Walker: Self-supervised Multiple Object Tracking by Walking on Temporal Appearance Graphs

Mattia Segu[1,2], Luigi Piccinelli[1], Siyuan Li[1],
Luc Van Gool[1,3], Fisher Yu[1], and Bernt Schiele[2]

[1] ETH Zurich, Switzerland
[2] Max Planck Institute for Informatics, Saarland Informatics Campus, Germany
[3] INSAIT, Bulgaria
https://github.com/mattiasegu/walker

Abstract. The supervision of state-of-the-art multiple object tracking (MOT) methods requires enormous annotation efforts to provide bounding boxes for all frames of all videos, and instance IDs to associate them through time. To this end, we introduce Walker, the first self-supervised tracker that learns from videos with sparse bounding box annotations, and no tracking labels. First, we design a quasi-dense temporal object appearance graph, and propose a novel multi-positive contrastive objective to optimize random walks on the graph and learn instance similarities. Then, we introduce an algorithm to enforce mutually-exclusive connective properties across instances in the graph, optimizing the learned topology for MOT. At inference time, we propose to associate detected instances to tracklets based on the max-likelihood transition state under motion-constrained bi-directional walks. Walker is the first self-supervised tracker to achieve competitive performance on MOT17, DanceTrack, and BDD100K. Remarkably, our proposal outperforms the previous self-supervised trackers even when drastically reducing the annotation requirements by up to 400x.

Keywords: Multiple Object Tracking · Self-supervised Learning

1 Introduction

Multiple object tracking (MOT) represents a cornerstone of modern perception systems for challenging computer vision applications, such as autonomous driving [13], video surveillance [12], and augmented reality [36]. Following the tracking-by-detection paradigm, multiple object trackers detect objects in all frames (object detection) while associating them through time (data association) to obtain tracklets. Modern trackers [1, 11, 46] achieve state-of-the-art performance by combining motion heuristics [4, 49, 56] with learned appearance descriptors [35, 49, 57] for data association. As such, the supervision of multiple object trackers requires annotating detection labels - *i.e.* bounding boxes - in every frame for all the objects of the categories of interest, and tracking labels

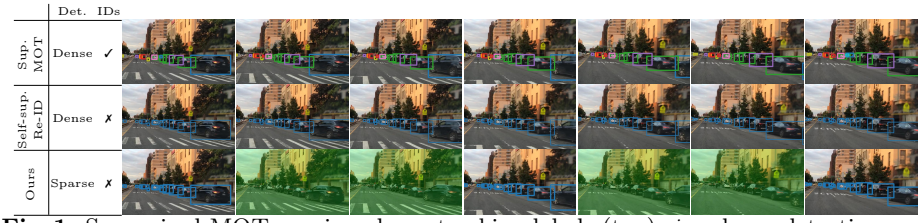

Fig. 1: Supervised MOT requires dense tracking labels (top), *i.e.* dense detection annotations at each frame and instance labels (shown by coloring boxes by instance ID) across frames. Self-supervised Re-ID assumes dense detection labels and no instance labels (middle). We explore self-supervised MOT in a more practical sparsely-annotated setting (bottom), with sparse detection annotations every k frames (here $k = 3$ for illustration purpose) and no instance labels. Fully-unlabeled frames in green.

as instance IDs to associate objects through time (Fig. 1, top). Thus, the annotation cost of MOT datasets [10, 41, 42, 44, 54] is linear in the number of frames, and labeling large video datasets can be prohibitive.

Self-supervised MOT - the problem of learning to track in the absence of the instance labels - represents an appealing solution to alleviate the enormous annotation cost. Nevertheless, the most common self-supervised MOT solutions [14, 19, 27, 39, 57] only rely on image-level self-supervision. By not leveraging the privileged temporal information of video streams, these approaches cannot learn appearance descriptors robust to view changes, and fail to close the gap with supervised MOT. Analogously, orthogonal research on self-supervised re-identification (Re-ID) [3, 15, 25, 51] traditionally assumes high-quality dense detection annotations in videos (Fig. 1, middle), hindering label-efficiency. We argue that video-level self-supervision should both enable discarding instance ID annotations and greatly sparsify the redundant detection labels (Fig. 1, bottom).

To this end, we introduce Walker, the first self-supervised multiple object tracker to learn from videos with sparse bounding box annotations and no tracking labels. Walker is a joint detection and tracking model composed of a detector and a cascaded embedding head. Inspired by [22], we design a temporal object appearance graph (TOAG) (Sec. 3.2) that connects object-level regions of interest (RoIs) on a pair of key/reference frames. During training, we propose to self-supervise appearance representations by walking on TOAGs. First, we introduce a novel multi-positive contrastive formulation to optimize cyclic random walks on the graph and learn instance similarities (Sec. 3.3). Then, we propose an algorithm to identify pseudo-matches between key and reference clusters of detections as the max-likelihood transition states over the cycle walks connecting them. Given such assignments, we enforce a mutually-exclusive graph connectivity across instances as required for MOT (Sec. 3.4). At inference time, we propose a more refined appearance similarity metric - namely the *biwalk* - to associate detections to tracklets by finding the max-likelihood transition state under the motion-constrained cycle walks connecting them (Sec. 3.6).

Moreover, we investigate the efficacy of self-supervised MOT by sparsifying the dense detection annotations requirement, *i.e.* providing ground-truth bounding boxes only every k frames in a video (Fig. 1, bottom). By relying on our video-level self-supervision, we find that Walker effectively leverages fully-unlabeled frames to learn superior appearance representations, significantly outperforming the frame-level self-supervised MOT state of the art [14] even when training with up to 400x less annotated frames (Fig. 4). Finally, experimental results on MOT17 [10], DanceTrack [42], and BDD100K [54] highlight that Walker is the first self-supervised tracker competitive with state-of-the-art supervised ones.

We summarize our contributions: (i) we introduce Walker, the first self-supervised multi-object tracker to learn appearance from sparsely annotated videos and no tracking labels; (ii) we propose a novel video-level self-supervision formulation that learns instance similarities with multi-positive and mutually-exclusive contrastive random walks on temporal object appearance graphs; (iii) Walker is the first self-supervised tracker competitive with state-of-the-art supervised MOT, while greatly reducing the annotation requirements.

2 Related Work

Multiple Object Tracking. Most MOT approaches rely on the tracking-by-detection paradigm, *i.e.* objects are detected in each frame while data association matches the detected instances across frames. *Motion-based* heuristics have long been used to associate objects through time [4, 37, 56]. SORT [4] first predicts the future location of the tracklets with a Kalman filter [23] and then matches predicted to detected boxes using Intersection over Union (IoU) as a measure of spatial similarity. ByteTrack [56] proposes a two-stage matching strategy to properly utilize low-score detections. However, motion-based trackers struggle under occlusions, low frame rates, and complex camera and objects motion [14]. DeepSORT [49], StrongSORT [11] and BoT-SORT [1] extend SORT with a standalone Re-ID module for occlusion-handling, and train it on an external pedestrian re-identification dataset [58] to extract *appearance-based* representations. However, their parallel Re-ID module undermines efficiency and is trained on external data. Recent *joint detection and tracking* models [14, 32, 35, 48, 57] extend the detector's feature extractor with an embedding head for efficient appearance extraction. QDTrack's [14, 35] quasi-dense contrastive formulation proved an effective in-domain appearance-learning scheme [14]. Queries in query-based trackers [34,38,43,55] are also implicit appearance representations. While appearance complements motion-based trackers, it comes with a high annotation cost. Training appearance extractors in-domain necessitates tracking datasets to provide detection and instance ID annotations for all frames in a video (Fig. 1, top). Our work overcomes these limitations by proposing a self-supervised appearance-learning algorithm that eliminates the need for instance-association labels, and allows for sparser detection annotations (Fig. 1, bottom).

