Tracking by Associating Clips

Sanghyun Woo[1], Kwanyong Park[1],
Seoung Wug Oh[2], In So Kweon[1], and Joon-Young Lee[2]

[1] KAIST
[2] Adobe Research

Abstract. The tracking-by-detection paradigm today has become the dominant method for multi-object tracking and works by detecting objects in each frame and then performing data association across frames. However, its sequential *frame-wise matching* property fundamentally suffers from the intermediate interruptions in a video, such as object occlusions, fast camera movements, and abrupt light changes. Moreover, it typically overlooks temporal information beyond the two frames for matching. In this paper, we investigate an alternative by treating object association as **clip-wise matching**. Our new perspective views a single long video sequence as multiple short clips, and then the tracking is performed both within and between the clips. The benefits of this new approach are two folds. First, our method is robust to tracking error accumulation or propagation, as the video chunking allows bypassing the interrupted frames, and the short clip tracking avoids the conventional error-prone long-term track memory management. Second, the multiple frame information is aggregated during the clip-wise matching, resulting in a more accurate long-range track association than the current frame-wise matching. Given the state-of-the-art tracking-by-detection tracker, QDTrack, we showcase how the tracking performance improves with our new tracking formulation. We evaluate our proposals on two tracking benchmarks, TAO and MOT17 that have complementary characteristics and challenges each other.

Keywords: Clip-based Tracking, Long-term Video Modeling

1 Introduction

Discriminating the identity of multiple objects in a scene and providing individual trajectories of their movements over time, namely multi-object tracking, is one of the fundamental computer vision problems, imperative to tackle many real-world problems, *e.g.* autonomous driving and surveillance. Despite being a rather classical vision task, it is still challenging to design a robust multi-object tracker capable of tracking a time-varying number of objects moving through unconstrained environments in the presence of many other complexities.

Early studies approached the multi-object tracking problem by breaking it into multiple sub-problems that could be tackled individually, typically starting with object detection, followed by association, track management, and post-processing [1,2,6,14,35,39]. Ever since, this tracking-by-detection paradigm has

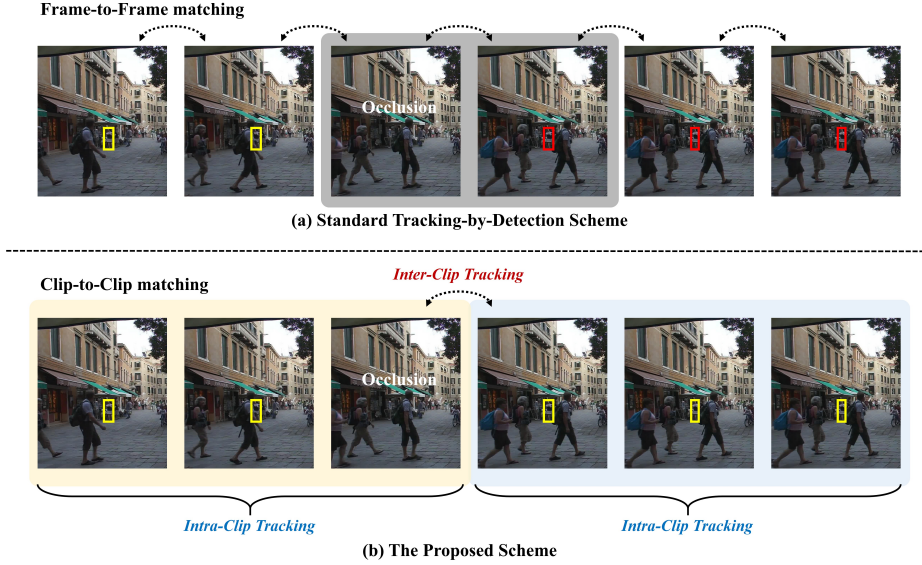**(a) Standard Tracking-by-Detection Scheme**

(b) The Proposed Scheme

Fig. 1: **The proposed Clip-to-Clip matching.** (a) The current tracking paradigm sequentially matches the object instances frame-by-frame, which is vulnerable to intermediate interruptions in a video and cannot faithfully exploit the temporal information during the matching, resulting in tracker drifting. (b) Unlike the standard frame-wise matching scheme, we introduce a new clip-wise tracking method. Our approach is robust to random interference in a video, as we can skip those frames by chunking the video into multiple clips. Moreover, the multiple frame information is exploited during the clip-wise matching. With our proposal, the object can be seamlessly tracked over time.

become the standard approach for multi-object tracking, and most of the state-of-the-art trackers follow this scheme [4,44,23,52,28,49,51].

The tracking-by-detection scheme is essentially a sequential frame-by-frame association approach (see Fig. 1-(a)). In practice, it is achieved by matching the temporally smoothed history, which can be either new locations based on the past motion records [51] or moving averaged RE-ID features of the trajectories [28], with the current predictions. However, this sequential frame-based matching fundamentally suffers from two prominent limitations. First, the error accumulation or propagation cannot be handled properly. For example, sudden object motion pattern changes or fast camera movements significantly disturb the previous motion records. Also, object deformation such as occlusion or blur corrupts the moving averaged RE-ID features, leading to track drifting. The phenomenon becomes even more severe when the input video frame rate is low. Second, it does not consider the overall temporal context during matching. Looking at only two frames for matching essentially has ambiguity, resulting in track fragmentation.

