SLAck: Semantic, Location, and Appearance Aware Open-Vocabulary Tracking

Siyuan Li[1], Lei Ke[1], Yung-Hsu Yang[1], Luigi Piccinelli[1], Mattia Segù[1], Martin Danelljan[1], and Luc Van Gool[1,2]

[1] Computer Vision Lab, ETH Zürich
[2] INSAIT

Abstract. Open-vocabulary Multiple Object Tracking (MOT) aims to generalize trackers to novel categories not in the training set. Currently, the best-performing methods are mainly based on pure appearance matching. Due to the complexity of motion patterns in the large-vocabulary scenarios and unstable classification of the novel objects, the motion and semantics cues are either ignored or applied based on heuristics in the final matching steps by existing methods. In this paper, we present a unified framework SLAck that jointly considers semantics location, and appearance priors in the early steps of association and learns how to integrate all valuable information through a lightweight spatial and temporal object graph. Our method eliminates complex post-processing heuristics for fusing different cues and boosts the association performance significantly for large-scale open-vocabulary tracking. Without bells and whistles, we outperform previous state-of-the-art methods for novel classes tracking on the open-vocabulary MOT and TAO TETA benchmarks. Our code is available at github.com/siyuanliii/SLAck.

Keywords: Open-Vocabulary · Multiple Object Tracking

1 Introduction

Multiple Object Tracking (MOT) has traditionally been confined to a limited vocabulary, focusing on categories like pedestrians and vehicles [4, 8, 12, 26, 34]. The dawn of open-vocabulary tracking [19] expands the horizon to a wider array of classes but also amplifies the challenges, given the diverse appearances, behaviors, and motion patterns across different object classes. Despite the challenges, developing such tracking systems is of great significance. Real-world applications, ranging from autonomous driving to augmented reality, necessitate trackers that transcend these vocabulary constraints for broader, more versatile functionality.

Current best-performing MOT methods for large-vocabulary tracking are based on pure appearances-based matching [18–20]. Motion-based MOT faces significant challenges since the prevailing motion-based MOT methods lean heavily on the Kalman Filter (KF) [2,5,11,37,43]. KF-based trackers rely on the linear movement assumption, which is valid in pedestrian or vehicle-centric datasets like

(a) Pedestrian MOT	(b) Open-Vocabulary MOT
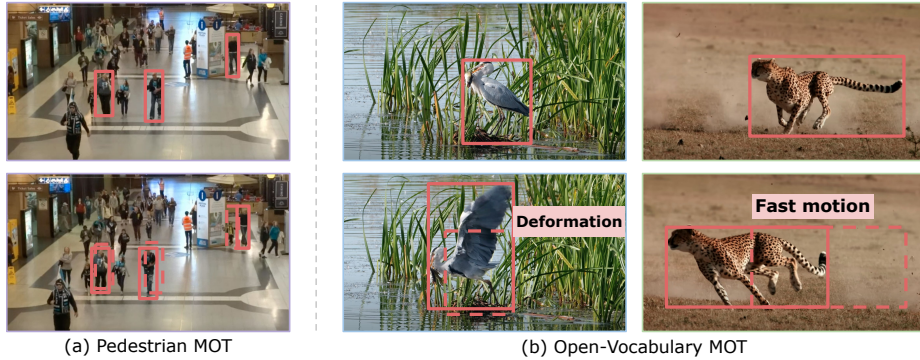
Fig. 1: Unlike conventionally pedestrian tracking such as MOT20 [8], open-world objects show high dynamicity on shape and motion, which poses significant challenges for motion-based trackers.
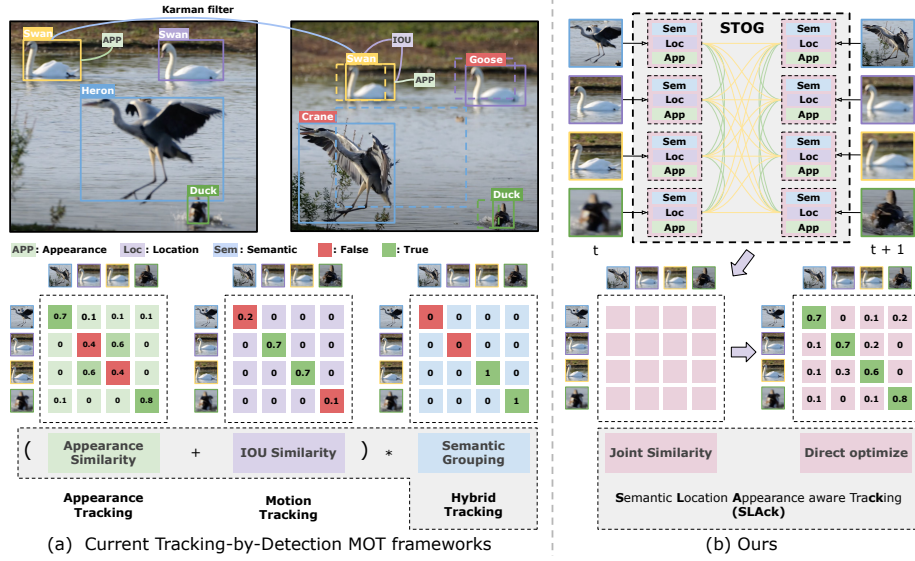
MOT Challenges [8,26]. Unfortunately, this assumption falters in complex open-vocabulary scenarios, characterized by non-linear object movements of varying object classes and motion patterns, as shown by Fig. 1.

Despite these challenges, motion and location prior can still provide important cues for robust MOT. Our approach diverges from explicit KF-based motion and, instead, proposes an implicit incorporation of location priors within the association learning process. By projecting location and shape information into the feature space and constructing an implicit spatial and temporal object graph through multi-head attention, our model learns implicit motion priors directly from data. This implicit modeling allows us to learn complex scene-level object structures spatially and enables us to capture both linear and non-linear motion temporally.