Self-supervised Re-ID. Self-supervised Re-ID [3, 15, 25, 51] is the problem of learning instance representations given ground-truth detections (Fig. 1, middle).

[8, 21, 25] learn Re-ID with image-level self-supervision via pre-text tasks - *e.g.* image rotation, puzzle solving, reconstruction, MoCo-v2 [7], BYOL [16]. Other techniques learn Re-ID directly on in-domain videos by means of weak clustering labels obtained with tracking algorithms [20, 24, 51], or cycle consistency [3, 15] on ground-truth bounding boxes. By assuming availability of ground-truth detections, such approaches are not designed for joint detection and tracking.

Self-supervised Multiple Object Tracking. Despite the recent advances in self-supervised correspondence learning in videos [17, 22, 45], frame-level self-supervision is the standard in MOT. QDTrack-S(tatic) [14] generates two views of the same frame with data augmentation and optimizes a contrastive loss on the embeddings of different instances. Due to its simplicity, this paradigm has been adopted in test-time adaptive [39], open-vocabulary [27,29,53] and foundational tracking [28]. However, MOT requires associating instances through time, and data augmentation cannot mimic the occlusions, pose changes, and distortions of real videos. By walking on temporal appearance graphs, our method benefits from the video information to learn superior appearance representations.

3 Walker

We introduce our novel self-supervised tracker, Walker. We report architectural details in Sec. 3.1, and define our proposed quasi-dense temporal object appearance graph (Sec. 3.2). We then introduce our techniques to train the TOAG and learn instance descriptors from unlabeled videos: a novel multi-positive contrastive objective to optimize random walks on the appearance graph - after which Walker is named - (Sec. 3.3); our approach to identify pseudo-assignments and optimize mutually-exclusive connectivity on the graph (Sec. 3.4). Finally, we detail Walker's data association scheme and introduce our biwalk similarity metric (Sec. 3.6) to track objects based on the learned appearance graph.

3.1 Architecture

Our tracker can be coupled with any two-stage and one-stage detector for end-to-end training. The object detector is composed of a feature extractor with a Feature Pyramid Network (FPN) to extract multi-scale feature maps and a bounding box head. An additional embedding head extracts deeper appearance representations for each RoI after RoIAlign [18]. For two-stage detectors, we treat the region proposals as RoIs; for one-stage detectors, the detections after non maximum suppression (NMS). Following state-of-the-art appearance- [14] and motion-based [56] trackers, we choose YOLOX as *detector*, while our *embedding head* is a 4conv-1fc head with group normalization [50] to extract 256-dimensional features as in QDTrack [14].

3.2 Temporal Object Appearance Graphs

We introduce a self-supervised formulation to learn instance similarities by walking on quasi-dense temporal object appearance graphs (TOAGs). Inspired by the

contrastive random walk for self-supervised pixel-level correspondences [22], we represent each video as a quasi-dense [14] directed appearance graph $\mathcal{G}$ where nodes are the quasi-dense RoIs, and weighted edges connect nodes in neighboring frames. Unlike [22], our work redefines the appearance graph to walk on quasi-dense object regions, introduces a new multi-positive self-supervised objective (Sec. 3.3), and enforces mutually-exclusive connective properties across instances (Sec. 3.4) to make the learned topology optimal for MOT.

Nodes Definition. We define the graph nodes for an image I_t at time t as its RoIs, and describe them by their appearance embeddings. Given the set of high-confidence detections $\mathcal{D}_t^{\text{high}} = \{d_t^i \mid \text{conf}(d_t^i) \geq \beta_{\text{obj}} = 0.3\}$ predicted by the detector on I_t, or the set of ground-truth boxes $\hat{\mathcal{D}}_t = \{\hat{d}_t^i\}$, we define a RoI as positive to a detection d_t^i if their IoU is higher than $\alpha_1 = 0.7$, negative if lower than $\alpha_2 = 0.3$. We use RoI Align [18] to pool feature maps at different levels in the FPN [30] according to the RoI scales. For each frame I_t, we select 128 positive RoIs $\mathbf{Q}_t^+$ and 128 negative $\mathbf{Q}_t^-$ ones, and describe the nodes $\mathbf{Q}_t = \mathbf{Q}_t^+ \cup \mathbf{Q}_t^-$ by the corresponding embeddings matrix $Q_t = [Q_t^+, Q_t^-]$ obtained by applying the embedding head on the pooled RoI features. In contrast to [22], our nodes are object-centric RoIs instead of patches to learn instance-specific representations.

Cluster Definition. Given the quasi-dense nature of our TOAG, multiple nodes can represent different views of the same object. We define the cluster $\mathcal{C}_t^i = \mathcal{C}_t(\mathbf{q}_t^i) = \{\mathbf{q}_t^j \in \mathbf{Q}_t \mid \text{IoU}(\mathbf{q}_t^j, \mathbf{q}_t^i) \geq \alpha_1 = 0.7\}$ as the set of nodes sufficiently overlapping with the i-th node $\mathbf{q}_t^i$ in I_t. Given the high overlap, all RoIs in a cluster $\mathcal{C}_t(\mathbf{q}_t^i)$ typically represent the same instance, *i.e.* a specific pedestrian.

Edges Definition. We define the edges $A_t^{t'}(i,j)$ connecting the nodes $\mathbf{q}_t^i$ and $\mathbf{q}_{t'}^j$ across I_t and $I_{t'}$ by the cosine similarities $c(q_t^i, q_{t'}^j) = (q_t^i \cdot q_{t'}^j)/(||q_t^i||||q_{t'}^j||)$ between the nodes' embeddings q_t^i and $q_{t'}^j$, transformed into non-negative affinities by a softmax with temperature τ over edges departing from each node $\mathbf{q}_t^i$ directed to all nodes $\mathbf{q}_{t'}^i \in \mathbf{Q}_{t'}$. $A_t^{t'}$ is the local transition matrix from $\mathbf{Q}_t$ to $\mathbf{Q}_{t'}$ on $\mathcal{G}$:

$$A_t^{t'}(i,j) = \texttt{softmax}_i(Q_t Q_{t'}^\top)(i,j) = \frac{exp(c(q_t^i, q_{t'}^j)/\tau)}{\sum_{l=1}^{N} exp(c(q_t^i, q_{t'}^l)/\tau)}, \tag{1}$$

Unlike [22] and since our edges represent the instance similarities used for tracking, the optimal topology of $\mathcal{G}$ for MOT must present mutually-exclusive connective properties across clusters of nodes - *i.e.* nodes from one instance can only transition to other nodes of the same instance - which we enforce in Sec. 3.4.