To effectively utilize the temporal information and make the tracker robust to intermediate interruptions in a video, we reformulate the standard tracking-by-detection scheme as a clip-to-clip matching problem (see Fig. 1-(b)). In particular, we chunk a video into multiple short clips and perform the tracking at the clip level. The video chunking allows skipping challenging frames for tracking, and the short-clip tracking avoids long-term track memory management such as long sequence temporal smoothing of RE-ID features along the video. Furthermore, the association between clips utilizes multi-frame information, which is more robust than the previous frame-based association. While not being sensitive to any specific design of tracker, we build our proposal upon the state-of-the-art RE-ID-based tracking-by-detection paradigm, QDTrack [28], as it can handle both the low [11] and high frame rate [27] video inputs better than the motion-based trackers [55,23,52,47,51].

We define two new basic tracking operations to implement the clip-wise tracker: intra- (within-) and inter- (between-) clip tracking. We target *semi-online scenarios* throughout this paper. Thus we only allow the intra-clip association to perform matching irrespective of the frame order. Meanwhile, inter-clip tracking incrementally merges the clip predictions until the complete video-level object trajectories are all retrieved. It requires associating objects at a track level, and thus the temporal information plays an important role here.

We investigate various viable implementations for both intra- and inter-clip tracking. Specifically, for the intra-clip tracking, we explore two feasible association approaches, directional and direction-free matching [53]. For the inter-clip tracking, we consider the IoU-based chaining [31] and temporal averaged feature matching. We further improve the inter-clip matching by designing a novel transformer-based clip tracker. It learns to temporally attend to all the past track embeddings of each object and predicts representative embeddings. These temporally attended embeddings are matched to associate the clip-level predictions. To make the model more robust at test time, we simulate hard positives and negatives during training: mislocalization and track drifting. We apply our inter-clip tracker recursively over time and merge the clip predictions into global object tracks incrementally.

With our proposals, we achieve state-of-the-art results on the challenging large vocabulary object tracking benchmark, TAO. We also validate our approach on the MOT17 benchmark to focus on the tracking performance more thoroughly. Compared to the baseline, we observe clear improvement in the association performance. Finally, we conduct extensive ablation studies and confirm that the proposals are effective and generic.

2   Related Work

Multi-Object Tracking. The tracking-by-detection paradigm [32] dominates the current state-of-the-art multi-object tracking frameworks [22]. It detects objects first and associates them frame-by-frame across time [1,2,6,14,35,39]. Recently, deep learning has contributed significantly to improving the performance of

multi-object tracking approaches by focusing on designing better detectors [34,33] or developing more effective association objectives [18,21,37,44,46,28]. Several works also have shown advances regarding detection and tracking as a joint learning task [13,41,47,55]. Nonetheless, these methods often formulate the multi-object tracking problem only with two consecutive frames and dismiss long-term temporal information, which is crucial for tackling various challenges in tracking.

While sequential frame-based matching is efficient and has shown promising results, they are fundamentally vulnerable to intermediate interruptions in a video and miss rich temporal information during matching. Unlike these methods, we target a semi-online scenario where we allow slight lagging and utilize several consecutive frames (*i.e.* clip) as a whole for the association. We show that our proposal is robust to intermediate interruptions in a video and produces better association quality.

Clip-level Modeling in Video. The simultaneous process of multiple frames (*i.e.* clip-level modeling) is a promising research direction in many video tasks including video instance segmentation [3,43,19,48], video object segmentation [29], optical flow estimation [20], and depth estimation [45]. Per-clip models enjoy the opportunities to extract rich temporal information from multiple frames, leading to effectively tackle ambiguities in videos (*e.g.* occlusions or blur). Most existing methods focus on extracting temporal information to improve recognition or segmentation quality, and there has been little consideration to the association part.

Here, we present to associate objects at clip-level, performing the tracking both within and in-between the clips. The clip-level matching provides two clear advantages. First, by chunking, the matching within the clips essentially becomes short-term tracking, avoiding long-term history averaging or memory management. Second, the matching in-between the clips allows exploiting the temporal context, resolving the temporally-local ambiguities of matching between the two consecutive frames. To achieve better long-term association with accurate inter-clip matching, we design a novel transformer architecture.

Transformers for Tracking. Recently, transformers [42] have shown impressive results in many computer vision tasks, such as image classification [12,24], object detection [7,56], segmentation [54], and image generation [30]. There are also several methods [40,26,49] adopting the transformers for multi-object tracking by extending DETR frameworks [7]. However, these methods are still limited at utilizing short-term temporal information and rather adopt conventional heuristics such as Intersection over Union (IoU) matching [40], formulate the problem as a two frames task [26], or rely on the frame-by-frame evolving track quries [49].

Apart from the previous works, we set up the clip-based tracking scenario. We see that the key to accurate inter-clip matching is to generate discriminative clip-level representations. Our main idea is to use a Transformer as a track history summarizer. In practice, given a short track sequence of object instance (*i.e.* intra-clip tracking results), we input this track sequence to the transformer and

---

Algorithm 1: Tracking by Associating Clips

Input: A video sequence $\mathtt{V} = (f_1, \ldots, f_N)$; object detector $\mathtt{Det}$; clip size $\mathtt{C}_S$;
   clip interval $\mathtt{C}_I$;

Output: Tracks $\mathcal{T}$ of the video

1 begin

2 | Initialization: $\mathcal{T} \leftarrow \emptyset$

3 | **for** $i = 1 : C_I : N$ **do**

4 | | /* Sample clip */

5 | | $C \leftarrow \mathtt{V}[i : i + \mathtt{C}_S]$

6 |

7 | | /* Predict detection boxes & scores */

8 | | $\mathcal{D}_{clip} \leftarrow \emptyset$

9 | | **for** *frame f_k in C* **do**

10 | | | $\mathcal{D}_k \leftarrow \mathtt{Det}(f_k)$

11 | | | $\mathcal{D}_{clip} \leftarrow \mathcal{D}_{clip} \cup \mathcal{D}_k$

12 |

13 | | /* Intra clip association*/

14 | | Initialization: $\mathcal{T}_{intra} \leftarrow \emptyset$

15 | | Associate $\mathcal{T}_{intra}$ and $\mathcal{D}_{clip}$

16 |

17 | | /* Inter clip association*/

18 | | Associate $\mathcal{T}$ and $\mathcal{T}_{intra}$

---

get the condensed feature. Finally, this track summary feature is used for the subsequent inter-clip matching.