The synergy between semantics and motion cues is evident, as motion patterns often correlate with object categories. For instance, if the model learns the motion patterns of horses during training, it can directly use such knowledge to transfer to its semantic similar classes such as zebra during novel class tracking. This observation fuels our argument for a joint modeling approach, where semantic information enriches the location prior, leading to superior performance in novel class tracking, even without the reliance on appearance cues.

Appearance features are widely recognized as essential for accurate tracking [19, 27]. In our unified framework, these features are seamlessly integrated into the matching process by inclusion in the fused embedding. This integration strategy contrasts sharply with conventional methods that often resort to the heuristic-based fusion at later stages as shown in Fig. 2.

Our integrated model, named **SLAck**, encapsulates **S**emantics, **L**ocations, and **A**ppearances within a unified association framework for open-vocabulary tra**ck**ing. Utilizing pre-trained detectors, SLAck extracts a comprehensive set of descriptors and employs an attention-based spatial and temporal object graph (STOG) to facilitate information exchange, both spatially within frames and

APP: Appearance   **Loc :** Location   **Sem :** Semantic   ■ : False   ■ : True

(a) Current Tracking-by-Detection MOT frameworks

(b) Ours

Fig. 2: (a) Current MOT methods mainly use motion or appearance cues for tracking. Motion-based trackers typically assume linear motion, employing Kalman Filters, which do not account for the complex, non-linear movements observed in various classes, leading to tracking failures. Appearance-based tracking, which overlooks location information, may confuse targets with similar looks. Semantic cues are either ignored or used to group instances within the same class as a gating function, which usually leads to errors due to poor classification accuracy in the open world. Hybrid trackers combine all three cues—semantic, location, and appearance in a heuristic way which suffers the high dynamicity of the open world. (b) Our method jointly fuses the semantic, location, and appearance cues in the early matching stage and yields a single joint similarity matrix that can be directly optimized and trained end-to-end for association.

temporally across frames. Similar processes have been shown useful in modeling spatial object-level relations for object detection [16] or computing correspondence between points [30]. Our motivation is to exchange objects' semantics, appearance, and location, and exploit synergies among these aspects for better novel object tracking. This process not only enhances the awareness of relative object positions but also aligns motion and appearance patterns with semantic information, resulting in more robust temporal object associations.

Extensive experiments on large-vocabulary tracking benchmarks affirm the efficacy of our approach, with notable improvements in tracking performance in open-vocabulary settings, especially for novel objects. Our results indicate that motion patterns and semantics, when learned implicitly, can significantly enhance tracking performance. With the additional integration of appearance cues, our method achieves a substantial boost in association accuracy, confirming the value of our contributions: the early integration of semantic information with appearance and motion-based tracking, the learning of intrinsic cue fusion in-

stead of external heuristics, and the establishment of an implicit spatial-temporal object graph that demonstrates significant generalization in novel class tracking.

2   Related Work

Open-Vocabulary MOT The advent of the TAO dataset [7]—the first to offer a large vocabulary for tracking with over 800 classes—has spurred research into the open-world MOT challenges [18–20, 22, 45]. TETA [18] highlights the inadequacy of conventional MOT metrics in large-vocabulary scenarios and introduces a novel evaluation metric. TAO-OW [22] introduces a new setting assessing open-world tracking on TAO. This setting, however, focuses only on recall and is class-agnostic, failing to accurately measure tracker precision and semantic categorization. OVTrack [19] proposes an open-vocabulary MOT task that evaluates both precision and recall as well as classification. MASA [20] learns universal appearance models for MOT in the open world from vast unlabelled images. Current best-performing open-vocabulary MOT methods are most appearance-based and overlook semantic and location information. We posit that incorporating them can significantly improve generalization in novel class tracking.

MOT with Different Cues Motion cues are pivotal in MOT, with methods like SORT [2] and ByteTrack [43] utilizing the Kalman Filter (KF) for trajectory prediction. Extensions like AM-SORT [17] and OC-SORT [5] use transformers to adjust KF parameters during occlusions. However, KF-based methods struggle in open-vocabulary scenarios with rapid movement and deformation.

Appearance cues are commonly used in MOT. [24, 27, 33, 39] use discriminative instance appearance embeddings, while TETer [18] and OVTrack [19] leverage semantic appearances and Stable Diffusion [29] for association learning. These methods often require extensive annotations and can overfit to base classes, reducing performance on novel classes.

Hybrid approaches combine appearance and motion cues, often through heuristic-based fusion at later stages. DeepSORT [37] enhances SORT with an appearance model for occlusion handling. Other methods [31, 36, 43, 44] use dual branches for KF-based IoU and appearance embeddings, combining them heuristically. End-to-end transformer approaches [25, 32, 35, 42] treat instances as queries but face challenges with incomplete open-world video annotations. Offline methods like NeuralSolver [3] and SUSHI [6] model trajectories with future information but require comprehensive annotations and are challenging for real-world deployments, such as robotics.

3   Method

We first review popular MOT methods that utilize semantic, motion, appearance, and hybrid cues and investigate their failures when directly employed in open-vocabulary tracking. Then, we introduce our approach **SLAck** for addressing those issues.
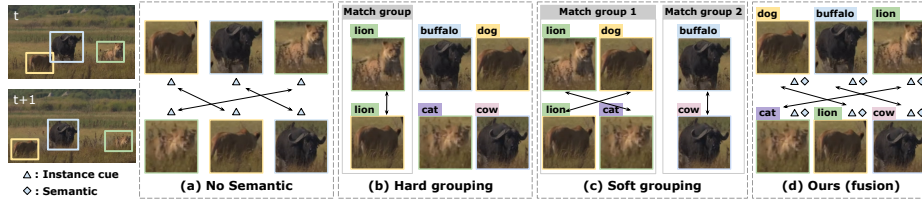
Fig. 3: Comparison of different strategies on utilizing semantic cues in open-vocabulary MOT. Instead of leveraging hard grouping or soft grouping [18], our method integrates semantic cues at the early stage to improve the association accuracy.