Temporal Appearance Graph Definition. An appearance graph $\mathcal{G}$ defined by the nodes and edges described above is a spatio-temporal Markov chain whose transition probabilities between its quasi-dense states are given by the non-negative affinity matrix $A_t^{t'}(i,j) = P(X_{t'} = j | X_t = i) = p_{X_{t'}|X_t}(j|i)$, where X_t is the state of a walker at time t and $P(X_t = i)$ is the probability of being at node i at time t. In Secs. 3.3 to 3.5 we show how to learn a mutually-exclusive TOAG, and in Sec. 3.6 how to use it for tracking.

Fig. 2: Multi-positive Cycle Consistency. Illustration of the proposed multi-positive cycle consistency on quasi-dense TOAGs (Sec. 3.3). We show the cycle walk departing from a given query node (yellow). The multiple positive (negative) nodes are in green (red). For ease of visualization, we only show the high-likelihood transitions.

3.3 Learning Instance Representations by Walking on Cyclic Object Appearance Graphs

In absence of instance ID labels, we propose to self-supervise instance similarities (edges) by optimizing multi-positive contrastive random walks on cyclic TOAGs. **Cycle Walk Definition.** Given a key image I_t and its bounding box annotations, we randomly sample an unlabeled reference image I_{t+k} from its temporal neighborhood, *i.e.* $k \in [-\hat{k}, \hat{k}]$, with $\hat{k}$ dataset-dependent. We build a cyclic appearance graph $\mathcal{G}$ (Fig. 2) as a walk from the positive nodes $\mathbf{Q}_t^+$ - likely to represent objects - in the key image I_t to all the nodes $\mathbf{Q}_{t+k}$ in the reference image I_{t+k} and back to all nodes $\mathbf{Q}_t = [\mathbf{Q}_t^+, \mathbf{Q}_t^-]$ in I_t. The resulting walk $\mathcal{G} : \mathbf{Q}_t^+ \rightarrow \mathbf{Q}_{t+k} \rightarrow \mathbf{Q}_t$ is a Markov chain described by the forward and backward transitions $A_{t^+}^{t+k}$ and A_{t+k}^{t}, whose chained transition $\bar{A}_{t^+}^{t}$ describes the cycle correspondence as a multi-step walk along the object appearance graph $\mathcal{G}$:

$$\bar{A}_{t^+}^{t} = A_{t^+}^{t+k} A_{t+k}^{t} = P_{\mathcal{G}}(X_t|X_{t+k}) P_{\mathcal{G}}(X_{t+k}|X_t^+) = P_{\mathcal{G}}(X_t|X_t^+). \tag{2}$$

Multi-positive Cycle Consistency. Cycle consistency is satisfied for a node $\mathbf{q}_t^i$ in I_t if $p^{\mathcal{G}}_{X_t|X_t^+}(i|i) > p^{\mathcal{G}}_{X_t|X_t^+}(j|i)\ \forall\, j \neq i$, *i.e.* a cycle walk on $\mathcal{G}$ starting from $\mathbf{q}_t^i$ ends on $\mathbf{q}_t^i$ itself. However, since the above-defined graph is quasi-dense, we can identify multiple positive targets Y_i^+ for the walk starting from $\mathbf{q}_t^i$ as the cluster $\mathcal{C}_t(\mathbf{q}_t^i)$ of nodes $\mathbf{q}_t^l$ sufficiently overlapping with the starting node $\mathbf{q}_t^i$, *i.e.*$Y_i^+ = \mathcal{C}_t(\mathbf{q}_t^i) = \{\mathbf{q}_t^j \in \mathbf{Q}_t \mid \text{IoU}(\mathbf{q}_t^j, \mathbf{q}_t^i) \geq \alpha_1 = 0.7\}$. All other nodes are considered negative targets to $\mathbf{q}_t^i$, *i.e.*$Y_i^- = \{\mathbf{q}_t^j \mid \mathbf{q}_t^i \notin Y_i^+\ \forall\, \mathbf{q}_t^j \in \mathbf{Q}_t\}$. Fig. 2 illustrates the positive (green) and negative (red) targets for a cycle walk starting from a query node (yellow). We consider *multi-positive cycle consistency* satisfied if:

$$p^{\mathcal{G}}_{X_t|X_t^+}(Y_i^+|i) = \sum_{\mathbf{q}_t^l \in Y_i^+} p^{\mathcal{G}}_{X_t|X_t^+}(l|i) > p^{\mathcal{G}}_{X_t|X_t^+}(j|i)\ \forall\, \mathbf{q}_t^j \notin Y_i^+. \tag{3}$$

Meaningful pairwise instance similarities must emerge to solve the cyclic walk on the graph, such that each node walks back to one of its multiple positive targets when a latent correspondence is found in I_{t+k}. In MOT, a desired latent

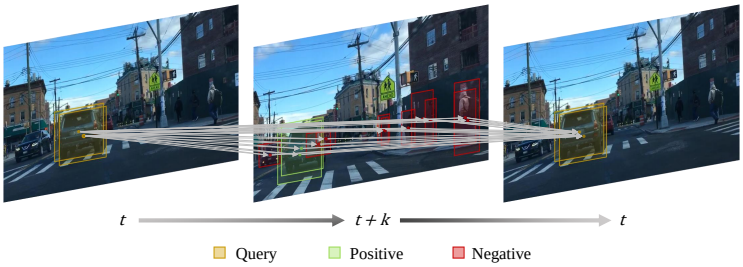

Fig. 3: Cluster-wise Forward Assignment. Illustration of the positive (green) and negative (red) forward pseudo-labels for an input query cluster (yellow), deriving from our cluster-wise forward assignment strategy described in Sec. 3.4.

correspondence in I_{t+k} to a RoI in I_t is a RoI representing the same instance. We introduce a novel *multi-positive contrastive loss* on the cycle probabilities to solve the quasi-dense cycle consistency problem and let latent matches emerge for all starting object nodes $\mathbf{q}_t^i \in \mathbf{Q}_t^+$, with $\bar{A}_{t^+}^t(i,j) = p^{\mathcal{G}}_{X_t|X_t^+}(j|i)$ probability of closing in $\mathbf{q}_t^j$ a cycle on $\mathcal{G}$ that starts from $\mathbf{q}_t^i$:

$$\mathcal{L}_{\text{cycle}} = \sum_{\mathbf{q}_t^i \in \mathbf{Q}_t^+} \log(1 + \sum_{\mathbf{q}_t^l \in Y_i^+} \sum_{\mathbf{q}_t^j \in Y_i^-} \exp(\bar{A}_{t^+}^t(i,j) - \bar{A}_{t^+}^t(i,l))). \tag{4}$$

3.4 Enforcing Mutually-exclusive Assignments

For a given starting node $\mathbf{q}_t^i$ in I_t, enforcing our multi-positive cycle consistency allows the emergence of multiple latent correspondences in the reference frame I_{t+k}, *i.e.* multiple nodes $\mathbf{q}_{t+k}^j$ with high transition probability $p^{\mathcal{G}}_{X_{t+k}|X_t,X_t^+}(j|Y_i^+,i)$ on the cycle walk $\mathcal{G}_i$. However, it is not guaranteed that all such correspondences belong to the same instance. In MOT, where the optimal graph topology must exhibit mutually-exclusive connective properties, having multiple instances in I_{t+k} linked to the same instance in I_t is undesirable. To this end, we propose to (i) identify cluster-wise forward assignments on our cyclic appearance graph (Fig. 3), and (ii) optimize the corresponding transition probabilities to satisfy mutually-exclusive connectivity. Pseudo-code is in the Appendix.