3   Tracking by Associating Clips

We propose a simple yet effective data association method, Tracking by Associating Clips. Unlike the standard tracking-by-detection scheme, which matches object instances frame-by-frame, we instead conduct association at a clip level as shown in Fig. 1. We provide the pseudo-code in Alg. (1).

The input is a video sequence $\mathtt{V}$ and an object detector $\mathtt{Det}$. Also, we set two hyper-parameters, clip size $\mathtt{C}_S$ and clip interval $\mathtt{C}_I$. The output is the tracks $\mathcal{T}$ of the video, and each track contains the bounding box and identity of the object in each frame.

After sampling the clip from the video, we predict the detection boxes and scores using the detector $\mathtt{Det}$ for each frame. Then, the first association is performed to link the detection boxes in the clip $\mathcal{D}_{clip}$, which results in local clip-level tracks $\mathcal{T}_{intra}$. The second association is performed between the local clip-level tracks $\mathcal{T}_{intra}$ and the global video-level tracks $\mathcal{T}$. The global video-level tracks $\mathcal{T}$ keep the history of the tracked objects and use this information to associate with the local clip-level tracks $\mathcal{T}_{intra}$. The object tracks are incrementally expanded over time by repeating the intra- and inter-clip associations. Notably,

the proposal degenerates to the standard tracking-by-detection scheme when the clip window size reduces to 1 without overlaps in between.

In this paper, we built our proposal upon the recently presented QDTrack [28], given its strong Re-ID association capability. It is more robust than motion-based trackers [55,23,52,47,51] to target both the low-frame-rate (*e.g.* 1FPS, TAO [11]) and high-frame-rate (*e.g.* 30FPS, MOT17 [27]) video inputs. In the following, we explore effective implementations of intra- and inter-clip association methods.

3.1   Intra Clip Association

Given the detection boxes of each frame in the clip, we examine two possible implementations, directional and direction-free matching (see Fig. 2-(a)).

Directional Matching. The standard way to associate the clip-level object predictions is to link them sequentially. We see that the matching direction, either left to the right or right to the left, provides similar performance; thus, we only consider the former case. To address object occlusions within the clip, we employ track rebirth [46,9] using a memory mechanism.

Direction-Free Matching. Within the clip, the motion patterns and appearances of objects are quite similar due to the inherent redundancy in a video. Given this fact, we can also associate the objects in a direction-free manner.

Specifically, we employ the heap-based hierarchical clustering algorithm [53]. At the beginning, we consider that every detection box in all frames has its own cluster. We calculate their pair-wise appearance distances [28]. We avoid the association of the objects in the same frame by setting their distance to be infinity. Subsequently, every cluster distance will be inserted into a priority queue based on a heap data structure. Afterward, the cluster pair associated with the smallest distance in the priority queue will be popped and will be merged together. For this new cluster, distances to other clusters have to be computed and inserted into the priority queue. To set the distance between the track pairs, we search the minimum distance. The clustering is repeated until there are only distances greater than a threshold left.

For the intra-clip tracking, which is essentially short-term tracking, we observe that the standard directional matching with the lost track management is already competitive. It outperforms the strong heap-based hierarchical clustering algorithm [53] both in accuracy and efficiency.

3.2   Inter Clip Association

This section explores how to match and merge the current clip-level predictions into the global video-level tracks. We consider the following two feasible implementations (see Fig. 2-(b)).

Iou-Based Matching. We adopt IoU-based chaining [31] that associates object tracks based on the IoU in the same frame. However, this method has three apparent drawbacks. First, it misses linking all the object tracks not revealed in

(a) Intra-Clip Tracking   **(b) Inter-Clip Tracking**

- Directional Matching with Memory   • Direction-Free Hierarchical Matching   • IoU-based Chaining
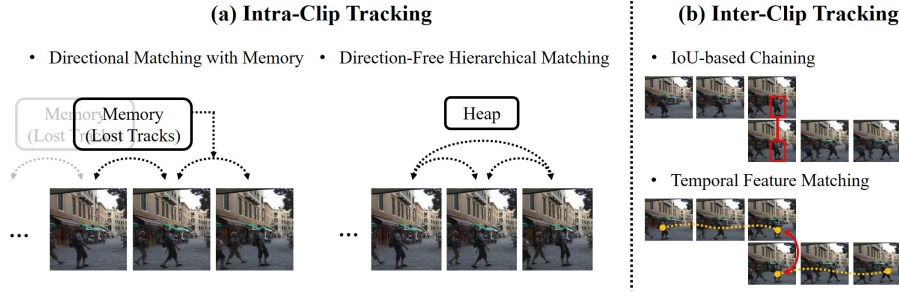

Fig. 2: **Two Basic Operations for Clip-based Tracking.** We define intra- and inter-clip tracking for instantiating the clip-based tracker. For intra-clip tracking, we consider both directional and direction-free matching approaches. The former is a standard frame-by-frame matching with a memory mechanism. The latter is implemented using the heap structure and hierarchically clusters the detection boxes across the frames without considering the frame order [53]. For the inter-clip tracking, we examine IoU-based chaining [31] and temporal average-pooled feature matching.

the overlapping frame, and this phenomenon becomes even more severe as clip window size increases. Second, as the number of overlapping frames increases, the single-frame-based IoU matching does not guarantee optimal matching. Third, the method is basically not applicable when there are no overlapping frames.

