attempt to utilize the TAO dataset with its discontinuous annotations by constructing pseudo-labels for training. However, SLAck still fails to make full use of the continuous video information. As shown in Fig. 2, issues such as occlusions, viewpoint transitions, and appearance deformation lead to significant differences in annotated objects across large intervals (*e.g.*, 30 frames). Consequently, the trained model may not perform very well, as its features (*e.g.*, appearance, motion) are not well-suited for frame-by-frame inference during testing.

From the above observation, in this work, *we aim to newly build and fully leverage the continuous data for OVMOT problem.* To address the limitations of exist-

<sup>†</sup>Corresponding author.

ing datasets, we introduce the C-TAO (Continuous TAO) dataset, which enhances the original TAO dataset. We manually label bounding boxes, category labels, and their IDs on frames that lack annotations, creating complete and continuous trajectories while significantly increasing the amount of available training data. Specifically, C-TAO enlarges the annotations of TAO by over 26 times (in terms of annotated frames). The continuity of C-TAO is obviously improved, which can be reflected by the significantly decreased object area change and distance movement, between two annotated targets. With the C-TAO dataset established, we can thus leverage this continuous data to extract multi-cue features for OVMOT. Specifically, besides classical appearance features [22], benefiting from the C-TAO dataset, we also incorporate motion features, which are commonly used in traditional multi-object tracking (MOT). Furthermore, we utilize object category-related semantic features as additional cues for tracking, distinguishing our approach from classical MOT methods that typically focus on a single category. To enhance the utility of these multi-cue features, we develop a novel multi-cue confidence estimation strategy to balance and aggregate them. This strategy includes intra-frame confidence to manage multiple cues within a single frame, and the inter-frame confidence derived from a cycle-consistency based self-correction mechanism across adjacent frames. The main contributions of this work are:

- We build C-TAO, to the best of our knowledge, which is the first continuously-labeled (frame-by-frame) training dataset for OVMOT. It is also valuable for all kinds of other multi-category MOT tasks.
- We develop a multi-cue (appearance, motion, and semantic) feature aggregation framework, in which we design a novel multi-cue confidence estimation strategy to dynamically and adaptively balance the multi-cue features.
- Comparisons of the same methods trained on TAO and C-TAO demonstrate the proposed dataset’s usefulness. Additionally, experiments on both datasets verify our method’s effectiveness, achieving a 6.5% improvement in OVMOT performance (novel TETA) over the baseline OVTrack [22].

## 2. Related Work

**Multiple Object Tracking.** The dominant paradigm in multiple object tracking (MOT) is the tracking-by-detection framework [1, 10, 16], where detections are linked over time through association. Early studies emphasized appearance features [2, 4, 11, 20, 32, 35, 42] for re-identification, serving as the primary cue for object association. Motion information also plays a critical role, with techniques utilizing Kalman Filtering [9, 33, 36, 43, 47] for trajectory prediction and 3D motion features [18, 19, 28, 31, 40] for enhanced dynamics capture. However, motion-based meth-

ods like SORT [3] struggle with rapid motion and severe occlusions.

Hybrid approaches combine appearance and motion cues for robust tracking. DeepSORT [42] enhances appearance-based associations with motion priors, while dual-branch architectures [37, 41, 45, 46] and transformer-based methods [29, 38, 44] integrate these cues through feature fusion, addressing challenges like occlusion. However, these methods focus primarily on simple categories (*e.g.*, humans) and lack generalization to diverse object categories. Moreover, when extended to open-vocabulary scenarios, the varying reliability of different cues across novel categories presents new challenges for effective feature fusion.

In this work, besides the commonly used appearance and motion, we also use category-aware semantic cues for OVMOT, which is specific to this problem and different from classical MOT. To effectively leverage different types of cues, we develop a novel feature fusion strategy to dynamically balance and integrate the appearance, motion and semantic cues.

**Open-World/Vocabulary MOT.** To expand the object categories in MOT, the TAO benchmark [6] was introduced to evaluate tracking under a long-tail class distribution. Methods like AOA [8], GTR [48], TET [21], and QD-Track [11] have been developed for generic object tracking on TAO. However, these methods are limited to predefined categories and cannot handle novel class objects absent from the training set.

To address this limitation, open-world MOT was proposed, aiming to detect and track objects not seen in the training set. Early works focused on class-agnostic detection and tracking [30, 31], while the TAO-OW benchmark [27] was introduced to evaluate open-world tracking. However, TAO-OW relies on class-agnostic metrics, which fail to identify the specific classes of unknown objects. A related problem, Open-Vocabulary Multi-Object Tracking (OVMOT), was proposed in [22], with OVTrack extending the tracking framework to open-vocabulary settings. Similarly, MASA [23] leveraged unlabeled image pairs to learn universal appearance models. However, both OVTrack and MASA rely heavily on appearance-based strategies, limiting their generalization to novel categories. SLAck [24] advanced the field by integrating semantic, motion, and appearance features, eliminating the need for heuristic post-processing. However, SLAck’s feature fusion via simple summation operation fails to fully exploit the complementary potential of semantic and motion information. In contrast, this work proposes a more efficient feature fusion method with intra- and inter-frame confidence, integrating semantic and motion cues into appearance-based representations. Besides, the proposed C-TAO dataset further enhances tracking performance for both our and other methods in diverse open-vocabulary scenarios.

### 3. C-TAO: Continuous TAO Training Dataset

#### 3.1. Motivation: Training Challenges with TAO

For the task of OVMOT, only two datasets exist, *i.e.*, TAO and OVT-B [25], in which OVT-B is a test dataset. The TAO dataset emerges as a uniquely valuable video resource for training due to its extensive category coverage, vastly surpassing traditional MOT datasets limited to specific categories (person, vehicle, *etc.*). However, its *sparse annotation strategy*—annotations provided every 30 frames—poses significant training challenges. The large temporal gaps hinder learning accurate motion patterns, as the model cannot observe smooth trajectories between annotated frames. Additionally, the lack of intermediate frame annotations makes it difficult to capture fine-grained appearance changes and effectively handle occlusions. To compensate, researchers often supplement training with image datasets like LVIS [14], constructing synthetic image pairs [22, 23] to simulate adjacent frames. However, these pairs fail to simulate continuous motion and realistic appearance changes, limiting model performance. Recent SLAck [24] generates pseudo labels via IoU matching to create continuous annotations in TAO, showing substantial improvements over image-pair-based methods. Yet, IoU matching lacks temporal consistency, especially in dynamic scenarios, leading to unreliable motion features. To address these limitations, a continuously annotated version of the TAO training set is urgently needed. Such a dataset would facilitate more effective training of OVMOT, potentially achieving new performance levels in OVMOT tasks.