Cluster-wise Forward Assignment. In Sec. A (Appendix), we prove that the probability of transitioning on a latent node $\mathbf{q}_{t+k}^j$ on the reference image I_{t+k} when starting from $\mathbf{q}_{t^+}^i$ in I_t and ending on $\mathbf{q}_t^l$ in I_t along the cycle walk $\mathcal{G}$ is:

$$p^{\mathcal{G}}_{X_{t+k}|X_t,X_t^+}(j|l,i) = p^{\mathcal{G}}_{X_t|X_{t+k}}(l|j)p^{\mathcal{G}}X_{t+k}, X_t^+(j|i)/C \tag{5}$$

$$= A_{t^+}^{t+k}(i,j)A_{t+k}^t(j,l)/C \tag{6}$$

where $C = \sum_{\mathbf{q}_{t+k}^m \in \mathbf{q}_{t+k}} p^{\mathcal{G}}_{X_t|X_{t+k}}(l|m)p^{\mathcal{G}}_{X_{t+k}|X_t^+}(m|i)$.

In our quasi-dense setting (Fig. 3), the cluster of nodes $\mathcal{C}_t^i$ around $\mathbf{q}_{t^+}^i$ in I_t shares the set of multiple targets $Y_i^+ = \mathcal{C}_t^i$ with cardinality $||Y_i^+||$ in I_t for the cycle walk $\mathcal{G}_i$. For a node $\mathbf{q}_{t^+}^i$ in I_t, we can thus refine the probability estimate of traversing a reference node by averaging over all cycles starting from $\mathcal{C}_t^i$ and ending on Y_i^+. Thus, we identify the max-likelihood transition state z_{t+k}^i on I_{t+k} for a cycle walk $\mathcal{G}_i$ starting from $\mathbf{q}_{t^+}^i$ in I_t:

$$p^{\mathcal{G}}_{X_{t+k}|X_t,X_t^+}(j|Y_i^+,Y_i^+) = \sum_{i,l} \frac{A_{t^+}^{t+k}(i,j)A_{t+k}^{t}(j,l)}{C||Y_i^+||} \tag{7}$$

$$z_{t+k}^i = \underset{\mathbf{q}_{t+k}^j \in \mathbf{q}_{t+k}}{\operatorname{argmax}} \; p^{\mathcal{G}}_{X_{t+k}|X_t,X_t^+}(j|Y_i^+,Y_i^+) \tag{8}$$

where $\mathbf{q}_{t^+}^i \in Y_i^+$ and $\mathbf{q}_t^l \in Y_i^+$. We identify $\mathcal{Z}_{t+k}^i = \mathcal{C}_{t+k}(z_{t+k}^i)$ as the cluster of RoIs on I_{t+k} matching to the cluster $\mathcal{C}_{t^+}(\mathbf{q}_{t^+}^i)$ of RoIs on I_t (Fig. 3).

Optimizing Mutually-exclusive Assignments. Given the set of positive nodes $\mathbf{Q}_t^+$ in I_t, we propose to enforce the desired mutually-exclusive connectivity property on $\mathcal{G}$ - *i.e.* one cluster $\mathcal{Z}_{t+k}^i$ in I_{t+k} is assigned to at most one $\mathcal{C}_t^i$ on I_t - by incrementally assigning the clusters $\mathcal{C}_t^i \; \forall \mathbf{q}_{t^+}^i \in \mathbf{Q}_t^+$ to previously unassigned pseudo-matches $\mathcal{Z}_{t+k}^i$ in I_{t+k}, and optimizing the corresponding transition probabilities. In particular, (i) we sort the unique clusters $\mathcal{C}_t^i$ by their cycle closure probability $p^{\mathcal{G}}_{X_t|X_t^+}(Y_i^+|\mathcal{C}_t^i) = \frac{1}{||Y_i^+||}\sum_{\mathbf{q}_{t^+}^m \in Y_i^+}\sum_{\mathbf{q}_t^i \in Y_i^+} p^{\mathcal{G}}_{X_t|X_t^+}(l|m)$; (ii) since low cycle closure probability means that a latent correspondence cannot be found, we filter out clusters with cycle closure probability less than a threshold β_{cycle}, *i.e.*$\mathcal{C}_t^{\text{valid}} = \{\mathcal{C}_t^i \mid p^{\mathcal{G}}_{X_t|X_t^+}(Y_i^+|\mathcal{C}_t^i) \geq \beta_{\text{cycle}} = 0.8; \; \forall \mathbf{q}_{t^+}^i \in \mathbf{Q}_t^+\}$; (iii) for each valid cluster $\mathcal{C}_t^i \in \mathcal{C}_t^{\text{valid}}$ in I_t we find a matching cluster $\mathcal{Z}_{t+k}^i \notin \mathcal{Z}_{t+k}^{\text{assigned}}$ in I_{t+k} that was not previously matched to another cluster, where $\mathcal{Z}_{t+k}^{\text{assigned}}$ is the set of already-assigned latent clusters; (iv) we optimize the forward transition probabilities $A_{t^+}^{t+k}$ using an L_2 loss, whose positive targets for nodes in a cluster $\mathcal{C}_t^l \in \mathcal{C}_t^{\text{valid}}$ are $\mathcal{Z}_{t+k}^i$, and all other nodes are negative targets:

$$\mathcal{L}_{\text{forward}} = \sum_{l,i,j}(p^{\mathcal{G}}_{X_{t+k}|X_t^+}(j|i) - I[\mathbf{q}_{t+k}^j \in \mathcal{Z}_{t+k}^i])^2 = \tag{9}$$

$$= \sum_{l,i,j}(A_{t^+}^{t+k}(i,j) - I[\mathbf{q}_{t+k}^j \in \mathcal{Z}_{t+k}^i])^2, \tag{10}$$

where $\{l,i,j|\mathcal{C}_{t^+}^l \in \mathcal{C}_{t^+}^{\text{valid}}, \mathbf{q}_{t^+}^i \in \mathcal{C}_{t^+}^l, \mathbf{q}_{t+k}^j \in \mathbf{Q}_{t+k}\}$ and $I[\cdot]$ is the indicator function. We sample three times more negative pairs than positive ones to balance the loss.