3.1   Preliminaries: Various cues for MOT

Semantic cues Fig. 3 summarizes different approaches to utilizing semantics in previous MOT literature. Semantic cues often play a minor role in multi-class MOT, typically utilized as a hard grouping where trackers associate objects of the same class based on detector predictions [27, 43]. This approach is effective for simple tasks like tracking humans and vehicles in datasets like KITTI [12] and nuScenes [4]. However, in open-vocabulary tracking, where classification is unreliable, this strategy falls short, as illustrated by Fig. 2. Relying on such uncertain classifications compromises tracking performance. TETer [18] proposes using contrastive class exemplar encodings for semantic comparison in feature space, shifting from hard to more dependable soft grouping. Yet, this method still relegates semantic information to a late-stage heuristic-based association. In contrast, we advocate for the early integration of semantic cues into the association process, harnessing their informative potential to enhance learning and association accuracy.

Motion cues Most motion-based MOT methods, including [2, 5, 11, 43], rely on the Kalman Filter (KF) under a linear-motion assumption. However, the dynamic nature of open environments, characterized by changing camera angles, rapid object movements, and diverse motion patterns across classes, challenges the efficacy of linear motion models. This complexity is evidenced both qualitatively, in Fig. 1, and quantitatively, through the poor performance of SORT-based trackers indicated in Table 4. Despite these challenges, object motion remains a valuable cue for tracking, given that spatial structures and proximities are generally consistent, even with non-linear motions.

   In response, we introduce an approach that leverages implicit motion modeling by establishing spatial and temporal relationships between objects. Specifically, we map the location and shape of each object into feature spaces, enabling intra-frame and inter-frame interactions through attention mechanisms. This process facilitates information exchange among objects about their locations, enhancing motion representation without relying on explicit linear assumptions.

Appearance cues Appearance-based methods, including [18–20], predominate in open-vocabulary tracking by utilizing appearance embeddings. These embeddings, derived from an added head to detectors and trained on static images or
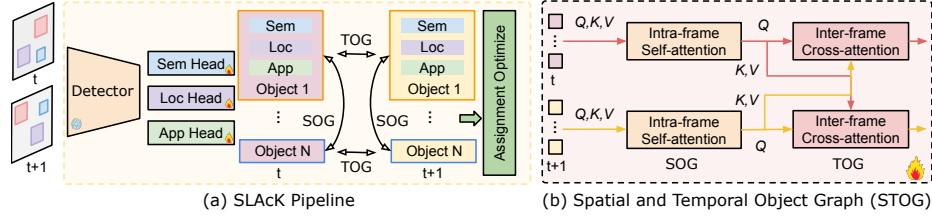
Fig. 4: Overall pipeline of SLAck. SLAck first extracts semantic, location, and appearance cues from pre-trained detectors, and then constructs the Spatial-Temporal Object Graph to fuse these cues and model the object dynamics for tracking.

video pairs via contrastive learning, are crucial for association in diverse tracking scenarios. Yet, reliance on appearance alone introduces challenges like occlusion sensitivity and demands extensive data for learning robust matches, often resulting in overfitting to base categories.

We integrate appearance with semantic information early in the feature-matching process, leveraging the high-level context from semantics while allowing the appearance head to focus on lower-level details which enables us to learn more generalized features across various conditions.

Hybrid cues Methods like JDE, FairMOT, DeepSORT, and ByteTrack [36, 37, 43, 44] utilize both appearance and motion cues, with motion modeled via the Kalman Filter and appearance derived from either a dedicated re-id network or jointly with detection features. Fusion of these cues occurs at the final matching step, where a spatial proximity matrix (from IoU) and an appearance similarity matrix (via dot product or cosine similarity) are combined through heuristics for Hungarian matching.

Our approach deviates by integrating all valuable information early on, culminating in a singular matching matrix. This early fusion sidesteps the heuristic complexity and enhances generalization, especially for novel classes.

3.2   Method Overview

Our method builds upon pre-trained open or large vocabulary detectors and extends them for tracking. We extract all information directly from the detectors such as semantics, location, and appearance. We then combine those cues with a spatial-temporal object graph to reason the association assignment. Our process is straightforward and end-to-end, without needing any extra heuristics to mix different cues. The model simply outputs an assignment matrix using the differential Sinkhorn-Knopp algorithm following [30]. Also, to deal with incomplete annotations in the TAO dataset, we directly use both predicted detection boxes and TAO's sparse ground truth as input for the association learning. We call this strategy Detection Aware Training. We show the overall pipeline in the Fig. 4.

3.3   Extract Semantic, Location, and Appearance Cues

We extract semantic, location, and appearance cues from pre-trained detectors. Recently, many strong object detectors [10, 13, 21, 41, 46] have emerged for universal object detection. To ensure fairness, we build our trackers on the same detectors used in [18, 19] for open and large vocabulary tracking. Importantly, we freeze all detector components during our association process to maintain the original strong open-vocabulary detection capabilities. We describe the semantic, location, and appearance heads in detail below.

Semantic Head For open-vocabulary tracking, we need to be able to configure the classes we are interested in without re-training. We use the same detector as OVTrack [19]. Directly using a CLIP [28] encoder for the semantic cues incurs high inference costs. Therefore, the semantic cues come from an adapted RCNN classification head which distillates the CLIP text encoder following [13]. Based on the output embedding of this CLIP-aligned classification head, we add a five-layer MLP to project the semantic feature to get the final semantic embedding E_{sem}. For the close-set setting, we use TETer's detectors [18] and use their CEM encoding as the input for our semantic head.