#### 3.2. Construction of C-TAO Dataset

To create a continuously annotated high-quality dataset, we retain all videos and annotated trajectories from the TAO training set to preserve the original diversity in categories, scenes, and motion patterns. The re-annotation process follows a rigorous three-stage pipeline: First, professional annotators label the bounding boxes and identify target objects for each frame based on the original (discontinuous) trajectories, with independent verification to ensure quality and temporal consistency. Second, we label the categories for each tracked object according to the ground-truth categories of the original annotations, accompanied by additional verification for quality assurance. Finally, we implement a federated annotation protocol that utilizes cross-validation among multiple annotators to verify the completeness of annotations. This process ensures that all instances within each trajectory are labeled throughout the video.

#### 3.3. Comparison and Analysis

**Statistical Comparison.** We compare our C-TAO dataset with the original TAO dataset in terms of annotation density, coverage, and continuity, as shown in Fig. 1.

Both datasets share the same videos and trajectories, C-TAO significantly enhances annotation density, with total annotated frames and bounding boxes increasing over 26 and 27 times, respectively, as illustrated in the left subplot. At the video level, C-TAO shows a dramatic improvement in annotated frame coverage and the number of annotations per video, ensuring a more comprehensive capture of object dynamics and scene evolution, as shown in the center subplot. For trajectory-level statistics, C-TAO exhibits denser temporal sampling, with remarkable increases in frames per track and annotations per track, providing more complete object motion information. Regarding continuity and temporal consistency, we analyze three key factors between consecutive annotated frames: average IoU, area change ratios, and pixel movement of the object center, shown in the right subplot. Results indicate significant improvements: cross-frame IoU increases by 2.2 times, area change rate reduces to 1/8 of the original, and average pixel displacement decreases from 101 to 6.6 pixels, a 93.5% reduction. These factors collectively demonstrate a substantial enhancement in the overall temporal coherence of the dataset.

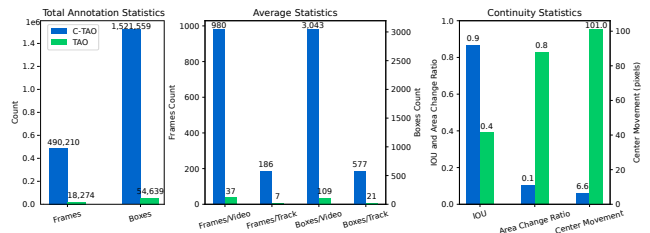


Figure 1. Visualization of annotation statistics comparison between our dataset (C-TAO) and TAO. Left: total number of annotated frames and bounding boxes. Middle: average statistics per video and per track. Right: continuity statistics between consecutive annotated frames.

**Qualitative Analysis.** To further demonstrate the advantages of C-TAO, we present three representative scenarios in Fig. 2. These examples illustrate how our continuous annotations capture crucial intermediate states that are missing in the original TAO dataset’s sparse annotations. First, as shown in Fig. 2(a), for the very common occlusion cases, our annotations capture the complete progression of the target moving through the occlusion, providing required training samples for handling complex occlusion scenes. Second, as shown in Fig. 2(b), under significant camera motion, the dense annotations reveal smooth viewpoint transitions that bridge the large perspective gaps present in sparse annotations. Third, as shown in Fig. 2(c), for objects undergoing appearance changes, our annotations record the continuous evolution of the target appearance, not only the object before and after the significant shape deformations.

Qualitative examples and statistical improvements demonstrate how our densely annotated data enhances the original TAO by providing more effective temporal infor-

mation, which is crucial for MOT systems to learn robust tracking features, especially in challenging scenarios.

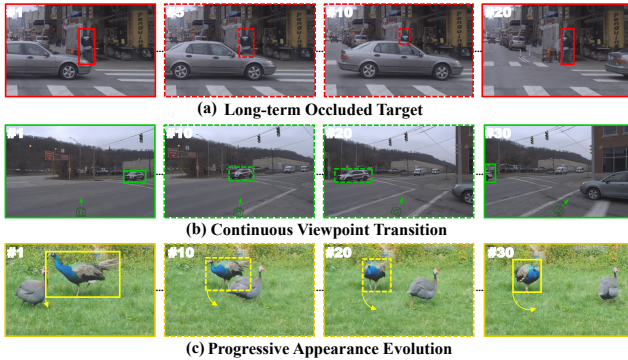


Figure 2. Annotation examples in challenging scenarios. Solid boxes represent original TAO annotations (30-frame intervals), while dashed boxes show our continuous annotations. (a) Occlusion process of a pedestrian behind a car. (b) Viewpoint changes of a vehicle under camera motion. (c) Appearance evolution during a bird’s pose transformation. C-TAO annotations capture crucial intermediate states that are missed in the original sparse annotations.

## 4. Proposed Method

### 4.1. Overview

As shown in Fig. 3, based on a pre-trained OV detector, we extract multiple tracking cues for each object, including the *appearance, location, and semantic features for object tracking*. Section 4.2 elaborates on the extraction process of each feature cue. In Section 4.3, we describe how to balance the confidence of different cues using intra- and inter-frame information. Section 4.4 presents the methodology for effectively incorporating location and semantic cues into appearance features. The training loss and gradient propagation paths are presented in Section 4.5.

### 4.2. Appearance, Location and Semantic Cues

For high efficiency, we extract multi-cue features using the backbone of a pre-trained detector. Specifically, we adopt the same detector as in OVTrack [22] and SLAck [24] for OV detection and freeze it during training. Based on it, as shown in Fig. 3, we construct the features for association from appearance, location, and semantic cues.

**Appearance head.** To capture visual details for effective object association, we extract Region-of-Interest (RoI) features from the detection proposals. These features are subsequently processed by a lightweight convolutional module followed by an MLP. The resulting transformation produces the  $i$ -th object’s appearance embedding,  $\mathbf{e}_{\text{app}}^i \in \mathbb{R}^d$ .

**Location head.** The location head focuses on encoding the spatial attributes of detected objects. Given a bounding box, we directly normalize its coordinates with respect to the image dimensions. The normalized values capture the object’s relative spatial position in a scale-invariant man-

ner, which are then input into a dedicated projection layer to yield the  $i$ -th object’s location embedding,  $\mathbf{e}_{\text{loc}}^i \in \mathbb{R}^d$ .