Temporal Feature Matching. Free from the above limitations, an alternative is based on feature matching. As a simple baseline, we employ temporally average-pooled features for the association.

While simple, temporal feature matching improves over the frame-based tracking baseline and outperforms the IoU-based chaining [31]. However, we found that it is not always the case; For the low frame rate video inputs (*i.e.* TAO), which inherently include severe appearance change across the frame, the simple averaged temporal matching is inferior to the baseline (see Table 3). To improve the clip-level association quality further, we designed a new Transformer-based clip tracker, described in the following section.

3.3   Clip Tracker

The most challenging aspect of the inter-clip association is to generate a discriminative *clip-level* embedding for any object tracks. Our key idea for this problem is to use transformer as a track history summarizer (see Fig. 3). By design, the transformer can process and relate set-based elements. Here, we consider the intra-clip results (*i.e.* short object track sequence) as an input for the transformer, and we attempt to temporally-pool their feature information into a condensed feature. During training, the clip tracker learns to produce a good summary of the given object track history, which can link the clip-level predictions over time at test time.
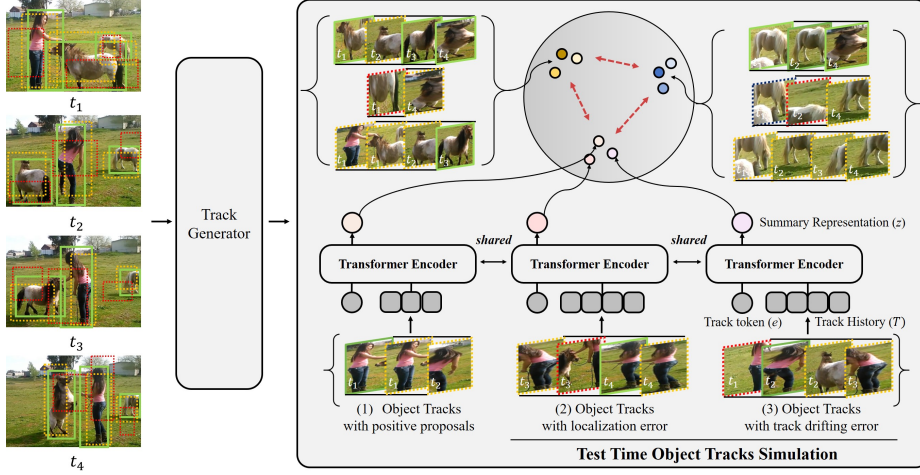
Fig. 3: **The overview of Clip Tracker.** The solid green line indicates the original ground truth boxes of each instance. The yellow and red dotted lines are positive and negative proposals, respectively. The transformer learns a *weighted temporal pooling* function of track history in a data-driven manner. *Best viewed in color.*

In practice, we design the clip tracker using transformer encoder layers. For object p, the clip tracker takes all the feature embeddings in the track $\mathcal{T}^p = \{x_{t_s}^p, \ldots, x_{t_e}^p\} \in R^{L \times C}$ and an extra learnable track token $e \in R^C$ as an input. Here, t_s and t_e denote the track initiation and termination time, respectively. L and C are the track length and the dimension of the feature embeddings, respectively. This input sequence, $[e, \mathcal{T}^p] \in R^{(L+1) \times C}$, is forwarded to the transformer, and we take the output of the track token, $z^p \in R^C$, as a condensed representation of the given object's track history.

This summary representation should be representative enough to match the same object instance over time consistently. Also, it should be discriminative enough to prevent it from being matched to other object instances. This can be achieved with the contrastive learning objective [16], where the positive can be oneself, and set the negatives other object embeddings. However, given that the number of objects appearing in a video is limited in typical, we significantly lack training samples. As a remedy, we introduce three strategies in the following for scaling the training data.

First, we allow using positive proposals [3], and concatenate them over the time axis to generate object tracks. This strategy has two clear advantages: 1) we can significantly increase the number of object tracks compared to only using ground truth boxes. 2) we can enjoy the natural jittering effect of the proposals, allowing the model to be robust to temporal appearance changes(see Fig. 3-(1)).

Second, we simulate hard positives and negatives. The first is mislocalization due to some boxes along the track that does not tightly cover the object. The

---

[3] The proposals that have IoU greater than the given threshold with the ground truths.

second is a track drifting due to the incorrect match to other instances in the middle of the track. To make the model robust to these errors, we augment the training samples with the following two new strategies.

1. **Negative Proposals.** For the localization error, we incorporate negative proposals that have loose IoU with the ground truths (in between 0.3 and 0.5). Given both the positive and negative proposals, three different types of object tracks are generated in equal probability during training; positive only, negative only, and hybrid.

2. **Object Track Mixup.** For the track drifting error, we propose to mix two different object tracks [50]. In particular, we switch random portions of the given object track embeddings with randomly selected different object track embeddings. To keep the original identity of the given track, we bound the mixup ratio to 0.3.

With these two new augmentation strategies, we can effectively simulate hard positives and negatives, which can help the model improve embedding quality (see Fig. 3-(2) and (3)). The proposal sampling is conducted in an order-free manner, which means the object tracks can be constructed with multiple boxes from the same image.