3.5 Total Loss

We optimize the entire network under $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{det}} + \gamma_1\mathcal{L}_{\text{cycle}} + \gamma_2\mathcal{L}_{\text{forward}}$. $\mathcal{L}_{\text{det}}$ is the loss for the chosen object detector on the key frame I_t, and $\gamma_1 = 1.0$, $\gamma_2 = 2.0$.

3.6 Tracking with Walker

We here detail Walker's inference-time data association pipeline used for tracking with a TOAG trained as in Secs. 3.3 to 3.5.

Biwalk Similarity. Inspired by the properties of our cyclic (bi-directional) walk on temporal object appearance graphs, we propose the *biwalk*, a novel appearance similarity metric. Let N be the set of detected objects in frame I_t with appearance embeddings $\mathbf{n}$, and M the matching candidates from the past K frames with appearance embeddings $\mathbf{m}$. We define $\mathcal{G} : N \rightarrow M \rightarrow N$ as the cycle transition walk from the detections to the matching candidates and back to the detections. $\mathcal{G}$ is described by the cycle transition matrix $\bar{A}_N^N = A_N^M A_M^N$, with A_N^M and A_M^N forward and backward transition matrices respectively. We then propose to measure the similarity between a detection N_i and a matching candidate M_j as the probability of traversing the corresponding node over a satisfied cycle transition $\mathcal{G}_i : N_i \rightarrow M \rightarrow N_i$. Analogously to Sec. 3.4, the biwalk similarity can be used to determine the most-plausible match in M as the max-likelihood transition state on the cyclic graph $\mathcal{G}_i$. We thus define the biwalk similarity $s_{i,j}^{\text{biwalk}}$ between a detection i and a matching candidate j as:

$$s_{i,j}^{\text{biwalk}} = p_{M|N,N}^{\mathcal{G}_i}(j|i,i) \cdot \text{I}[p_{N|N}^{\mathcal{G}_i}(i|i) \geq \beta_{\text{cycle}}] = \tag{11}$$

$$= A_N^M(i,j) A_M^N(j,i)/C \cdot \text{I}[\bar{A}_N^N(i,i) \geq \beta_{\text{cycle}}], \tag{12}$$

where $p_{M|N,N}^{\mathcal{G}_i}(j|i,i) = A_N^M(i,j) A_M^N(j,i)/C$ as shown in Sec. 3.4. The higher $s_{i,j}^{\text{biwalk}}$, the stronger the similarity. Enforcing that the cycle transition is satisfied - *i.e.* $p_{N|N}^{\mathcal{G}_i}(i|i) \geq \beta_{\text{cycle}}$ - allows to reject false positive matches. We ablate on the superiority of our biwalk similarity over other appearance match metrics in Sec. 4.5.

Data Association. Inspired by BYTE [56], we adopt a two-stage data association scheme. In our first association stage, we propose to associate high-confidence detections to tracklets based on the max-likelihood transition state under motion-constrained bi-directional walks. We then follow the original BYTE implementation for the second association stage. Pseudo-code in the Appendix.

We here describe in details our first association stage. We define a novel gating function W for Hungarian assignment of detections to matching candidates based on motion-constrained appearance similarity. In particular, we combine our appearance similarity metric *biwalk* with spatial proximity between the detected objects N and the matching candidates M refined by Kalman filtering.

First, we adopt the Kalman filter [23] to predict the future location of the matching candidates. We estimate the *motion cost* via the IoU distance $d_{i,j}^{\text{IoU}} = 1 - IoU(M_i, N_j)$ between the i-th predicted bounding box and j-th detected one. We estimate the *appearance cost* via the biwalk distance $d_{i,j}^{\text{biwalk}} = 1 - s_{i,j}^{\text{biwalk}}$. Similarly to [1], we reject appearance-based matches for objects that are spatially far-apart - *i.e.* $d_{i,j}^{\text{IoU}} \geq \beta_{\text{IoU}}$ - or with dissimilar appearance - *i.e.* $d_{i,j}^{\text{biwalk}} \geq \beta_{\text{biwalk}}$ - by setting their cost to 1:

$$\hat{d}_{i,j}^{\text{biwalk}} = \begin{cases} d_{i,j}^{\text{biwalk}}, & \text{if } (d_{i,j}^{\text{IoU}} < \beta_{\text{IoU}}) \wedge (d_{i,j}^{\text{biwalk}} < \beta_{\text{biwalk}}) \\ 1, & \text{otherwise} \end{cases} \tag{13}$$

Finally, we fuse the appearance- $\hat{d}_{i,j}^{\text{biwalk}}$ and motion-based $d_{i,j}^{\text{IoU}}$ costs as their element-wise minimum: $W_{i,j} = \min\{\lambda_{\text{biwalk}} \cdot \hat{d}_{i,j}^{\text{biwalk}}, d_{i,j}^{\text{IoU}}\}$, with λ_{biwalk} relative weight of the appearance cost wrt. motion. We use the fused cost matrix W for Hungarian assignment of detections to matching candidates.

4 Experiments

We provide details on our evaluation protocol for self-supervised MOT methods (Sec. 4.1). We report implementation details in Sec. 4.2. We compare our method with the state of the art in MOT on sparsely (Sec. 4.3) and densely (Sec. 4.4) annotated videos. Finally, we conduct ablation studies in Sec. 4.5.

4.1 Evaluation Protocol

We aim to evaluate the effectiveness of self-supervised MOT methods for learning appearance and their sensitivity to different annotation sparsity levels.

Datasets. *MOT17* [10] is one of the most popular pedestrian tracking datasets, annotated at 14 $\sim$ 30 FPS and featuring 7 training and 7 test sequences in crowded street scenes. *DanceTrack* [42] is a challenging tracking dataset for pedestrians in uniform appearance and diverse motion. Annotated at 20 FPS, it includes 40 videos for training, 25 for validation, and 35 for testing. Its appearance uniformity provides a challenging setting for appearance-based trackers, and even more for self-supervised ones. *BDD100K* [54] is a driving dataset annotated at 5 FPS, counting 1400 sequences for training, 200 for validation, and 400 for testing. Featuring 8 classes, it allows to validate MOT methods in a multi-class setting. We report the most popular metrics for each dataset.

Annotation Sparsity. We evaluate self-supervised MOT under two detection annotation settings during training, *i.e.* dense and sparse. Tracking labels are never provided. In the *sparse* setting, detection annotations are provided for only one every k frames. This is the most practical setting, as it is undesirable to annotate all frames in a video. We thus compare self-supervised trackers trained with detection annotations at 0.1 FPS, a value sensitively below the minimal annotation rate in tracking datasets (1 FPS [9]) and sparser than the average object living time in a video. In the *dense* setting, detection annotations are provided for all frames to compare self-supervised to supervised MOT.