Location Head The location head takes the output of bounding boxes head from the detector and projects them into feature space. Then, the bounding box coordinates are normalized relative to the image dimensions to ensure scale invariance. Given a set of bounding boxes with coordinates $[x_{\mathrm{min}}, y_{\mathrm{min}}, x_{\mathrm{max}}, y_{\mathrm{max}}]$ for an image of dimensions $[H, W]$, the normalization process involves scaling and translating the coordinates relative to the image center and size. The image center, (C_x, C_y), is computed as $C_x = \frac{W}{2}$, $C_y = \frac{H}{2}$. The scaling factor is determined as 70% of the maximum image dimension, i.e., $scaling = 0.7 \times \max(H, W)$. Using this scaling factor, the normalized bounding box coordinates are computed as:

$$(x'_{\mathrm{min}}, y'_{\mathrm{min}}, w', h') = \left(\frac{x_{\mathrm{min}} - \frac{W}{2}}{s}, \frac{y_{\mathrm{min}} - \frac{H}{2}}{s}, \frac{x_{\mathrm{max}} - x_{\mathrm{min}}}{s}, \frac{y_{\mathrm{max}} - y_{\mathrm{min}}}{s} \right) \quad (1)$$

This normalization step ensures that the spatial location of objects within the image is preserved relative to the image dimensions, facilitating scale-invariant location features. We then feed the normalized coordinates into the location head to get a location embedding E_{loc}. For the close-set setting, we include corresponding confidences c with the box coordinates.

Appearance Head The appearance head takes the RoI features embeddings as input and outputs appearance embeddings that are optimized for the association. The appearance head is a simple four-layer convolution with one additional MLP. The RoI embeddings are transformed by the appearance head and get an appearance embedding E_{app}.

3.4   Spatial-Temporal Object Graph (STOG)

Spatial-Temporal Object Graph (STOG) accomplishes this by leveraging a combination of intra-frame self-attention and inter-frame cross-attention mechanisms

following [30] across different objects, encoding rich semantic, location, and appearance patterns that are essential for understanding object dynamics during open-world tracking.

Feature Fusion in STOG Before STOG, each object within a frame is represented by a set of distinct yet complementary features: semantic embeddings (E_{sem}), location (E_{loc}), and appearance (E_{app}). These embeddings capture various aspects of an object, such as its visual characteristics, spatial position, and class information, which are crucial for accurate tracking. The STOG model initiates its process by fusing the appearance, location, and semantic features of each object into a unified representation. The fusion is conducted through a simple yet effective summation operation:

$$E_{\text{fused}}^i = E_{\text{app}}^i + E_{\text{loc}}^i + E_{\text{sem}}^i, \tag{2}$$

where E_{fused}^i represents the fused embedding for the i-th object in the frame. This operation ensures that the information from all three modalities is integrated, allowing the STOG to process each object with a comprehensive understanding of its appearance, spatial context, and semantic meaning.

The fused embeddings serve as the input to the STOG model, where they undergo further processing through the intra-frame self-attention and inter-frame cross-attention mechanisms. The self-attention mechanism within each frame allows the model to refine its understanding of the objects based on their relative positions and visual similarities, while the cross-attention mechanism across frames captures the temporal evolution of objects, facilitating robust tracking even in complex dynamic scenes.

Intra-Frame Self-Attention The intra-frame self-attention mechanism independently processes objects within both the key and reference frames to analyze their spatial relationships and interactions. For objects in the keyframe (K) and reference frame (R), the self-attention (SA) operations are defined as follows:

$$\text{SA}_K(Q_K, K_K, V_K) = \sigma\left(\frac{Q_K K_K^T}{\sqrt{d}}\right) V_K, \text{SA}_R(Q_R, K_R, V_R) = \sigma\left(\frac{Q_R K_R^T}{\sqrt{d}}\right) V_R, \tag{3}$$

where σ denotes the Softmax operation, Q_K, K_K, V_K and Q_R, K_R, V_R denote the query, key, and value vectors for objects within the key and reference frames, respectively. d represents the dimensionality of the vectors. This step enhances the model's understanding of the complex intra-frame object arrangements and interactions by focusing on the most contextually relevant features of each object.

Inter-Frame Cross-Attention After processing intra-frame relationships, the model applies inter-frame cross-attention (CA) to both the key and reference frames independently, aiming to align and update object features across frames. This captures the temporal dependencies essential for tracking objects. The cross-attention operations for objects transitioning from the key to the reference frame and vice versa are formalized as:

$$\text{CA}_{K \to R}(Q_K, K_R, V_R) = \sigma\left(\frac{Q_K K_R^T}{\sqrt{d}}\right) V_R, \text{CA}_{R \to K}(Q_R, K_K, V_K) = \sigma\left(\frac{Q_R K_K^T}{\sqrt{d}}\right) V_K, \tag{4}$$

where Q_K and Q_R are the query vectors derived from objects in the key and reference frames, respectively, while K_R, V_R and K_K, V_K are the key and value vectors from the opposite frames. σ denotes the Softmax operation. This step is crucial for tracking objects over time, enabling the model to emphasize features that are most relevant for understanding the temporal evolution of each object. By first applying self-attention within each frame and then cross-attention across frames, the STOG model effectively captures both spatial and temporal object dynamics, enhancing its ability to track multiple objects across frames in a video.

3.5   Association Loss

Given a pair of frames, the objective is to compute a loss that encourages correct matching between objects across these frames. This process involves generating a target match matrix and then applying the Sinkhorn algorithm to compute a differentiable approximation of the optimal transport problem. The loss is calculated as follows:

Target Match Matrix Computation For each pair of key and reference frames, we construct a binary target match matrix, $\mathbf{T}$, to represent matches between objects across frames. Each entry in $\mathbf{T}$, denoted as T_{ij}, is set to 1 if the i-th object in the key frame matches the j-th object in the reference frame, and 0 otherwise. This method relies on ground truth match indices to determine the correspondence between objects. To accommodate objects that do not find a match in the counterpart frame, indicative of objects disappearing or newly appearing, we use a special 'dustbin' class. This adjustment expands the matrix to $\mathbf{T}'$ with dimensions $(M+1) \times (N+1)$, where M and N signify the count of objects in the key and reference frames, respectively.