**Semantic head.** To produce versatile semantic representations without re-training, we initially consider leveraging CLIP [34] but opt to distill its semantic head due to high inference costs. Following a distillation process similar to [13], we fuse CLIP’s text and image embeddings via element-wise summation, refine them with an MLP, and obtain the final  $i$ -th object’s semantic embedding  $\mathbf{e}_{\text{sem}}^i \in \mathbb{R}^d$ . This dual-path design with both textual context and visual details for semantic features is verified in [13, 22].

By explicitly modeling appearance, location, and semantic cues, our framework constructs a rich feature representation for the open-vocabulary tracking (association) task.

### 4.3. Intra/Inter-Frame Cue Confidence

Appearance is the most commonly used feature in previous OVMOT and MOT tasks. While appearance features serve as a solid foundation, incorporating semantic and location cues can significantly enhance the discriminative capability. However, effectively fusing these features from different cues presents significant challenges.

Given the three feature embeddings  $\mathbf{e}_{\text{app}}^i$ ,  $\mathbf{e}_{\text{loc}}^i$ , and  $\mathbf{e}_{\text{sem}}^i$  extracted in Section 4.2, a straightforward fusion approach, as adopted in SLAck [24], is direct summation:  $\mathbf{e}_{\text{fused}}^i = \mathbf{e}_{\text{app}}^i + \mathbf{e}_{\text{loc}}^i + \mathbf{e}_{\text{sem}}^i$ . While simple, this equal-weighting strategy faces substantial challenges in OVMOT. Since object localization and classification are inherently difficult tasks, the reliability of location and semantic cues varies significantly across different scenarios. This reliability variation becomes particularly problematic for novel categories, where classification accuracy often remains in single digits. Consequently, such naive feature fusion can introduce considerable noise and destabilize tracking performance.

To address these fusion reliability issues, we propose a dual-perspective confidence estimation strategy that dynamically determines the contribution weights of different cues. Our approach evaluates feature reliability from two complementary perspectives: First, *intra-frame confidence* assesses the reliability of semantic and location features within a single frame by learning their mutual relationships with appearance features. Second, *inter-frame confidence* leverages temporal consistency between adjacent frames to evaluate the stability of each feature type across time. This dual-perspective design enables adaptive feature weighting that accounts for both spatial and temporal reliability variations.

**Intra-frame cue confidence.** The  $i$ -th object’s intra-frame cue confidence, denoted as  $c_{\text{sem},i}^{\text{intra}}$  and  $c_{\text{loc},i}^{\text{intra}}$ , is computed using the Self-attentive Gated Network (SGN). This network evaluates the relative reliability of the semantic and location cues within a single frame, ensuring that their impacts on the final fused feature are adaptively weighted.

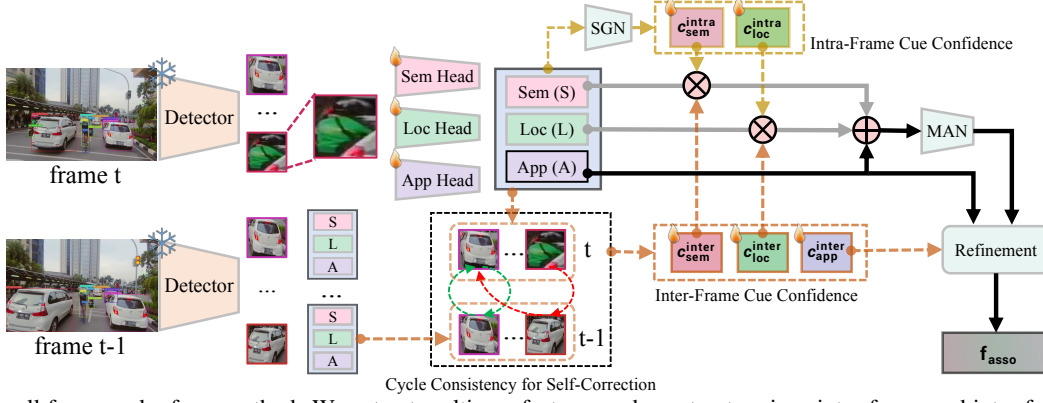


Figure 3. Overall framework of our method. We extract multi-cue features and construct various intra-frame and inter-frame confidence scores that effectively facilitate multi-cue feature fusion.  $\otimes$  and  $\oplus$  represent the scaling multiplication and concatenation operation.

As shown at the top of Fig. 3, on the current frame  $t$ , the network takes three cues as input, which interact at the feature level through concatenation operations. Then the network learns the *mutual confidence relationships* between these cues to construct effective intra-frame confidence scores of the  $i$ -th object as

$$[c_{loc,i}^{\text{intra}}, c_{sem,i}^{\text{intra}}] = \text{sigmoid}(\text{SGN}(\text{concat}(\mathbf{e}_{\text{app}}^i, \mathbf{e}_{loc}^i, \mathbf{e}_{sem}^i))), \quad (1)$$

where SGN is designed as two fully connected layers with a ReLU activation, and the sigmoid activation ensures that the gate values  $c_{loc,i}^{\text{intra}}$  and  $c_{sem,i}^{\text{intra}}$  are within the range  $[0, 1]$ .

These intra-frame confidence scores *dynamically adjust the (relative) influence of the location and semantic cues for appearance*. Higher confidence values indicate that the corresponding cue is more reliable and should contribute more to the final fused feature. This mechanism ensures that the fusion process is robust to noise and uncertainty, from a single-frame level.

**Inter-frame cue confidence.** To evaluate the temporal reliability of different cues across continuous frames, as shown in the lower part of Fig. 3, we design an inter-frame cycle consistency estimation module. Given a feature of appearance, location, and semantic, here we take the appearance for example, let  $\mathbf{E}_{\text{app}}^t \in \mathbb{R}^{n \times d}$  and  $\mathbf{E}_{\text{app}}^{t-1} \in \mathbb{R}^{m \times d}$  represent the feature matrices at frames  $t$  and  $t-1$  respectively, where  $n$  and  $m$  denote the number of objects in each frame, and  $d$  is the feature dimension. We first compute the pairwise similarity matrix between objects in consecutive frames as  $\mathbf{S}_{\text{app}} = \mathbf{E}_{\text{app}}^t \cdot (\mathbf{E}_{\text{app}}^{t-1})^\top \in \mathbb{R}^{n \times m}$ . To enhance the discriminative power of similarity scores, we apply adaptive temperature scaling  $\alpha = \frac{\log(\frac{\delta}{1-\delta} \cdot \max(n,m))}{\epsilon}$ , where  $\alpha$  is the temperature parameter,  $\delta$  and  $\epsilon$  are hyperparameters. The scaled similarity scores are normalized using softmax as  $\hat{\mathbf{S}}_{\text{app}} = \text{softmax}(\alpha \cdot \mathbf{S}_{\text{app}})$ . Next, we implement the cycle consistency for inter-frame confidence estimation. Specifically, we compute the cycle consistency matrix as

$$\mathbf{C}_{\text{app}}^{\text{cycle}} = \hat{\mathbf{S}}_{\text{app}} \cdot (\hat{\mathbf{S}}_{\text{app}})^\top. \quad (2)$$