Self-supervised Baselines. We evaluate all models using the YOLOX detector, a 4conv-1fc *embedding head*, and QDTrack's [14] appearance-only data association scheme. First, we compare across all settings to QDTrack-S [14], which uses data augmentation for image-level self-supervision. Then, we ablate against the self-supervised Re-ID literature (Tab. 4) by extending MvMHAT [15] and ReMOTS [51] to the joint detection and tracking setting. Moreover, Moreover, we apply the original contrastive random walk for pixel correspondences [22] on our quasi-dense TOAG defined in Sec. 3.2. We refer to it as QD-CRW. Finally, we introduce an appearance-only variant of Walker that follows QDTrack's data association scheme, namely QD-Walker. Details in the Appendix.

4.2 Implementation Details

In the *sparse* setting, we select positive nodes for our appearance graph (Sec. 3.2) by their IoU with high-confidence detections, and with the available ground-truth boxes in the *dense* setting. We train Walker using a batch size of 16 and an initial

Table 1: State of the art on DanceTrack. We compare existing methods on DanceTrack's test set under sparse (0.1 FPS) and dense (20 FPS) annotations. Methods in black use self-supervised appearance.

	Self. Sup.	Method	HOTA	AssA	DetA	MOTA	IDF1
Sparse		QDTrack-S [14]	29.2	12.3	70.2	79.3	22.6
	✓	QD-Walker (ours)	41.0	23.2	**72.6**	85.8	39.9
		Walker (ours)	**45.9**	**29.5**	71.9	**86.2**	**49.0**
Dense	✗	FairMOT [57]	39.7	23.8	66.7	82.2	40.8
		CenterTrack [59]	41.8	22.6	78.1	86.8	35.7
		TransTrack [43]	45.5	27.5	75.9	88.4	45.2
		ByteTrack [56]	47.7	32.1	71.0	89.6	53.9
		QDTrack [14]	54.2	36.8	80.1	87.7	50.4
		MOTR [55]	54.2	40.2	73.5	79.7	51.5
		OC-SORT [6]	55.1	38.3	80.3	92.0	54.6
	✓	QDTrack-S	38.3	19.8	77.2	85.4	33.6
		QD-Walker (ours)	49.8	32.2	**77.3**	89.4	49.3
		Walker (ours)	**52.4**	**36.1**	76.5	**89.7**	**55.7**

learning rate of 0.00025, decayed with a cosine schedule after a one-epoch warm-up. We initialize the detector from a COCO pre-trained model. We train on 8 GPUs NVIDIA RTX 3090. On MOT17, we follow the private detector half-train/half-val protocol, training for 50 epochs on the union of CrowdHuman [40] and MOT17 [6, 14, 56]. On DanceTrack and BDD100K, we train for 12 and 25 epochs. On MOT17, we apply offline tracklet interpolation [1, 14, 56].

4.3 Sparse Annotations - Comparison with the State of the Art

The sparse setting is the most relevant for assessing self-supervised MOT (Sec. 4.1). We here consider a 0.1 FPS annotation rate and ablate on the effect of different annotation sparsity rates on self-supervised trackers in Sec. 4.5.

Dancetrack. DanceTrack challenges appearance-based trackers by featuring dancing people with uniform appearance. While previous work [14,55] shows that supervised methods can rely on fine details to learn meaningful appearance, the same has never been shown for self-supervised ones. Our experiments (Tab. 1, **Sparse**) show that Walker and QD-Walker significantly outperform QDTrack-S by +16.7 HOTA [33] and with more than twice the association accuracy (AssA) (29.5 vs. 12.3). We argue that Walker's remarkable improvement over QDTrack-S is due to its access to the unlabeled video stream during self-supervision, which allows Walker to learn how to match under the rapid pose changes across DanceTrack's neighboring frames. Since QDTrack-S is only exposed to individual frames during training, it cannot deal with rapid pose changes.

BDD100K. Similar observations hold for BDD100K (Tab. 2, **Sparse**). Walker learns more discriminative multi-class appearance descriptors than QDTrack-S.

4.4 Dense Annotations - Comparison with the State of the Art

Although Walker learns appearance representations in a self-supervised way, we show that it impressively reports competitive performance with the supervised

Table 2: State of the art on BDD100K. We compare with existing methods on the BDD100K test set under sparse (0.1 FPS) and dense (5 FPS) annotations. Methods in black use self-supervised appearance.

	Self. Sup.	Method	mMOTA	mIDF1	MOTA	IDF1
Sparse		QDTrack-S [14]	37.1	49.7	63.5	64.0
	✓	QD-Walker (ours)	37.8	52.3	64.7	67.2
		Walker (ours)	**39.0**	**54.1**	**68.2**	**70.1**
Dense		Yu *et al.* [54]	26.3	44.7	58.3	68.2
		DeepSORT [49]	31.6	38.7	56.9	56.0
	✗	TETer [26]	37.4	53.3	-	-
		ByteTrack [56]	40.1	55.8	69.9	71.3
		QDTrack [14]	42.4	55.6	68.4	73.9
		QDTrack-S [14]	38.7	50.3	65.2	66.8
	✓	QD-Walker (ours)	39.6	53.4	65.9	69.7
		Walker (ours)	**41.2**	**56.1**	**68.3**	**72.1**

Table 3: State of the art on MOT17. We compare methods with private detectors on MOT17's test set under dense annotations (14 $\sim$ 30 FPS). Methods in black use self-supervised appearance.

	Self. Sup.	Method	HOTA	AssA	DetA	MOTA	IDF1
Dense		CenterTrack [59]	52.2	51.0	53.8	67.8	64.7
		FairMOT [57]	59.3	58.0	60.9	73.7	72.3
		TransTrack [43]	54.1	47.9	61.6	63.9	74.5
		ByteTrack [56]	63.1	62.0	64.5	77.3	80.3
	✗	QDTrack [14]	63.5	62.6	64.5	78.7	77.5
		MOTR [55]	57.8	55.7	60.3	68.6	73.4
		OC-SORT [6]	63.2	63.2	-	77.5	78.0
		StrongSORT++ [11]	64.4	64.4	64.6	79.5	79.6
		BoT-SORT [1]	64.6	-	-	79.5	80.6
		QDTrack-S [14]	58.9	59.2	62.6	74.4	74.0
	✓	QD-Walker (ours)	61.7	60.6	63.1	75.4	74.2
		Walker (ours)	**63.6**	**63.0**	**64.0**	**78.2**	**77.4**

state of the art on MOT17 [10], DanceTrack [42], and BDD100K [54]. Walker's training follows the dense protocol (Sec. 4.1).

Dancetrack. (Tab. 1, **Dense**) shows that our self-supervised appearance-only Walker outperforms several popular trackers, including ByteTrack. Its high-quality appearance representations make Walker competitive with other supervised methods such as QDTrack [14] and MOTR [55], even achieving the highest IDF1 across all methods.

BDD100K. On the multi-class dataset BDD100K (Tab. 2, **Dense**), Walker outperforms the supervised appearance-based TETer [26] and improves over ByteTrack [56], demonstrating the importance of appearance descriptors in tracking.

MOT17. The relatively linear motion of pedestrians in MOT17 (Tab. 3) makes the benchmark particularly suitable for motion-based trackers. Nevertheless, our self-supervised appearance-only baseline QD-Walker approaches supervised appearance-only trackers such as QDTrack and MOTR, and the full Walker further improves it and reports competitive performance.