Training Loss The whole open-vocabulary tracking framework can be trained in an end-to-end manner using the Sinkhorn algorithm, which provides a differentiable solution to the optimal transport problem. Given a score matrix $\mathbf{S}$ from the model, representing the predicted affinities between objects in the key and reference frames, and the augmented target match matrix $\mathbf{T}'$, the Sinkhorn loss $\mathcal{L}_{\text{Sinkhorn}}$ is computed as follows:

$$\mathcal{L}_{\text{Sinkhorn}} = -\sum_{i,j} T'_{ij} \log(S'_{ij}), \tag{5}$$

where S'_{ij} is the softmax-normalized score matrix after applying the Sinkhorn iterations, akin to a probabilistic matching. The final loss $\mathcal{L}$ for a batch of frames is the average of $\mathcal{L}_{\text{Sinkhorn}}$ over all frame pairs, encouraging the model to predict matches that align with the ground truth correspondences encoded in $\mathbf{T}'$.

3.6   Detection Aware Training (DAT)

Open-vocabulary tracking faces a significant challenge due to limited training data, with TAO [7] being the sole large-vocabulary MOT dataset, characterized by its incomplete annotations. This issue arises from the impracticality of

exhaustively annotating every object in video sequences, compounded by the expense of video annotation and the complexity of class definitions, which may vary in granularity. To address the incomplete annotation problem without compromising the detection capabilities of our pre-trained open-vocabulary detector, we freeze the detector's weights during training on TAO videos. This ensures the detector's performance remains unchanged. To adapt to the sparse annotations, we employ a strategy where the detector first infers bounding boxes on training videos, maintaining consistency in input data during both training and testing phases. We only compute the association loss when there is a matching between these predicted boxes and available ground truth, disregarding unmatched pairs. This method proves highly effective for end-to-end training in MOT tasks, allowing for significant performance gains by closely replicating test conditions during the training process.

4   Experiment

4.1   Evaluation Metrics

Following [19,38], we use **TETA** metrics [18] for the large-scale multiple-class multiple object tracking. TETA disentangles MOT evaluation into three sub-factors: Localization (LocA), Association (AssocA) and Classification (ClsA). Additionally, TETA supports evaluation with incomplete annotations, which gives more accurate tracking measurement for the open-vocabulary MOT [19].

4.2   Benchmarks

We evaluate our methods using the TAO dataset [7], the only large-scale, open-vocabulary MOT dataset. With over 800 categories, TAO stands as the most comprehensive MOT dataset regarding class diversity to date. It encompasses 500, 988, and 1,419 40-second annotated videos across train, validation, and test sets, respectively. TAO offers various benchmarks, highlighting distinct characteristics and requirements. We only use the TAO training set for training.
Open-Vocabulary MOT [19] This benchmark challenges trackers to generalize across novel classes without training on their annotations. TAO's adherence to the LVIS [14] taxonomy means the split between base and novel classes aligns with established open-vocabulary detection protocols [13], designating frequent and common classes as the base and rare ones as the novel. This setup reflects real-world conditions where trackers encounter seldom-seen categories, testing their adaptability and stability in tracking uncommon objects.
TAO TETA [18] TAO TETA benchmark is a closed-set MOT challenge, permitting training on all class annotations within TAO. It focuses on association quality, rewarding trackers that generate precise trajectories without overlaps.

4.3   Implementation Details

Our model, SLAck-OV, is constructed atop pre-trained open and large vocabulary detectors. For open-vocabulary MOT, we employ the same Faster R-CNN

Table 1: Effectiveness of semantic-aware matching. We examine the effectiveness of the semantic cues in novel class tracking. We focus on AssocA, which evaluates the association's performance. We observe consistent improvement when we add the semantic cues into the early association process.

Method	Cues			Novel			
TAO-val	Semantic	Location	Appearance	TETA	LocA	AssocA	ClsA
OC-SORT	-	✓	-	23.7	49.6	20.4	1.1
Lck (Motion-only)	-	✓	-	27.8	54.3	28.3	0.9
SLck (Semantic-aware Motion)	✓	✓	-	30.5 (+2.2)	55.1	35.4 (+7.1)	1.0
OVTrack	-	-	✓	27.8	48.3	33.6	1.5
Ack (Appearance-only)	-	-	✓	29.4	54.3	32.7	1.1
SAck (Semantic-aware Appearance)	✓	-	✓	29.7 (+0.3)	53.2	35.1 (+2.4)	0.9
LAck (Motion-Appearance)	-	✓	✓	30.8	54.6	36.4	1.4
SLAcK (Full model)	✓	✓	✓	**31.1** (+0.3)	55.0	**37.8** (+1.4)	1.3

detector as OVTrack [19], featuring a ResNet-50 [15] backbone trained on LVIS base classes. For the close-set TAO TETA benchmark, we utilize Faster R-CNN detectors with a class exemplar head, akin to TETer [18]. The variants SLAck-T and SLAck-L deploy Swin-Tiny and Swin-Large backbones [23], respectively. Training images are randomly resized, keeping the aspect ratio, with the shorter side between 640 and 800 pixels. Pairs of adjacent frames, within a maximum interval of 3 seconds, are selected for training. Further details are provided in the appendix.

4.4 Ablation Study

We provide comprehensive ablation experiments to validate the effectiveness of our proposed models and training strategy. Ablation tables report novel tracking performance on the open-vocabulary MOT benchmark [19] if not specified.

Effectiveness of semantic aware matching Previously, semantic cues are largely ignored by MOT methods. The association is done by either using motion or appearance cues. We validate the importance of semantic information for large-scale open-vocabulary tracking, especially in novel classes. We also add the pure motion-based OC-SORT [5] and pure appearance-based OVTrack [19] as baselines for easier comparison. Table 1 presents the results. Our model with location cues only (Lck) simulates the motion-based trackers whose association is done purely based on location cues. Adding semantics improves the association performance significantly by an increase of +7.1 on AssocA. It shows that modelling the semantics together with location or motion cues can effectively enhance the novel class tracking performance. With our semantic-aware motion tracker (SLck), we already outperform the appearance-based state-of-the-art method OVTrack [19] by +1.8 on AssocA. This further demonstrates our assumption that semantic and motion patterns are highly related. The motion patterns learned on base classes can effectively transfer to the unseen but semantically similar novel classes, to improve the generalization of trackers. Adding semantics cues into appearance-base matching (SAck) also increases the AssocA by +2.4 compared to the appearance-only method (Ack). Finally, adding the semantic cues can also

help the hybrid tracker (LAck) which includes both motion and appearance cues by further +1.4 improvement on AssocA. This demonstrates the importance of incorporating semantic cues into associations.