The rationale of the inter-frame confidence estimation is to leverage the *self-correction mechanism*. Specifically, by computing the feature cycle consistency between different frames from each cue, higher confidence reflects the corresponding cue has stronger stability and usability, thus accounting for more impact during the feature fusion. Finally, for each object  $i$  at frame  $t$ , we extract its confidence score from  $\hat{\mathbf{C}}_{\text{app}}$  as  $c_{\text{app},i}^{\text{inter}} = \left[ \text{diag}(\mathbf{C}_{\text{app}}^{\text{cycle}}) \right]_i$ ,  $i \in \{1, \dots, n\}$ .

Similarly, this pipeline can be applied to location and semantic features to obtain  $c_{loc,i}^{\text{inter}}$  and  $c_{sem,i}^{\text{inter}}$  respectively.

#### 4.4. Multi-Cue Feature Fusion

**Multi-cue feature aggregation.** With the above confidence, for the  $i$ -th object, we construct a multi-cue feature representation by combining the original appearance feature with confidence-weighted location and semantic features, which is modulated by their corresponding intra-frame and inter-frame confidence scores as

$$\tilde{\mathbf{e}}_{loc}^i = c_{loc,i}^{\text{intra}} \cdot c_{loc,i}^{\text{inter}} \cdot \mathbf{e}_{loc}^i, \quad \tilde{\mathbf{e}}_{sem}^i = c_{sem,i}^{\text{intra}} \cdot c_{sem,i}^{\text{inter}} \cdot \mathbf{e}_{sem}^i. \quad (3)$$

After that, these features are concatenated to the appearance to form a multi-cue representation, which is input into a Multi-cue Aggregation Network (MAN) to integrate information from all cues for producing the integrated feature with the original embedding dimension

$$\mathbf{f}_{\text{m-cue}}^i = \text{MAN}(\text{concat}(\mathbf{e}_{\text{app}}^i, \tilde{\mathbf{e}}_{loc}^i, \tilde{\mathbf{e}}_{sem}^i)) \in \mathbb{R}^d, \quad (4)$$

where MAN is designed as two fully connected layers with a ReLU activation. This aggregated feature  $\mathbf{f}_{\text{m-cue}}^i$  effectively encodes the appearance, location, and semantic information with their intra-/inter-frame relationships.

**Feature refinement with re-connection.** In the second stage, we consider the usage of the multi-cue aggregated feature  $\mathbf{f}_{\text{m-cue}}^i$  and the main appearance feature  $\mathbf{e}_{\text{app}}^i$ . Specifically, using the inter-frame confidence score of appearance features  $c_{\text{app},i}^{\text{inter}}$  as a guidance, we adaptively combine the  $\mathbf{f}_{\text{m-cue}}^i$  with  $\mathbf{e}_{\text{app}}^i$  as the final fused association feature  $\mathbf{f}_{\text{asso}}^i$

for the association in the OVMOT framework:

$$\mathbf{f}_{\text{asso}}^i = c_{\text{app},i}^{\text{inter}} \cdot \mathbf{e}_{\text{app}}^i + (1 - c_{\text{app},i}^{\text{inter}}) \cdot \mathbf{f}_{\text{m-cue}}^i \in \mathbb{R}^d. \quad (5)$$

This refinement mechanism adaptively balances between the original appearance feature and the aggregated feature based on the *appearance self-correction*. When appearance features demonstrate strong temporal consistency (high  $c_{\text{app},i}^{\text{inter}}$ ), the model places greater emphasis on  $\mathbf{e}_{\text{app}}^i$ . Conversely, when appearance features show weak temporal consistency (low  $c_{\text{app},i}^{\text{inter}}$ ), the model relies more on the information-rich aggregated feature  $\mathbf{f}_{\text{m-cue}}^i$  which incorporates multiple complementary cues. This adaptive mechanism ensures robust feature fusion by dynamically adjusting the impacts of each cue.

#### 4.5. Association Loss

To train the feature fusion framework, we compute the similarity between objects in consecutive frames using the final fused association features, *i.e.*,  $\mathbf{F}_{\text{asso}}^t \in \mathbb{R}^{n \times d}$  and  $\mathbf{F}_{\text{asso}}^{t-1} \in \mathbb{R}^{m \times d}$ , as  $\mathbf{S} = \mathbf{F}_{\text{asso}}^t \cdot (\mathbf{F}_{\text{asso}}^{t-1})^\top \in \mathbb{R}^{n \times m}$ . The association loss is formulated as

$$\mathcal{L}_{\text{asso}} = - \sum_{i,j} \mathbf{Y}_{ij} \log(\hat{\mathbf{S}}_{ij}), i \in \{1, \dots, n\}, j \in \{1, \dots, m\} \quad (6)$$

where  $\mathbf{Y}$  is the ground-truth association label, and  $\hat{\mathbf{S}}$  represents the normalized  $\mathbf{S}$  with softmax operation.

**Training Analysis.** The association loss enables end-to-end training of all confidence scores through backpropagation. Specifically, five learnable confidence scores are trained in different stages of feature fusion. During the multi-cue feature aggregation stage, four confidence scores work collaboratively to adjust the location and semantic features, including two intra-frame confidence scores ( $c_{\text{loc}}^{\text{intra}}$ ,  $c_{\text{sem}}^{\text{intra}}$ ) and two inter-frame scores ( $c_{\text{loc}}^{\text{inter}}$ ,  $c_{\text{sem}}^{\text{inter}}$ ). In the refinement stage, the inter-frame appearance confidence  $c_{\text{app}}^{\text{inter}}$  guides the adaptive refinement. During training, these scores are optimized through the continuous gradient flow as  $\frac{\partial \mathcal{L}_{\text{asso}}}{\partial c} = \frac{\partial \mathcal{L}_{\text{asso}}}{\partial \mathbf{S}} \cdot \frac{\partial \mathbf{S}}{\partial \mathbf{f}_{\text{asso}}} \cdot \frac{\partial \mathbf{f}_{\text{asso}}}{\partial \mathbf{f}_{\text{m-cue}}^i} \cdot \frac{\partial \mathbf{f}_{\text{m-cue}}^i}{\partial \mathbf{e}} \cdot \frac{\partial \mathbf{e}}{\partial c}$ . This fully differentiable design ensures that all confidence scores can be effectively trained. Through back-propagation, all the confidence weights are *learned to be optimally balanced among different cues, without any explicit supervision*. The end-to-end training allows these scores to automatically adapt to the mutual relation of different cues and the temporal self-correction along the video.