Table 4: Comparison to self-supervised Re-ID (†) and self-supervised correspondence (‡) approaches on DanceTrack val. For a fair comparison, all baselines share the same architecture and inference algorithm as our appearance-only QD-Walker.

	Method	HOTA	AssA	DetA
Sparse	QD-CRW‡ [22]	18.4	4.8	**72.7**
	MvMHAT† [15]	40.7	23.4	71.6
	ReMOTS† [51]	41.0	23.5	71.8
	QD-Walker (Ours)	42.2	24.7	71.7
	Walker (Ours)	**47.6**	**31.0**	71.5
Dense	QD-CRW‡ [22]	19.2	5.1	74.1
	MvMHAT† [15]	44.6	26.9	**75.0**
	ReMOTS† [51]	45.2	27.5	74.8
	QD-Walker (Ours)	49.0	32.8	73.6
	Walker (Ours)	**53.0**	**38.6**	73.1

Fig. 4: Self-supervised MOT under different annotation sparsity rates (FPS) during training. We compare video-level (Walker; QD-Walker) and frame-level (QDTrack-S) self-supervision. †: reference QDTrack fully-supervised at 20 FPS.

4.5 Ablation Study

Annotations Sparsity. We argue that a good self-supervised MOT method must fully utilize the available unlabeled data to learn meaningful appearance representations. Thus, we compare in Fig. 4 the sensitivity to different annotation sparsity levels during training for representative self-supervised MOT methods. We compare: QDTrack-S [14], which relies on image-level self-supervision by augmenting static images; QD-Walker, which shares the same architecture and appearance-only tracking algorithm with QDTrack-S but utilizes video-level self-supervision; Walker, which further combines motion cues to appearance ones. All methods use YOLOX as detector. We assess their AssA at different annotation frame rates - varying from 0.05 to 20 FPS - on the DanceTrack validation set. We find that our video-level self-supervision is considerably more robust to annotation sparsity, and it can outperform image-level self-supervision even when reducing the number of annotated frames by 400x. Moreover, complementing appearance with motion, Walker's performance remains remarkably stable at any annotation frame rate, outperforming the fully supervised QDTrack despite not using tracking labels and even with up to 10x less annotated frames.

Self-supervised Re-ID. As motivated in Sec. 4.1, we extend baselines from the self-supervised Re-ID [15, 51] and correspondence learning [22] literature to the joint detection and tracking problem. For a fair comparison, all methods share the same architecture and inference algorithm as the appearance-only QD-Walker. Walker additionally uses motion to reject unlikely appearance-based associations. Compared to all other baselines, both QD-Walker and Walker show stark superiority in association accuracy, proving the superiority of our self-supervised appearance-learning scheme. Moreover, the comparison to QD-CRW

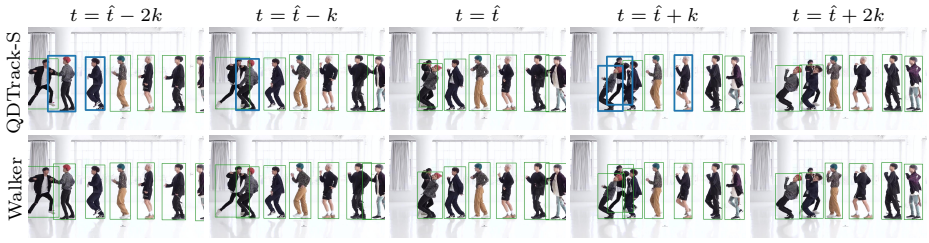

Fig. 5: We analyze 5 frames spaced by 0.2 seconds of the DanceTrack sequence *0058*. Compared to image-level self-sup. (QDTrack-S [14]), Walker effectively utilizes the temporal information to reduce ID switches (blue). Correctly tracked boxes in green.

indicates that the original single-positive contrastive random walk is suboptimal on quasi-dense TOAGs. We argue that: (i) a single positive formulation introduces several false negatives in the optimized loss; (ii) by not enforcing mutual exclusivity its assignments are ambiguous for MOT, where one detection must be assigned to at most one tracklet. This further validates the importance of our contributions towards learning an optimal TOAGs topology for MOT.

Qualitative Results. We analyze 5 frames spaced by 0.2 seconds of the DanceTrack sequence 0058 (Fig. 5). Walker eliminates the ID switches caused by occlusions and rapid pose changes, further validating that - unlike QDTrack-S - Walker can effectively learn to disambiguate non-rigid objects under rapidly varying poses by learning from the temporal stream.

5 Conclusion

This paper introduces Walker, the first self-supervised multiple object tracker that learns from sparse detection annotations and no instance IDs. Walker self-supervises appearance representations by optimizing the topology of a cleverly-designed temporal object appearance graph (Sec. 3.2). We let meaningful instance similarities (edges) emerge by optimizing our multi-positive contrastive random walks (Sec. 3.3), and enforce the mutually-exclusive graph connectivity necessary to downstream association (Sec. 3.4). By relying on video-level self-supervision, Walker effectively makes use of the unlabeled frames in sparsely annotated datasets. As a result, Walker significantly outperforms previous state-of-the-art self-supervised trackers [14] even when trained with 400x less annotated frames. Remarkably, Walker is the first self-supervised tracker to achieve competitive performance with state-of-the-art supervised trackers on a variety of benchmarks. We hope that our work will inspire future research in downstream tracking applications dealing with limited labels, *e.g.* open-world and open-vocabulary tracking [27, 52], domain adaptation [39], continual learning [31]. Finally, by replacing the commonly-used frame-level self-supervision with our video-level self-supervision, we believe that our contributions will enable training stronger foundational models for multiple object tracking [28].

Acknowledgements

This work was supported in part by the Max Plank ETH Center for Learning Systems.

References

1. Aharon, N., Orfaig, R., Bobrovsky, B.Z.: Bot-sort: Robust associations multi-pedestrian tracking. arXiv preprint arXiv:2206.14651 (2022)
2. Athar, A., Luiten, J., Voigtlaender, P., Khurana, T., Dave, A., Leibe, B., Ramanan, D.: Burst: A benchmark for unifying object recognition, segmentation and tracking in video. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1674–1683 (2023)
3. Bastani, F., He, S., Madden, S.: Self-supervised multi-object tracking with cross-input consistency. Advances in Neural Information Processing Systems **34**, 13695–13706 (2021)
4. Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: 2016 IEEE international conference on image processing (ICIP). pp. 3464–3468. IEEE (2016)
5. 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)
6. Cao, J., Weng, X., Khirodkar, R., Pang, J., Kitani, K.: Observation-centric sort: Rethinking sort for robust multi-object tracking. arXiv preprint arXiv:2203.14360 (2022)
7. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
8. Collicott, B., Sarvaiya, M., Weston, B.: Self-supervised feature learning for online multi-object tracking
9. Dave, A., Khurana, T., Tokmakov, P., Schmid, C., Ramanan, D.: Tao: A large-scale benchmark for tracking any object. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16. pp. 436–454. Springer (2020)
10. Dendorfer, P., Osep, A., Milan, A., Schindler, K., Cremers, D., Reid, I., Roth, S., Leal-Taixé, L.: Motchallenge: A benchmark for single-camera multiple target tracking. International Journal of Computer Vision **129**, 845–881 (2021)
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. Elhoseny, M.: Multi-object detection and tracking (modt) machine learning model for real-time video surveillance systems. Circuits, Systems, and Signal Processing **39**(2), 611–630 (2020)
13. Ess, A., Schindler, K., Leibe, B., Van Gool, L.: Object detection and tracking for autonomous navigation in dynamic environments. The International Journal of Robotics Research **29**(14), 1707–1725 (2010)
14. Fischer, T., Pang, J., Huang, T.E., Qiu, L., Chen, H., Darrell, T., Yu, F.: Qdtrack: Quasi-dense similarity learning for appearance-only multiple object tracking. arXiv preprint arXiv:2210.06984 (2022)