Finally, we treat the object instances from other videos in the training batch as negatives. This naturally increases the negatives in the contrastive learning context, leading to more discriminative feature embeddings [10,17].

Objective Function. Under this setup, we can apply the following contrastive learning formula,

$$\mathcal{L} = \log[1 + \sum_{\mathbf{z}^+} \sum_{\mathbf{z}^-} \exp(\mathbf{z} \cdot \mathbf{z}^- - \mathbf{z} \cdot \mathbf{z}^+)]. \tag{1}$$

Where $\mathbf{z}$ denotes an anchor, $\mathbf{z}^+$ and $\mathbf{z}^-$ are corresponding positives and negatives. We note the formula allows multiple positives, which is an extension of the standard single-positive formula, and thus can learn more discriminative embeddings [28].

Track History Management. At test-time, the clip tracker holds the global video-level object tracks and expands those incrementally by merging the outputs of the intra-clip tracker. In specific, the affinity matrix is computed based on the *summary* feature similarities between clip and global tracks. We associate the tracks with the matching threshold of 0.5. If the match is found, we save all the frame-level predictions of the objects in the memory buffer for the subsequent matching.

Memory Buffer Size. In computing the summary embedding of the global tracks, we set a proper memory buffer size. Intuitively, the larger the memory buffer size, the more the model can exploit the temporal context up to its temporal abstraction capacity. In practice, we set to 30 and 10 frames for MOT17 and TAO, respectively.

4   Experiments

In this section, we conduct experiments to demonstrate the benefits of our proposals using the TAO [11] and MOT17 [27] benchmarks. We investigate the results focusing mainly on the association aspect. Specifically, we use Track AP for the evaluation on the TAO benchmark. Also, Track AP_S, Track AP_M, and Track AP_L are adopted which indicates the Track AP of (short ≤ 3), ($3 \leq$ medium ≤ 10), and (long ≥ 10) length object tracks. We use MOTA [5], IDF1 [36], and the recently presented HOTA metrics [25] for the evaluation on the MOT17 benchmark. To see the association quality more thoroughly, we also present AssA performance.

4.1   Datasets

TAO [11] is the first video benchmark for large vocabulary object tracking. It annotates 482 classes in total, which are the subset of the LVIS [15] dataset. Its training set has 400 videos, covering 216 classes. The validation set consists of 988 videos, spanning 302 classes. The test set has 1419 videos, covering 369 classes. The videos are annotated in 1 FPS.

MOT17 [27] is the most popular multi-object tracking benchmark, which consists of 7 training and 7 testing videos. While it has only pedestrian annotations, sequences include several challenges such as frequent occlusions and crowd scenes. The video framerate is relatively high, ranging from 14 to 30 FPS. For ablation experiments, we split each training sequence into two halves, and use the first half-frames for training and the second for validation.

Due to the difference in frame rate (1FPS *v.s.* 30FPS) and annotated object categories (482 *v.s.* 1), TAO and MOT17 cover significantly different motion patterns and video contents. We show that our proposal works well on both.

4.2   Implementation Details

Clip Tracker Architecture.   The Clip Tracker model consists of 3 transformer encoder layers, where each layer has 8 heads with embedding dimensions of 256 (for MOT17) and 512 (for TAO). The input of Clip Tracker is the track embedding sequence of each object instance and the trainable track token. We apply a single linear layer to embed the input sequence. Finally, the output of the track token passes through a single MLP layer to produce the summary of the object track history. The positional encodings are not used in this work.

Hyperparameters.   The proposed clip-based tracking pipeline introduces two new hyperparameters, clip size and clip interval (see Alg. (1)). The clip size and interval directly affect the video chunking patterns. We observe that overlapping the clip prediction eases the inter-clip association. In practice, we considered the frame rate of the given datasets. For example, for the TAO, which provides low

Method	TAO val			TAO test		
	TrackAP_{50}	TrackAP_{75}	$\text{TrackAP}_{50:95}$	TrackAP_{50}	TrackAP_{75}	$\text{TrackAP}_{50:95}$
SORT [11]	13.2	-	-	10.2	4.4	4.9
QDTrack [28]	13.4	4.9	6.1	12.6	4.5	5.6
Ours	**17.7**	**5.8**	**7.3**	**15.8**	**5.9**	**6.6**

Table 1: Tracking Results on TAO *val* and *test*. We use clip size and interval of 6 and 3, respectively.

Method	MOT val			MOT test		
	MOTA	IDF1	HOTA	MOTA	IDF1	HOTA
QDTrack [28]	68.1	69.5	57.5	71.1	70.2	57.8
Ours	**68.9**	**72.4**	**59.7**	**71.6**	**72.7**	**59.0**

Table 2: Tracking Results on MOT17. We use clip size and interval of 10 and 5, respectively.

frame rate video inputs, we set clip size and interval value to 6 and 3, respectively. For the MOT, we set clip size and interval to 10 and 5, respectively.

Training.   We use MMdetection and MMTracking frameworks [8]. For the TAO experiments, COCO-style training schedule of $2\times$ and $1\times$ are adopted for LVIS pre-training and TAO fine-tuning, respectively. For the MOT experiments, we pre-trained the model using CrowdHuman dataset [38] for 12 epochs and then fine-tuned on the MOT17 train for 4 epochs. During pre-training, the detection part is learned. A batch size of 16 (2 per GPU) and an initial learning rate of 0.02 are used. We use the backbone ResNet-101 (ResNet-50) for the TAO (MOT17) experiment following the previous studies [11,28]. We train the Clip Tracker by sampling 2 and 6 frames under the frame range of 1 and 30 for TAO and MOT17 experiments, respectively.