#### 4.6. Implementation Details

Our model employs the same backbone architecture as [22, 24], utilizing a Faster R-CNN detector with ResNet-50 [17]. Following OVTrack’s training protocol, the detector is trained on base classes from the LVIS dataset. For association, we utilize our proposed C-TAO dataset for training,

where  $\delta$  and  $\epsilon$  in Eq. (2) are set to 0.5 and 0.1, respectively. The model is trained for 10 epochs on 4 RTX 3090 GPUs.

For inference, we employ class-agnostic NMS for object filtering, with a maximum of 80 detected objects per frame. We use  $\mathbf{f}_{\text{asso}}$  to associate objects and adopt the bi-softmax matching strategy as [22]. The matching process uses a threshold of 0.35 and a memory queue length of 30.

## 5. Experiments

### 5.1. Datasets and Metrics

Following [22, 24], we conduct our evaluation using the same dataset and metrics. Specifically, we utilize the dataset TAO, which shares a similar category division scheme with LVIS [14] for OVMOT evaluation. We designate the rare categories in LVIS as novel classes, while the remaining categories serve as base classes. Comparative experiments are conducted on the validation and test sets of TAO. For performance evaluation, we adopt the standard OVMOT metric, tracking-everything accuracy (TETA) [21], which evaluates localization accuracy (LocA), classification accuracy (ClsA), and association accuracy (AssocA). To comprehensively demonstrate our algorithm’s performance, we evaluate base and novel classes respectively.

### 5.2. Comparison to State-of-the-Arts

We compare our method with current mainstream and state-of-the-art tracking methods on both the validation and test sets of TAO. For a fair comparison, all methods utilize ResNet-50 as the backbone architecture. The comparison includes closed-set baselines trained on all categories, established off-the-shelf trackers such as ByteTrack [46], OC-SORT [5], and MASA [23], as well as specialized OVMOT methods like OVTrack [22], and the current state-of-the-art method, SLack [24]. As shown in Table 1, we can first see that our method significantly outperforms all other methods in terms of TETA on both the validation and test sets. Notably, we achieve improvements of 3.5% and 4.4% in novel AssocA and base AssocA, respectively, compared to the state-of-the-art SLack. On the test set, our method maintains superior performance, with substantial improvements in AssocA - surpassing SLack by 6.5% for base classes and 2.6% for novel classes, demonstrating the effectiveness of our approach for the temporal tracking problem.

### 5.3. Ablation Study

**Effectiveness of cue confidence.** The first five rows in Table 2 demonstrate that the proposed cue confidence mechanisms effectively enhance association performance. We can see that the motion-related cues show a more significant impact compared to semantic-related cues, which validates the necessity of using continuous TAO for learning better temporal continuity features. Furthermore, rows six and seven

Table 1. Comparison of tracking performance on validation and test sets. We compare the methods on open-vocabulary TAO benchmark [22]. All methods use the same backbone. † represents using the same detector.

| Method                | Venue (year) | Classes |      | Novel       |             |             |            | Base        |             |             |             |
|-----------------------|--------------|---------|------|-------------|-------------|-------------|------------|-------------|-------------|-------------|-------------|
|                       |              | Novel   | Base | TETA        | LocA        | AssocA      | ClsA       | TETA        | LocA        | AssocA      | ClsA        |
| <b>Validation set</b> |              |         |      |             |             |             |            |             |             |             |             |
| QDTrack [11]          | TPAMI (2023) | ✓       | ✓    | 22.5        | 42.7        | 24.4        | 0.4        | 27.1        | 45.6        | 24.7        | 11.0        |
| TETer [21]            | ECCV (2022)  | ✓       | ✓    | 25.7        | 45.9        | 31.1        | 0.2        | 30.3        | 47.4        | 31.6        | 12.1        |
| DeepSORT (ViLD) [42]  | ICIP (2017)  | -       | ✓    | 21.1        | 46.4        | 14.7        | 2.3        | 26.9        | 47.1        | 15.8        | 17.7        |
| Tracktor++ (ViLD) [2] | ICCV (2019)  | -       | ✓    | 22.7        | 46.7        | 19.3        | 2.2        | 28.3        | 47.4        | 20.5        | 17.0        |
| ByteTrack† [46]       | ECCV (2022)  | -       | ✓    | 22.0        | 48.2        | 16.6        | 1.0        | 28.2        | 50.4        | 18.1        | 16.0        |
| OC-SORT† [5]          | CVPR (2023)  | -       | ✓    | 23.7        | 49.6        | 20.4        | 1.1        | 28.9        | 51.4        | 19.8        | 15.4        |
| OVTrack† [22]         | CVPR (2023)  | -       | ✓    | 27.8        | 48.8        | 33.6        | 1.5        | 35.5        | 49.3        | 36.9        | <b>20.2</b> |
| MASA (R50)† [23]      | CVPR (2024)  | -       | -    | 30.0        | 54.2        | 34.6        | 1.0        | 36.9        | 55.1        | 36.4        | 19.3        |
| SLAck† [24]           | ECCV (2024)  | -       | ✓    | 31.1        | 54.3        | 37.8        | 1.3        | 37.2        | 55.0        | 37.6        | 19.1        |
| Ours†                 | -            | -       | ✓    | <b>34.3</b> | <b>58.2</b> | <b>41.3</b> | <b>3.5</b> | <b>39.6</b> | <b>57.3</b> | <b>42.0</b> | 19.6        |
| <b>Test set</b>       |              |         |      |             |             |             |            |             |             |             |             |
| QDTrack [11]          | TPAMI (2023) | ✓       | ✓    | 20.2        | 39.7        | 20.9        | 0.2        | 25.8        | 43.2        | 23.5        | 10.6        |
| TETer [21]            | ECCV (2022)  | ✓       | ✓    | 21.7        | 39.1        | 25.9        | 0.0        | 29.2        | 44.0        | 30.4        | 10.7        |
| DeepSORT (ViLD) [42]  | ICIP (2017)  | -       | ✓    | 17.2        | 38.4        | 11.6        | 1.7        | 24.5        | 43.3        | 14.6        | 15.2        |
| Tracktor++ (ViLD) [2] | ICCV (2019)  | -       | ✓    | 18.0        | 39.0        | 13.4        | 1.7        | 26.0        | 44.1        | 19.0        | 14.8        |
| OVTrack† [22]         | CVPR (2023)  | -       | ✓    | 24.1        | 41.8        | 28.7        | 1.8        | 32.6        | 45.6        | 35.4        | 16.9        |
| SLAck† [24]           | ECCV (2024)  | -       | ✓    | 27.1        | 49.1        | 30.0        | 2.0        | 34.7        | 52.5        | 35.6        | 16.1        |
| Ours†                 | -            | -       | ✓    | <b>28.9</b> | <b>50.9</b> | <b>32.6</b> | <b>3.3</b> | <b>37.9</b> | <b>54.5</b> | <b>42.1</b> | <b>17.2</b> |