Effectiveness on DAT We evaluate the effectiveness of our proposed detection-aware training (DAT) strategy. Table 2 shows that training with DAT significantly improves the association performance by (+13.7 AssocA) compared to the training directly with the sparse and incomplete GroundTruth. DAT exploits the sparse supervision and, in the meantime, simulates the training with inference by using all predicted bounding boxes for training which significantly eliminates the gap between training and inference.

Comparison of different ways of utilizing semantic cues We benchmark against existing methods for utilizing semantic cues on the TAO TETA benchmark. We conduct these experiments based on TETer-T. Our investigation reveals that traditional hard grouping [27, 43], which associates objects sharing identical predicted class labels, underperforms significantly, showing a -4.6 decrease in AssocA. This drop is attributed to the instability of class predictions in large vocabularies, where hard classification constraints can hurt the overall tracking accuracy. Alternatively, soft grouping, which clusters semantically similar objects based on feature similarity, offers varied performance improvements depending on the features employed. Utilizing CEM encoding, a contrastive semantic feature trained on LVIS, soft grouping yields a +1.5 increase in AssocA over the no-semantic baseline. However, using BERT encoding hurt the performance. We add our semantic-aware Matching (Ours-SAck) to the TETer-T by integrating CEM encoding and appearance features directly within the STOG association process, eliminating late-stage grouping or gating mechanisms. This approach results in a substantial improvement, outperforming the no-semantic baseline by +2.8 and the CEM soft grouping by +1.3 in AssocA. This gain underscores the efficiency of our semantic-aware matching strategy in leveraging semantic information for large-vocabulary tracking.

Are semantic cues alone enough for tracking? As shown in Table 5, we also check whether the semantic cues alone contain sufficient appearance information that is distinguishable enough for tracking. Compared to the appearance alone (Ack-STOG), the association accuracy drops significantly -4.4 AssocA using semantic cues (Sck-STOG) alone. It suggests that although semantic cues contain instance appearance, however, they are not as distinguishable as the task-specific appearance features.

Effectiveness of STOG We assess the impact of spatial (SOG) and temporal (TOG) object graphs on trackers using different cues: semantic (Sck), location (Lck), appearance (Ack), and combination (SLAck). Table 5 compares trackers with and without SOG and TOG. Ind means without STOG. It uses simple MLP and addition to obtain the final tracking features. The results in Table 5 indicate TOG is important for semantic (+2.4 in AssocA) or location cues (+4.1 in AssocA). TOG provides insight into which objects are present in subsequent frames, thereby refining the features for better associations. For example, location-based trackers can anticipate potential object movements, while

Table 2: Training with incomplete annotated GT vs Ours DAT.

Method	TETA	LocA	AssocA	ClsA
SLAck with sparse GT	24.6	47.1	24.1	**2.5**
SLAck with DAT	**31.1**	**54.3**	**37.8** (+13.7)	1.3

Table 3: Comparison of different ways of integrating semantic cues for association. The comparison is done on the TAO TETA benchmark and we only focus on the AssocA. The study is based on TETer-T.

	No semantic	Hard Grouping (Class)	Soft Grouping (BERT Embed)	Soft Grouping (CEM Embed)	Semantic-aware Matching (Ours-SAck)
AssocA	35.2	30.6	31.2	36.7	**38.0**

Table 4: Open-vocabulary MOT comparison. We compare our method on open-vocabulary MOT benchmark [19]. We indicate the classes and data the methods trained on. All methods use ResNet50 as the backbone. [†] represents use the same detector.

Method	Classes		Novel				Base			
Validation set	Novel	Base	TETA	LocA	AssocA	ClsA	TETA	LocA	AssocA	ClsA
QDTrack [27]	✓	✓	22.5	42.7	24.4	0.4	27.1	45.6	24.7	11.0
TETer [18]	✓	✓	25.7	45.9	31.1	0.2	30.3	47.4	31.6	12.1
DeepSORT (ViLD) [37]	-	✓	21.1	46.4	14.7	**2.3**	26.9	47.1	15.8	17.7
Tracktor++ (ViLD) [1]	-	✓	22.7	46.7	19.3	2.2	28.3	47.4	20.5	17.0
ByteTrack[†] [43]	-	✓	22.0	48.2	16.6	1.0	28.2	50.4	18.1	16.0
OC-SORT[†] [5]	-	✓	23.7	49.6	20.4	1.1	28.9	51.4	19.8	15.4
MASA (R50)[†] [20]	-	-	30.0	54.2	34.6	1.0	36.9	55.1	36.4	19.3
OVTrack[†] [19]	-	✓	27.8	48.8	33.6	1.5	35.5	49.3	36.9	**20.2**
SLAck-OV[†]	-	✓	**31.1**	**54.3**	**37.8**	1.3	**37.2**	**55.0**	**37.6**	19.1
Test set	Novel	Base	TETA	LocA	AssocA	ClsA	TETA	LocA	AssocA	ClsA
QDTrack [27]	✓	✓	20.2	39.7	20.9	0.2	25.8	43.2	23.5	10.6
TETer [18]	✓	✓	21.7	39.1	25.9	0.0	29.2	44.0	30.4	10.7
DeepSORT (ViLD) [37]	-	✓	17.2	38.4	11.6	1.7	24.5	43.8	14.6	15.2
Tracktor++ (ViLD) [1]	-	✓	18.0	39.0	13.4	1.7	26.0	44.1	19.0	14.8
OVTrack[†] [19]	-	✓	24.1	41.8	28.7	1.8	32.6	45.6	**35.4**	**16.9**
SLAck-OV[†]	-	✓	**27.1**	**49.1**	**30.0**	**2.0**	**34.7**	**52.5**	35.6	16.1

semantic-based trackers can refine the semantic encodings based on temporal information. Appearance-based trackers benefit more from SOG (+0.9 in AssocA compared to TOG's +0.3). SOG facilitates awareness of co-located objects' appearances, encouraging feature differentiation to improve association accuracy. When all three cues are integrated with both SOG and TOG (STOG), we observe a substantial boost in association performance (+3.7 in AssocA). This underscores the synergistic effect of STOG with multiple tracking cues, significantly enhancing association accuracy.