Testing.   Our method processes video frames recursively at clip-level, generating object tracks for the given clip and merging them into the global tracks incrementally. We use standard directional matching for intra- and Clip-Tracker for inter-clip tracking. The overall tracking procedure follows Alg. (1). While the frame-level association capability within the clip is similar to that of the baseline, video chunking along with the clip-wise association allows tracking performance improvement. We use resized frames of 1080×1080 for testing.

4.3   Main Results

Upon the state-of-the-art tracking-by-detection framework, QDTrack [28], we instantiate our proposal, clip-based tracking. The main results on TAO and MOT17 are summarized in Table 1 and Table 2. We observe our proposal pushes the tracking performances on both datasets. The results imply that the clip-based association is generic over various scenarios.

Method	MOT17			TAO			
	IDF1	HOTA	AssA	T_{50}	T_S	T_M	T_L
QDtrack [28]	69.5	57.5	58.0	13.4	6.7	9.6	17.2
Direc-free [53]	70.3	57.9	58.9	13.1	6.1	8.9	16.8
Direc	70.3	58.6	59.4	12.3	5.1	8.3	16.5

(a) Intra Clip Association Methods

Method	MOT17			TAO			
	IDF1	HOTA	AssA	T_{50}	T_S	T_M	T_L
QDtrack [28]	69.5	57.5	58.0	13.4	6.7	9.6	17.2
IoU-chain [31]	69.1	56.4	56.1	10.2	3.4	6.5	14.4
Tmp avg.	70.3	58.6	59.4	12.3	5.1	8.3	16.8
Clip-Tracker	72.4	59.7	61.9	17.7	10.9	14.6	21.3

(b) Inter Clip Association Methods

Table 3: **Impact of Clip-based Tracking.** We use both TAO *val* and MOT *val* to evaluate the impact of the new clip-based tracking approach. We investigate various feasible implementations for both intra and inter clip associations described in Sec. 3.1 and Sec. 3.2. Here, T is an abbreviation of Track AP.

4.4   Ablation Studies

To empirically confirm the effectiveness of our proposals, we conduct ablation studies using the TAO and MOT17 benchmarks. We focus on the association quality; thus, we use the metrics of IDF1, HOTA, and ASSA for MOT, and Track AP for TAO.

Simple Clip-based Tracker.   We first analyze how the standard QDTrack framework [28] improves by changing the tracking scheme from frame-based to clip-based. To do so, we design a simple clip-based tracking baseline, which takes the clip as input, performs intra-clip tracking, and links the clip predictions based on the temporal averaged feature matching. The results are shown in Table 3a. We observe that the clip-based tracking scheme improves the association quality in the MOT17 benchmark but not TAO. We see this because the motion change is significant in TAO due to the low frame rate; thus, the simple temporal averaged feature performs poorly, which motivates the design of Clip Tracker.

Intra-clip association.   Among the intra-clip variants, we observe that the standard scheme is already competitive and runs in a much faster time. Here, we use directional matching with memory mechanism for the intra-clip tracking.

Inter-clip association.   We perform ablations on inter-clip association methods. The results are summarized in Table 3b. We observe that IoU-based chaining performs inferior, even lower than the baseline. This is because the method misses all the object tracks not in the overlapping frame, leading to significant track fragmentation. Finally, the presented Clip Tracker significantly outperforms the simple temporal averaged feature matching. The results show that learning-based temporal pooling is more robust than simple averaging. We use Clip Tracker as our inter-clip association method in the following experiments.

Parameter Analysis on Clip Size and Interval.   We conducted parameter analysis on the two newly introduced hyperparameters, clip size and interval. These hyperparameters affect the video chunking pattern and, as a result, impact both the intra- and inter-clip association quality. The results are summarized

Method	$[C_S, C_I]$	IDF1	HOTA	AssA
QDtrack [28]		69.5	57.5	58.0
Clip-Tracker	[5,5]	71.6	59.0	60.5
	[10,10]	71.1	58.6	59.8
	[10,5]	72.4	59.7	61.9
	[20,10]	71.9	59.5	61.3

(a) Clip Size and Interval

Loc. Err	Drift. Err	IDF1	HOTA	AssA
		70.3	58.8	60.2
✓		71.9	59.3	61.3
✓	✓	72.4	59.7	61.9

(b) Track Augmentations

Method	IDF1	HOTA	AssA
Two-clip based.	71.8	59.2	60.4
Moving avg.	71.7	59.3	60.9
Feature bank. (Ours)	72.4	59.7	61.9

(c) Global Track Management

Table 4: **Component Analysis on Clip Tracker** using MOT *val.* First, we conduct parameter analysis on two newly introduced hyperparameters: clip size C_S and clip interval C_I. Second, We study the impact of two track augmentation strategies, negative proposal sampling (for localization error simulation) and object track mixup (for track drifting simulation), during the Clip Tracker training. Lastly, we investigate three different baseline methods to manage global video-level tracks.
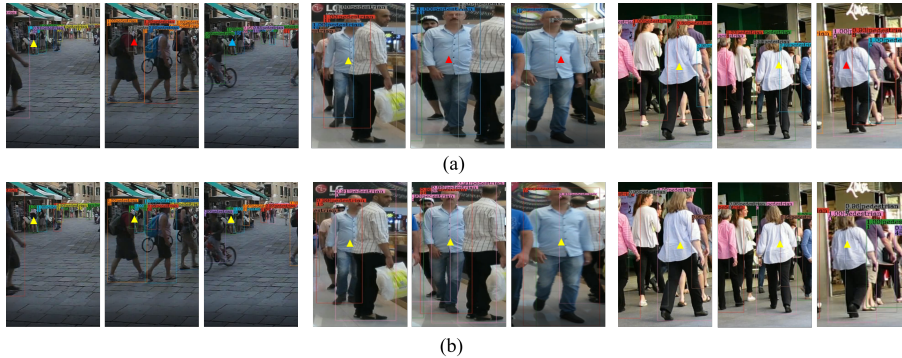