Table 2. Ablation study results on the validation set. We compare the results of different ablation methods on both novel and base classes.

| Ablation Method   | Novel       |             |             |            | Base        |             |             |             |
|---|-------------|-------------|-------------|------------|-------------|-------------|-------------|-------------|
|   | TETA        | LocA        | AssocA      | ClsA       | TETA        | LocA        | AssocA      | ClsA        |
| ① w/o $c_{sem}^{intra}$                                 | 33.3        | 57.3        | 40.6        | 2.1        | 39.1        | 57.3        | 41.2        | 18.9        |
| ② w/o $c_{loc}^{intra}$                                 | 32.8        | 57.6        | 37.7        | 3.1        | 38.7        | 57.4        | 39.5        | 19.1        |
| ③ w/o $c_{sem}^{inter}$                                 | 33.1        | 57.8        | 39.1        | 2.3        | 38.6        | 57.6        | 39.6        | 18.6        |
| ④ w/o $c_{loc}^{inter}$                                 | 32.8        | 57.8        | 37.3        | 3.2        | 38.2        | 57.4        | 38.4        | 18.8        |
| ⑤ w/o $c_{app}^{inter}$                                 | 33.4        | 57.8        | 39.2        | 3.3        | 38.6        | 57.4        | 39.2        | 19.1        |
| ⑥ w/o $c_{sem}^{intra} + c_{loc}^{intra}$ (intra conf.) | 32.3        | 57.7        | 36.8        | 2.3        | 38.5        | 57.4        | 39.3        | 18.8        |
| ⑦ w/o $c_{sem}^{inter} + c_{loc}^{inter}$ (inter conf.) | 32.1        | 57.7        | 35.7        | 2.8        | 37.8        | 57.7        | 37.3        | 18.3        |
| ⑧ w/o $e_{app}$ in Eq. (1)                              | 32.0        | 57.4        | 36.5        | 2.2        | 38.4        | 57.3        | 39.2        | 18.7        |
| ⑨ w/o $\tilde{e}_{sem}$ in Eq. (4)                      | 33.0        | 57.0        | 39.7        | 2.4        | 38.2        | 57.1        | 40.1        | 17.5        |
| ⑩ w/o $\tilde{e}_{loc}$ in Eq. (4)                      | 33.0        | 57.5        | 38.4        | 3.2        | 38.3        | 57.1        | 39.2        | 18.5        |
| ⑪ w/o $e_{app}$ in Eq. (5)                              | 33.3        | 57.8        | 39.1        | 2.9        | 38.6        | 57.6        | 39.4        | 18.7        |
| Ours  | <b>34.3</b> | <b>58.2</b> | <b>41.3</b> | <b>3.5</b> | <b>39.6</b> | <b>57.3</b> | <b>42.0</b> | <b>19.6</b> |

reveal that both the intra-frame and inter-frame cue confidence mechanisms, as an entirety, contribute more significantly to improving the features’ discriminative ability.

**Effectiveness of multi-cue fusion.** We then analyze the effectiveness of feature fusion. Row 8 validates the crucial role of appearance features as input during the intra-frame confidence learning process in SGN in Eq. (1). The absence of appearance features for object representations leads to a substantial decrease in confidence estimation effectiveness, particularly for novel classes. Additionally, the next two rows confirm that both semantic and location features in Eq. (4) contribute to the performance improvement, with location features showing a more important impact on the association. The last row also reveals that refinement operation in Eq. (5) can effectively improve association results.

**Comparisons of feature fusion methods.** In Table 3, the first four rows compare the feature fusion methods of ours and SLAck’s direct feature addition manner. The first two rows reveal that while SLAck’s association per-

formance deteriorates significantly when trained on original sparse TAO data, our method maintains robust performance by effectively balancing the weights between different cues, achieving a 10.7% improvement in AssocA over SLAck. In rows three and four, we conduct ablation experiments using SLAck’s pseudo-label generation through IoU-based matching between adjacent frames to provide continuous labels. Our method also outperforms SLAck in tracking performance. These comparisons validate the effectiveness and generalization of our method, showing substantial improvements regardless of whether using sparse annotations or continuous pseudo labels as supervision.

**Effectiveness of C-TAO.** Moreover, to validate the effectiveness and usefulness of the proposed C-TAO dataset, we evaluate the performance of various OVMOT methods trained with various data. We can see from the last row of Table 3 that our method using C-TAO annotations achieves the 6.5% and 6.2% improvement in novel and base AssocA, respectively, compared to the original TAO.

Table 3. Comparison of SLAck and our method using different supervisions for training on TAO/C-TAO.

| Method   | Novel |      |              |      | Base |      |             |      |
|--|-------|------|--------------|------|------|------|-------------|------|
|  | TETA  | LocA | AssoA        | ClsA | TETA | LocA | AssoA       | ClsA |
| SLAck trained on (original) TAO                | 24.6  | 47.1 | 24.1         | 2.5  | -    | -    | -           | -    |
| Ours trained on (original) TAO                 | 31.0  | 55.5 | 34.8 (+10.7) | 2.8  | 36.3 | 54.4 | 35.8        | 18.7 |
| SLAck trained on (original) TAO w pseudo label | 31.1  | 54.3 | 37.8         | 1.3  | 37.2 | 55.0 | 37.6        | 19.1 |
| Ours trained on (original) TAO w pseudo label  | 32.8  | 56.4 | 39.2         | 2.9  | 38.3 | 56.6 | 38.9        | 19.3 |
| Ours trained on C-TAO                          | 34.3  | 58.2 | 41.3 (+6.5)  | 3.5  | 39.6 | 57.3 | 42.0 (+6.2) | 19.6 |