4.5   Comparison to State-of-the-Art

Open-vocabulary MOT Table 4 presents our evaluation of the validation and test splits for open-vocabulary MOT [19]. To maintain comparability, we utilize a ResNet-50 backbone across all methods. QDTrack [27] and TETer [18], serving as closed-set baselines, train using all available classes. Competing methods enhance existing open-vocabulary detectors by incorporating off-the-shelf trackers like ByteTrack, OC-SORT, and MASA. OVTrack, previously the top performer in open-vocabulary MOT, leverages Stable Diffusion [29] to augment LVIS images for base class training. On val split, our model, SLAck-OV, significantly

Table 5: Ablation on the STOG. Ind: without STOG, SOG: Spatial Object Graph. TOG: Temporal Object Graph, Sck: using only Semantic, Lck: using only Location. SLAck: our full model.

Method	Spatial	Temporal	TETA	LocA	AssocA	ClsA
Sck - Ind	-	-	27.6	52.9	28.7	1.1
Sck - SOG	✓	-	27.5	53.3	28.1	1.0
Sck - TOG	-	✓	28.6	53.4	31.1	1.1
Sck - STOG	✓	✓	27.6	53.3	28.4	1.1
Lck - Ind	-	-	26.8	54.2	24.8	1.4
Lck - SOG	✓	-	27.3	53.9	26.6	1.5
Lck - TOG	-	✓	28.2	54.5	28.9	1.2
Lck - STOG	✓	✓	27.8	54.3	28.3	0.9
Ack - Ind	-	-	28.4	53.8	30.4	0.8
Ack - SOG	✓	-	28.8	54.2	31.3	0.8
Ack - TOG	-	✓	28.4	53.3	30.7	1.2
Ack - STOG	✓	✓	29.4	54.3	32.7	1.1
SLAck - Ind	-	-	29.6	53.9	34.1	0.8
SLAck - SOG	✓	-	30.1	53.9	35.5	0.8
SLAck - TOG	-	✓	29.7	53.5	34.7	0.9
SLAck - STOG	✓	✓	31.1	54.3	37.8	1.3

Table 6: Comparison on TAO TETA Benchmark [18]. Our models achieve better performance compared to SOTA methods using the same detectors. † indicates using the same object detector as TETer-T [18].

Method	TETA	LocA	AssocA	ClsA
SORT [2]	24.8	48.1	14.3	12.1
Tracktor [1]	24.2	47.4	13.0	12.1
DeepSORT [37]	26.0	48.4	17.5	12.1
Tracktor++ [1]	28.0	49.0	22.8	12.1
QDTrack [27]	30.0	50.5	27.4	12.1
OVTrack [19]	34.7	49.3	36.7	18.1
ByteTrack† [43]	27.6	48.3	20.2	14.4
OC-SORT† [5]	28.6	49.7	21.8	14.3
MASA(R50)† [20]	34.1	52.1	35.7	15.0
MASA(Detic)† [20]	34.7	51.9	36.4	15.8
MASA(GroundingDINO)† [20]	34.9	51.8	37.6	15.4
TETer-T† [18]	34.6	52.1	36.7	15.0
SLAck-T†	**35.5**	**52.2**	**38.9**	**15.6**
TETer-L [18]	40.1	56.3	39.9	24.1
SLAck-L	**41.1**	**56.3**	**41.8**	**25.1**
Train with additional tracking labels				
AOA [9]	25.3	23.4	30.6	21.9
UNINEXT(R50) [40]	31.9	43.3	35.5	17.1
GLEE-Plus(SwinL) [38]	41.5	52.9	40.9	30.8

surpasses all existing methods in tracking novel classes, particularly in association accuracy (AssocA), with a +4.2 improvement over OVTrack. This gap underscores OVTrack's limitation as a pure appearance-based tracker.

TAO TETA We also compare our methods on the close-set setting with all existing works in the established TAO TETA benchmark. As shown in Table 6, our SLAck-L outperforms all existing methods in AssocA, including methods trained with additional tracking annotations such as UNINEXT [40] and GLEE [38]. Using the same detectors as TETer-T and TETer-L [18], our models outperform each by +2.2 and +1.9 on the AssocA. Moreover, we outperform the second-best GLEE-Plus with the Swin-Large backbone by +0.9 in AssocA. Note that GLEE is a foundation vision model trained with 10 million images. This demonstrates the strong generalization and superior association accuracy of our methods.

5   Conclusion

We introduce SLAck, a unified open-vocabulary MOT approach that ingeniously integrates semantic and location cues with appearance features for object association. This framework, underpinned by a dynamic spatial-temporal object graph, adeptly captures the intricate relationships between objects, leveraging both their visual attributes and contextual information. By removing traditional late-stage fusion heuristics and in favor of early integration, the method not only simplifies the computational process but also enhances tracking accuracy, especially for novel object classes. The promising results on the large-vocabulary MOT benchmarks underscore the superior generalization capabilities of the proposed method.