(a)

(b)

Fig. 4: **Qualitative Results.** (a) Tracking by detection. (b) Tracking by associating clips. The colored triangle represents the predicted person id. *Best Viewed in Colors.*

in Table 4a. We observe two tendencies. First, the association improves when the overlapping exists in-between the clips, *i.e.* $C_S > C_I$. Intuitively, the overlapping makes the feature embedding sequence quite similar, and thus the Clip Tracker can more easily find the correct match. Second, we see a slight performance drop when the clip size increases, *e.g.* $[5, 5] \rightarrow [10, 10]$ or $[10, 5] \rightarrow [20, 10]$. We see this phenomenon is mainly due to the increased false positives with the large clip size, and we mitigate this issue by introducing the overlapping frames, $[10, 10] \rightarrow [10, 5]$. In practice, proper clip size and interval values should be set, and here we found the clip size of 10 and the interval of 5 strikes a balance.

Track Augmentations.   In order to improve the robustness of the Clip Tracker, we introduced two new hard training sample generation strategies, negative proposal sampling and object track mixup. We investigate their impacts on the tracking performance in Table 4b. From the baseline of using only the positive proposals, the track performance improves meaningfully with the negative proposal sampling. The object track mixup further improves the score.

Track Management Strategies.   The inter-clip tracker manages global video-level track over time (see Alg. (1)). It saves the tracked object's history for matching and merging subsequent clip prediction. Here, we examine three different feasible implementations for this. First is a two-clip-based matching, which only saves the immediately preceding object track features. The second is a moving-average-based matching, which uses smoothed object prototype embeddings. The third is feature bank based matching, which saves all the object track features up to pre-setted buffer size. As can be shown in Table 4c, we observe that the feature bank approach produces the best performance. This is because it has the broadest temporal view and can directly utilize the raw feature sequence rather than the temporally diluted features.

4.5   Qualitative Results

The visual comparisons of tracking results between the standard tracking scheme and our proposal are in Fig. 4. We select 3 different difficult sequences from the half validation set of MOT17. This include small objects, occlusion, and crowd scene. We see that our approach consistently tracks the objects without identity changes for long-term.

5   Conclusion

In this paper, we point out the fundamental limitations of the current tracking-by-detection scheme. Due to its sequential frame-based matching property, the method suffers from intermediate disturbances in a video and overlooks the temporal context around the target matching frame. As an alternative, we present a new tracking method based on the clip. For the implementation, we define two new tracking operations, intra- and inter-clip tracking. The new formulation views a single long video sequence as multiple short clips. In this way, the intra-clip tracking turns out to be short-term tracking, and we observe that the standard tracking approaches with memory are already competitive in this setup. For the inter-clip tracking, feature-based matching is more stable than IoU-based chaining. We further improve the association quality by designing a novel transformer-based approach, namely Clip Tracker. By connecting together, we show that our proposal achieves new state-of-the-art results on the challenging large vocabulary object tracking benchmark, TAO. We also confirm that our method significantly improves the association quality of the baseline in MOT17. We hope our work opens a new view of the current tracking paradigm.

References

1. Andriyenko, A., Schindler, K.: Multi-target tracking by continuous energy minimization. In: CVPR 2011. pp. 1265–1272. IEEE (2011) 1, 3
2. Andriyenko, A., Schindler, K., Roth, S.: Discrete-continuous optimization for multi-target tracking. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1926–1933. IEEE (2012) 1, 3
3. Athar, A., Mahadevan, S., Osep, A., Leal-Taixé, L., Leibe, B.: Stem-seg: Spatio-temporal embeddings for instance segmentation in videos. In: European Conference on Computer Vision. pp. 158–177. Springer (2020) 4
4. Bergmann, P., Meinhardt, T., Leal-Taixe, L.: Tracking without bells and whistles. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 941–951 (2019) 2
5. Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: the clear mot metrics. EURASIP Journal on Image and Video Processing **2008**, 1–10 (2008) 10
6. 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) 1, 3
7. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020) 4
8. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., et al.: Mmdetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019) 11
9. Chen, L., Ai, H., Zhuang, Z., Shang, C.: Real-time multiple people tracking with deeply learned candidate selection and person re-identification. In: 2018 IEEE international conference on multimedia and expo (ICME). pp. 1–6. IEEE (2018) 6
10. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020) 9
11. Dave, A., Khurana, T., Tokmakov, P., Schmid, C., Ramanan, D.: Tao: A large-scale benchmark for tracking any object. In: ECCV. pp. 436–454. Springer (2020) 3, 6, 10, 11
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 4
13. Feichtenhofer, C., Pinz, A., Zisserman, A.: Detect to track and track to detect. In: Proceedings of the IEEE international conference on computer vision. pp. 3038–3046 (2017) 4
14. Fortmann, T., Bar-Shalom, Y., Scheffe, M.: Sonar tracking of multiple targets using joint probabilistic data association. IEEE journal of Oceanic Engineering **8**(3), 173–184 (1983) 1, 3
15. Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. In: CVPR. pp. 5356–5364 (2019) 10
16. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). vol. 2, pp. 1735–1742. IEEE (2006) 8

17. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020) 9

18. Hu, H.N., Cai, Q.Z., Wang, D., Lin, J., Sun, M., Krahenbuhl, P., Darrell, T., Yu, F.: Joint monocular 3d vehicle detection and tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5390–5399 (2019) 4

19. Hwang, S., Heo, M., Oh, S.W., Kim, S.J.: Video instance segmentation using inter-frame communication transformers. Advances in Neural Information Processing Systems 34 (2021) 4

20. Janai, J., Guney, F., Ranjan, A., Black, M., Geiger, A.: Unsupervised learning of multi-frame optical flow with occlusions. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 690–706 (2018) 4

21. Leal-Taixé, L., Canton-Ferrer, C., Schindler, K.: Learning by tracking: Siamese cnn for robust target association. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 33–40 (2016) 4

22. Leal-Taixé, L., Milan, A., Schindler, K., Cremers, D., Reid, I., Roth, S.: Tracking the trackers: an analysis of the state of the art in multiple object tracking. arXiv:1704.02781 (2017) 3

23. Liang, C., Zhang, Z., Lu, Y., Zhou, X., Li, B., Ye, X., Zou, J.: Rethinking the competition between detection and reid in multi-object tracking. arXiv preprint arXiv:2010.12138 (2020) 2, 3, 6

24. 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) 4

25. 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(2), 548–578 (2021) 10

26. Meinhardt, T., Kirillov, A., Leal-Taixe, L., Feichtenhofer, C.: Trackformer: Multi-object tracking with transformers. arXiv preprint arXiv:2101.02702 (2021) 4

27. Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K.: Mot16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831 (2016) 3, 6, 10

28. 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) 2, 3, 4, 6, 9, 11, 12, 13

29. Park, K., Woo, S., Oh, S.W., Kweon, I.S., Lee, J.Y.: Per-clip video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1352–1361 (2022) 4

30. Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., Tran, D.: Image transformer. In: International Conference on Machine Learning. pp. 4055–4064. PMLR (2018) 4

31. Peng, J., Wang, C., Wan, F., Wu, Y., Wang, Y., Tai, Y., Wang, C., Li, J., Huang, F., Fu, Y.: Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking. In: European conference on computer vision. pp. 145–161. Springer (2020) 3, 6, 7, 12

32. Ramanan, D., Forsyth, D.A.: Finding and tracking people from the bottom up. In: CVPR. vol. 2, pp. II–II. IEEE (2003) 3

33. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016) 4

34. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015) 4

35. Rezatofighi, S.H., Milan, A., Zhang, Z., Shi, Q., Dick, A., Reid, I.: Joint probabilistic data association revisited. In: Proceedings of the IEEE international conference on computer vision. pp. 3047–3055 (2015) 1, 3

36. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: European conference on computer vision. pp. 17–35. Springer (2016) 10

37. Sadeghian, A., Alahi, A., Savarese, S.: Tracking the untrackable: Learning to track multiple cues with long-term dependencies. In: Proceedings of the IEEE international conference on computer vision. pp. 300–311 (2017) 4

38. 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) 11

39. Streit, R.L., Luginbuhl, T.E.: Maximum likelihood method for probabilistic multihypothesis tracking. In: Signal and Data Processing of Small Targets 1994. vol. 2235, pp. 394–405. International Society for Optics and Photonics (1994) 1, 3

40. 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) 4

41. Sun, S., Akhtar, N., Song, H., Mian, A., Shah, M.: Deep affinity network for multiple object tracking. IEEE transactions on pattern analysis and machine intelligence **43**(1), 104–119 (2019) 4

42. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017) 4

43. Wang, Y., Xu, Z., Wang, X., Shen, C., Cheng, B., Shen, H., Xia, H.: End-to-end video instance segmentation with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8741–8750 (2021) 4

44. Wang, Z., Zheng, L., Liu, Y., Li, Y., Wang, S.: Towards real-time multi-object tracking. In: European Conference on Computer Vision. pp. 107–122. Springer (2020) 2, 4

45. Watson, J., Mac Aodha, O., Prisacariu, V., Brostow, G., Firman, M.: The temporal opportunist: Self-supervised multi-frame monocular depth. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1164–1174 (2021) 4

46. 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) 4, 6

47. Wu, J., Cao, J., Song, L., Wang, Y., Yang, M., Yuan, J.: Track to detect and segment: An online multi-object tracker. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12352–12361 (2021) 3, 4, 6

48. Wu, J., Jiang, Y., Zhang, W., Bai, X., Bai, S.: Seqformer: a frustratingly simple model for video instance segmentation. arXiv preprint arXiv:2112.08275 (2021) 4

49. Zeng, F., Dong, B., Wang, T., Zhang, X., Wei, Y.: Motr: End-to-end multiple-object tracking with transformer. arXiv preprint arXiv:2105.03247 (2021) 2, 4

50. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017) 9

51. 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. arXiv preprint arXiv:2110.06864 (2021) 2, 3, 6

52. 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**(11), 3069–3087 (2021) 2, 3, 6

53. Zhang, Z., Wu, J., Zhang, X., Zhang, C.: Multi-target, multi-camera tracking by hierarchical clustering: Recent progress on dukemtmc project. arXiv preprint arXiv:1712.09531 (2017) 3, 6, 7, 12

54. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6881–6890 (2021) 4

55. Zhou, X., Koltun, V., Krähenbühl, P.: Tracking objects as points. In: European Conference on Computer Vision. pp. 474–490. Springer (2020) 3, 4, 6

56. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020) 4