15. Gan, Y., Han, R., Yin, L., Feng, W., Wang, S.: Self-supervised multi-view multi-human association and tracking. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 282–290 (2021)
16. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems **33**, 21271–21284 (2020)
17. Gupta, A., Wu, J., Deng, J., Fei-Fei, L.: Siamese masked autoencoders. arXiv preprint arXiv:2305.14344 (2023)
18. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
19. Heigold, G., Minderer, M., Gritsenko, A., Bewley, A., Keysers, D., Lučić, M., Yu, F., Kipf, T.: Video owl-vit: Temporally-consistent open-world localization in video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13802–13811 (2023)
20. Ho, K., Kardoost, A., Pfreundt, F.J., Keuper, J., Keuper, M.: A two-stage minimum cost multicut approach to self-supervised multiple person tracking. In: Proceedings of the Asian Conference on Computer Vision (2020)
21. Huang, K., Lertniphonphan, K., Chen, F., Li, J., Wang, Z.: Multi-object tracking by self-supervised learning appearance model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3162–3168 (2023)
22. Jabri, A., Owens, A., Efros, A.: Space-time correspondence as a contrastive random walk. Advances in neural information processing systems **33**, 19545–19560 (2020)
23. Kalman, R.E.: A new approach to linear filtering and prediction problems (1960)
24. Karthik, S., Prabhu, A., Gandhi, V.: Simple unsupervised multi-object tracking. arXiv preprint arXiv:2006.02609 (2020)
25. Kim, S., Lee, J., Ko, B.C.: Ssl-mot: self-supervised learning based multi-object tracking. Applied Intelligence **53**(1), 930–940 (2023)
26. Li, S., Danelljan, M., Ding, H., Huang, T.E., Yu, F.: Tracking every thing in the wild. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII. pp. 498–515. Springer (2022)
27. Li, S., Fischer, T., Ke, L., Ding, H., Danelljan, M., Yu, F.: Ovtrack: Open-vocabulary multiple object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5567–5577 (2023)
28. 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)
29. Li, S., Ke, L., Yang, Y.H., Piccinelli, L., Segu, M., Danelljan, M., Van Gool, L.: Slack: Semantic, location and appearance aware open-vocabulary tracking. In: Computer Vision–ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings. Springer (2024)
30. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
31. Liu, Z., Segu, M., Yu, F.: Cooler: Class-incremental learning for appearance-based multiple object tracking. arXiv preprint arXiv:2310.03006 (2023)
32. Lu, Z., Rathod, V., Votel, R., Huang, J.: Retinatrack: Online single stage joint detection and tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14668–14678 (2020)

33. Luiten, J., Osep, A., Dendorfer, P., Torr, P., Geiger, A., Leal-Taixé, L., Leibe, B.: Hota: A higher order metric for evaluating multi-object tracking. International journal of computer vision **129**, 548–578 (2021)
34. 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)
35. Pang, J., Qiu, L., Li, X., Chen, H., Li, Q., Darrell, T., Yu, F.: Quasi-dense similarity learning for multiple object tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 164–173 (2021)
36. Park, Y., Lepetit, V., Woo, W.: Multiple 3d object tracking for augmented reality. In: 2008 7th IEEE/ACM International Symposium on Mixed and Augmented Reality. pp. 117–120. IEEE (2008)
37. Reid, D.: An algorithm for tracking multiple targets. IEEE transactions on Automatic Control **24**(6), 843–854 (1979)
38. 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)
39. 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)
40. Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X., Sun, J.: Crowdhuman: A benchmark for detecting human in a crowd. arXiv preprint arXiv:1805.00123 (2018)
41. 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)
42. Sun, P., Cao, J., Jiang, Y., Yuan, Z., Bai, S., Kitani, K., Luo, P.: Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20993–21002 (2022)
43. 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)
44. Sun, T., Segu, M., Postels, J., Wang, Y., Van Gool, L., Schiele, B., Tombari, F., Yu, F.: Shift: a synthetic driving dataset for continuous multi-task domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21371–21382 (2022)
45. Wang, X., Jabri, A., Efros, A.A.: Learning correspondence from the cycle-consistency of time. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2566–2576 (2019)
46. Wang, Y.H.: Smiletrack: Similarity learning for multiple object tracking. arXiv preprint arXiv:2211.08824 (2022)
47. Wang, Z., Zhao, H., Li, Y.L., Wang, S., Torr, P., Bertinetto, L.: Do different tracking tasks require different appearance models? Advances in Neural Information Processing Systems **34**, 726–738 (2021)
48. Wang, Z., Zheng, L., Liu, Y., Li, Y., Wang, S.: Towards real-time multi-object tracking. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16. pp. 107–122. Springer (2020)

49. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: 2017 IEEE international conference on image processing (ICIP). pp. 3645–3649. IEEE (2017)
50. Wu, Y., He, K.: Group normalization. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
51. Yang, F., Chang, X., Dang, C., Zheng, Z., Sakti, S., Nakamura, S., Wu, Y.: Remots: Self-supervised refining multi-object tracking and segmentation. arXiv preprint arXiv:2007.03200 (2020)
52. Yang, J., Gao, M., Li, Z., Gao, S., Wang, F., Zheng, F.: Track anything: Segment anything meets videos. arXiv preprint arXiv:2304.11968 (2023)
53. Yang, L., Fan, Y., Xu, N.: Video instance segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5188–5197 (2019)
54. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2636–2645 (2020)
55. Zeng, F., Dong, B., Zhang, Y., Wang, T., Zhang, X., Wei, Y.: Motr: End-to-end multiple-object tracking with transformer. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVII. pp. 659–675. Springer (2022)
56. Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X.: Bytetrack: Multi-object tracking by associating every detection box. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII. pp. 1–21. Springer (2022)
57. Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W.: Fairmot: On the fairness of detection and re-identification in multiple object tracking. International Journal of Computer Vision **129**, 3069–3087 (2021)
58. Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., Tian, Q.: Mars: A video benchmark for large-scale person re-identification. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14. pp. 868–884. Springer (2016)
59. Zhou, X., Koltun, V., Krähenbühl, P.: Tracking objects as points. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV. pp. 474–490. Springer (2020)