References

1. Bergmann, P., Meinhardt, T., Leal-Taixe, L.: Tracking without bells and whistles. In: ICCV (2019)
2. Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: ICIP (2016)
3. Brasó, G., Leal-Taixé, L.: Learning a neural solver for multiple object tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6247–6257 (2020)
4. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020)
5. Cao, J., Pang, J., Weng, X., Khirodkar, R., Kitani, K.: Observation-centric sort: Rethinking sort for robust multi-object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9686–9696 (2023)
6. Cetintas, O., Brasó, G., Leal-Taixé, L.: Unifying short and long-term tracking with graph hierarchies. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22877–22887 (June 2023)
7. Dave, A., Khurana, T., Tokmakov, P., Schmid, C., Ramanan, D.: TAO: A large-scale benchmark for tracking any object. In: ECCV (2020)
8. Dendorfer, P., Rezatofighi, H., Milan, A., Shi, J., Cremers, D., Reid, I., Roth, S., Schindler, K., Leal-Taixé, L.: Mot20: A benchmark for multi object tracking in crowded scenes. arXiv preprint arXiv:2003.09003 (2020)
9. Du, F., Xu, B., Tang, J., Zhang, Y., Wang, F., Li, H.: 1st place solution to eccv-tao-2020: Detect and represent any object for tracking. arXiv preprint arXiv:2101.08040 (2021)
10. Du, Y., Wei, F., Zhang, Z., Shi, M., Gao, Y., Li, G.: 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. pp. 14084–14093 (2022)
11. Du, Y., Zhao, Z., Song, Y., Zhao, Y., Su, F., Gong, T., Meng, H.: Strongsort: Make deepsort great again. IEEE Transactions on Multimedia (2023)
12. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The International Journal of Robotics Research **32**(11), 1231–1237 (2013)
13. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. arXiv preprint arXiv:2104.13921 (2021)
14. Gupta, A., Dollar, P., Girshick, R.: LVIS: A dataset for large vocabulary instance segmentation. In: CVPR (2019)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
16. Hu, H., Gu, J., Zhang, Z., Dai, J., Wei, Y.: Relation networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3588–3597 (2018)
17. Kim, V., Jung, G., Lee, S.W.: Am-sort: Adaptable motion predictor with historical trajectory embedding for multi-object tracking. arXiv preprint arXiv:2401.13950 (2024)
18. Li, S., Danelljan, M., Ding, H., Huang, T.E., Yu, F.: Tracking every thing in the wild. In: ECCV. Springer (2022)

19. Li, S., Fischer, T., Ke, L., Ding, H., Danelljan, M., Yu, F.: Ovtrack: Open-vocabulary multiple object tracking. In: CVPR (2023)
20. Li, S., Ke, L., Danelljan, M., Piccinelli, L., Segu, M., Van Gool, L., Yu, F.: Matching anything by segmenting anything. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18963–18973 (2024)
21. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)
22. Liu, Y., Zulfikar, I.E., Luiten, J., Dave, A., Ramanan, D., Leibe, B., Ošep, A., Leal-Taixé, L.: Opening up open world tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19045–19055 (2022)
23. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
24. Liu, Z., Segu, M., Yu, F.: Cooler: Class-incremental learning for appearance-based multiple object tracking. In: DAGM German Conference on Pattern Recognition. pp. 443–458. Springer (2023)
25. Meinhardt, T., Kirillov, A., Leal-Taixe, L., Feichtenhofer, C.: Trackformer: Multi-object tracking with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8844–8854 (2022)
26. Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K.: MOT16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831 (2016)
27. Pang, J., Qiu, L., Li, X., Chen, H., Li, Q., Darrell, T., Yu, F.: Quasi-dense similarity learning for multiple object tracking. In: CVPR (2021)
28. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
29. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
30. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4938–4947 (2020)
31. Segu, M., Piccinelli, L., Li, S., Van Gool, L., Yu, F., Schiele, B.: Walker: Self-supervised multiple object tracking by walking on temporal appearance graphs. In: Computer Vision–ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings. Springer (2024)
32. Segu, M., Piccinelli, L., Li, S., Yang, Y.H., Schiele, B., Van Gool, L.: Samba: Synchronized set-of-sequences modeling for end-to-end multiple object tracking. arXiv preprint (2024)
33. Segu, M., Schiele, B., Yu, F.: Darth: holistic test-time adaptation for multiple object tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9717–9727 (2023)
34. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2446–2454 (2020)

35. Sun, P., Cao, J., Jiang, Y., Zhang, R., Xie, E., Yuan, Z., Wang, C., Luo, P.: Transtrack: Multiple object tracking with transformer. arXiv preprint arXiv:2012.15460 (2020)
36. Wang, Z., Zheng, L., Liu, Y., Wang, S.: Towards real-time multi-object tracking. The European Conference on Computer Vision (ECCV) (2020)
37. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: ICIP (2017)
38. Wu, J., Jiang, Y., Liu, Q., Yuan, Z., Bai, X., Bai, S.: General object foundation model for images and videos at scale. arXiv preprint arXiv:2312.09158 (2023)
39. Wu, J., Liu, Q., Jiang, Y., Bai, S., Yuille, A., Bai, X.: In defense of online models for video instance segmentation. ECCV (2022)
40. Yan, B., Jiang, Y., Wu, J., Wang, D., Yuan, Z., Luo, P., Lu, H.: Universal instance perception as object discovery and retrieval. In: CVPR (2023)
41. Ye, M., Ke, L., Li, S., Tai, Y.W., Tang, C.K., Danelljan, M., Yu, F.: Cascade-detr: delving into high-quality universal object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6704–6714 (2023)
42. Zeng, F., Dong, B., Zhang, Y., Wang, T., Zhang, X., Wei, Y.: Motr: End-to-end multiple-object tracking with transformer. In: European Conference on Computer Vision. pp. 659–675. Springer (2022)
43. Zhang, Y., Sun, P., Jiang, Y., Yu, D., Yuan, Z., Luo, P., Liu, W., Wang, X.: Bytetrack: Multi-object tracking by associating every detection box. In: ECCV (2022)
44. Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W.: FairMOT: On the fairness of detection and re-identification in multiple object tracking. IJCV (2021)
45. Zheng, G., Lin, S., Zuo, H., Fu, C., Pan, J.: Nettrack: Tracking highly dynamic objects with a net. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19145–19155 (2024)
46. Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., Misra, I.: Detecting twenty-thousand classes using image-level supervision. In: ECCV (2022)