Table 4. Comparison of OVTrack and our method under different ratios of data for training.

| Method  | Annotation ratio    | Novel |      |              |      | Base |      |             |      |
|---------|---------------------|-------|------|--------------|------|------|------|-------------|------|
|         |                     | TETA  | LocA | AssoA        | ClsA | TETA | LocA | AssoA       | ClsA |
| OVTrack | 3.7% (original TAO) | 25.4  | 48.8 | 25.6         | 1.9  | 33.5 | 49.7 | 30.9        | 19.8 |
|         | 3.7% × 2 (P-TAO)    | 29.5  | 50.2 | 36.5         | 1.7  | 36.5 | 51.7 | 38.2        | 19.6 |
|         | 100% (C-TAO)        | 30.5  | 50.2 | 39.3 (+13.7) | 2.1  | 37.4 | 52.7 | 39.3 (+8.4) | 20.1 |
| Ours    | 3.7% (original TAO) | 31.0  | 55.5 | 34.8         | 2.8  | 36.3 | 54.4 | 35.8        | 18.7 |
|         | 3.7% × 2 (P-TAO)    | 33.6  | 57.9 | 39.9         | 3.1  | 38.6 | 56.9 | 39.9        | 19.0 |
|         | 100% (C-TAO)        | 34.3  | 58.2 | 41.3         | 3.5  | 39.6 | 57.3 | 42.0        | 19.6 |

**In-depth insight of C-TAO.** The original TAO only annotates about 3.7% frames for training. As our experiments demonstrate, directly training on TAO fails to produce a viable OV tracker. To understand the incremental benefits, we conduct experiments with Pairwise-TAO (P-TAO), providing one continuous label after each TAO frame, resulting in  $3.7\% \times 2$  annotation ratio. The TAO→P-TAO→C-TAO progression validates our key assumption: OVMOT training *benefits significantly from continuously annotated video data*. However, an unexpected insight emerges: OVMOT training *does not heavily depend on fully-continuous annotations*. As shown in Table 4, even minimal continuous annotation (P-TAO) provides substantial improvements (13.7% in novel AssoA for OVTrack). This suggests that pairwise continuous annotation (low cost) can effectively help training, while fully-continuous annotations offer diminishing returns. This observation is insightful that significant performance gains are achievable with moderate annotation costs.

#### 5.4. Qualitative Analysis

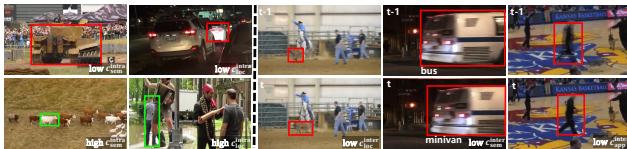


Figure 4. Visualization of intra- and inter-frame confidence.

**Intra-frame confidence analysis.** Fig. 4 (left) illustrates the visualization of intra-frame confidence. For semantic cues, low-confidence objects typically have ambiguous categories in complex backgrounds, while high-confidence objects are those with discriminative categories and simple backgrounds. Regarding location cues, low confidence is often associated with unclear positions. The last example in Fig. 4 (left) shows a high-confidence one, which has accurate locations while the appearance is disturbed. This way, the location cue can assist the tracking under this case.

**Inter-frame confidence analysis.** Fig. 4 (right) shows

several cases with low inter-frame confidence. Low  $c_{loc}^{inter}$  values typically occur with rapid object motion causing significant inter-frame position variations. Low  $c_{sem}^{inter}$  values appear when object categories are inconsistent across consecutive frames. Low  $c_{app}^{inter}$  values predominantly appear during sudden appearance changes, such as defocus or motion blur.

**Real-world visualization results.** Figure 5 presents challenging Internet videos with novel category objects experiencing dense scenarios, rapid deformation, and severe occlusion, demonstrating our algorithm’s robust tracking performance.

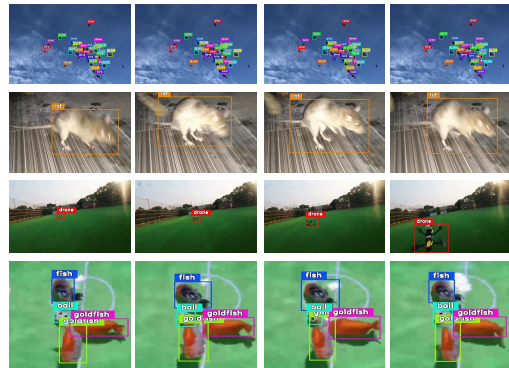


Figure 5. Visualization results of COVTrack.

#### 6. Conclusion

We have re-annotated the TAO training set to provide continuous labels and build C-TAO for OVMOT. We also propose a multi-cue collaborative framework for OVMOT. In this framework, we construct a novel intra-frame and inter-frame confidence estimation strategy to balance the weights among appearance, location, and semantic features for guiding multi-cue feature fusion. Extensive experimental results have verified the usefulness of continuous annotation (even a small amount) on TAO and the effectiveness of the multi-cue collaborative OVMOT framework. Through this work, we hope to provide new insights into using continuous video data and various cues for improving OVMOT.

## Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62402490, the Emerging Direction Cultivation Project of Interdisciplinary Center (Intelligent Protection and Utilization of Digital Cultural Heritage) from Tianjin University, the Guangdong Basic and Applied Basic Research Foundation under Grant 2025A1515010101, and the Hong Kong Research Grants Council under Grants 11219324 and 11219422.

## References

- [1] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. People-tracking-by-detection and people-detection-by-tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008. 2
- [2] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *IEEE/CVF International Conference on Computer Vision*, pages 941–951, 2019. 2, 7
- [3] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Uperoft. Simple online and realtime tracking. In *IEEE International Conference on Image Processing*, pages 3464–3468, 2016. 2
- [4] Jiarui Cai, Mingze Xu, Wei Li, Yuanjun Xiong, Wei Xia, Zhuowen Tu, and Stefano Soatto. MeMOT: Multi-object tracking with memory. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8090–8100, 2022. 2
- [5] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khrodkar, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9686–9696, 2023. 6, 7
- [6] Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. TAO: A large-scale benchmark for tracking any object. In *European Conference on Computer Vision*, pages 436–454, 2020. 1, 2
- [7] 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. 1
- [8] 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. 2
- [9] Yunhao Du, Zhicheng Zhao, Yang Song, Yanyun Zhao, Fei Su, Tao Gong, and Hongying Meng. StrongSORT: Make DeepSORT great again. *IEEE Transactions on Multimedia*, 25:8725–8737, 2023. 2
- [10] Wei Feng, Feifan Wang, Ruize Han, Yiyang Gan, Zekun Qian, Junhui Hou, and Song Wang. Unveiling the power of self-supervision for multi-view multi-human association and tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(1):351–368, 2025. 2
- [11] Tobias Fischer, Thomas E Huang, Jiangmiao Pang, Linlu Qiu, Haofeng Chen, Trevor Darrell, and Fisher Yu. QD-Track: Quasi-dense similarity learning for appearance-only multiple object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):15380–15393, 2023. 2, 7
- [12] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 1
- [13] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*, 2022. 4
- [14] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5356–5364, 2019. 1, 3, 6
- [15] Ruize Han, Wei Feng, Yujun Zhang, Jiewen Zhao, and Song Wang. Multiple human association and tracking from ego-centric and complementary top views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5225–5242, 2022. 1
- [16] Ruize Han, Wei Feng, Feifan Wang, Zekun Qian, Haomin Yan, and Song Wang. Benchmarking the complementary-view multi-human association and tracking. *International Journal of Computer Vision*, 132:118–136, 2023. 2
- [17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 6
- [18] Kuan-Chih Huang, Ming-Hsuan Yang, and Yi-Hsuan Tsai. Delving into motion-aware matching for monocular 3d object tracking. In *IEEE/CVF International Conference on Computer Vision*, pages 6909–6918, 2023. 2
- [19] Jan Krejčí, Oliver Kost, Ondřej Straka, and Jindřich Duník. Pedestrian tracking with monocular camera using unconstrained 3D motion model. *arXiv preprint arXiv:2403.11978*, 2024. 2
- [20] Laura Leal-Taixé, Cristian Canton-Ferrer, and Konrad Schindler. Learning by tracking: Siamese cnn for robust target association. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 33–40, 2016. 2
- [21] Siyuan Li, Martin Danelljan, Henghui Ding, Thomas E Huang, and Fisher Yu. Tracking every thing in the wild. In *European Conference on Computer Vision*, pages 498–515, 2022. 2, 6, 7
- [22] Siyuan Li, Tobias Fischer, Lei Ke, Henghui Ding, Martin Danelljan, and Fisher Yu. Ovtrack: Open-vocabulary multiple object tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5567–5577, 2023. 1, 2, 3, 4, 6, 7
- [23] Siyuan Li, Lei Ke, Martin Danelljan, Luigi Piccinelli, Mattia Segu, Luc Van Gool, and Fisher Yu. Matching anything by segmenting anything. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18963–18973, 2024. 2, 3, 6, 7

- [24] Siyuan Li, Lei Ke, Yung-Hsu Yang, Luigi Piccinelli, Mattia Segù, Martin Danelljan, and Luc Van Gool. Slack: Semantic, location, and appearance aware open-vocabulary tracking. In *European Conference on Computer Vision*, pages 1–18, 2024. 1, 2, 3, 4, 6, 7
- [25] Haiji Liang and Ruize Han. OVT-B: A new large-scale benchmark for open-vocabulary multi-object tracking. In *Advances in Neural Information Processing Systems*, pages 14849–14863, 2024. 3
- [26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pages 740–755, 2014. 1
- [27] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19045–19055, 2022. 2
- [28] Jonathon Luiten, Tobias Fischer, and Bastian Leibe. Track to reconstruct and reconstruct to track. *IEEE Robotics and Automation Letters*, 5(2):1803–1810, 2020. 2
- [29] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8844–8854, 2022. 2
- [30] Dennis Mitzel and Bastian Leibe. Taking mobile multi-object tracking to the next level: People, unknown objects, and carried items. In *European Conference on Computer Vision*, pages 566–579, 2012. 2
- [31] Aljoša Ošep, Wolfgang Mehner, Paul Voigtlaender, and Bastian Leibe. Track, then decide: Category-agnostic vision-based multi-object tracking. In *IEEE International Conference on Robotics and Automation*, pages 3494–3501, 2018. 2
- [32] Jiangmiao Pang, Linlu Qiu, Xia Li, Haofeng Chen, Qi Li, Trevor Darrell, and Fisher Yu. Quasi-dense similarity learning for multiple object tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 164–173, 2021. 2
- [33] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17939–17948, 2023. 2
- [34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021. 4
- [35] Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Tracking the untrackable: Learning to track multiple cues with long-term dependencies. In *IEEE/CVF International Conference on Computer Vision*, pages 300–311, 2017. 2
- [36] Fatemeh Saleh, Sadegh Aliakbarian, Hamid Reza Tofighi, Mathieu Salzmann, and Stephen Gould. Probabilistic tracking scoring and inpainting for multiple object tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14329–14339, 2021. 2
- [37] Mattia Segu, Luigi Piccinelli, Siyuan Li, Luc Van Gool, Fisher Yu, and Bernt Schiele. Walker: Self-supervised multiple object tracking by walking on temporal appearance graphs. In *European Conference on Computer Vision*, pages 1–18, 2024. 2
- [38] Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple object tracking with transformer. *arXiv preprint arXiv:2012.15460*, 2020. 2
- [39] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20993–21002, 2022. 1
- [40] Li Wang, Xinyu Zhang, Wenyuan Qin, Xiaoyu Li, Jinghan Gao, Lei Yang, Zhiwei Li, Jun Li, Lei Zhu, Hong Wang, and Huaping Liu. CAMO-MOT: Combined appearance-motion optimization for 3D multi-object tracking with camera-LiDAR fusion. *IEEE Transactions on Intelligent Transportation Systems*, 24(11):11981–11996, 2023. 2
- [41] Zhongdao Wang, Liang Zheng, Yixuan Liu, Yali Li, and Shengjin Wang. Towards real-time multi-object tracking. In *European Conference on Computer Vision*, pages 107–122. Springer, 2020. 2
- [42] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *IEEE International Conference on Image Processing*, pages 3645–3649, 2017. 2, 7
- [43] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *European Conference on Computer Vision*, pages 466–481, 2018. 2
- [44] Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. MOTR: End-to-end multiple-object tracking with transformer. In *European Conference on Computer Vision*, pages 659–675, 2022. 2
- [45] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. FairMOT: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, 129:3069–3087, 2021. 2
- [46] 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 *European Conference on Computer Vision*, pages 1–21, 2022. 2, 6, 7
- [47] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *European Conference on Computer Vision*, pages 474–490, 2020. 2
- [48] Xingyi Zhou, Tianwei Yin, Vladlen Koltun, and Philipp Krähenbühl. Global tracking transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8771–8780, 2022. 2