# Rethinking Similar Object Interference in Single Object Tracking

Yipei Wang
School of Instrument Science and Engineering, Southeast University
Institute of Automation, Chinese Academy of Sciences
Nanjing, China

Shiyu Hu
University of Chinese Academy of Sciences
Institute of Automation, Chinese Academy of Sciences
Beijing, China

Xin Zhao*
Institute of Automation, Chinese Academy of Sciences
University of Chinese Academy of Sciences
Beijing, China

## ABSTRACT

Similar object interference (SOI) problem challenges the single object tracking (SOT) task, leading to the failure of feature-based trackers and subsequent performance degradation. Unfortunately, current generic SOT benchmarks do not effectively tackle this critical challenge, while popular SOT algorithms consistently underestimate the influence they have on tracking performance. To bridge this gap and further enhance the investigation of similar object interference in SOT, we adopt the following viewpoints: (1) By examining the operational principles of mainstream trackers and their performance on representative SOT datasets, we redefine similar objects, taking into account the cognitive bias that exists between trackers and humans when dealing with this challenge. (2) Subsequently, we develop a mining methodology that enables the extraction of the SOI sub-dataset from SOT datasets without relying on human intervention. This methodology comprises two main components: determining the SOI challenge and screening the SOI sequences. The SOI dataset is acquired from representative SOT dataset using our proposed approach, known as SOI2023. This dataset serves as an ideal environment to facilitate the investigation of challenges related to similar object interference. (3) Additionally, we conduct extensive tracking experiments with 20 typical trackers and their variants on SOI2023 and analyze their performance for similar object interference scenes in several dimensions. The experimental results demonstrate the effectiveness of our proposed mining method, while revealing the strengths and weaknesses of current trackers when faced with the challenge of similar object interference. We hope this work can provide inspiration to the tracking community and also provide support and insights for robust tracking under the SOI challenge.

## CCS CONCEPTS

• **Computing methodologies → Tracking**.

## KEYWORDS

Visual Tracking, Similar Object Interference, Data Mining, Transformer-based Tracker

*CSAI 2023, December 08–10, 2023, Beijing, China*

ACM ISBN 979-8-4007-0868-8/23/12…$15.00
https://doi.org/10.1145/3638584.3638644

**ACM Reference Format:**
Yipei Wang, Shiyu Hu, and Xin Zhao*. 2023. Rethinking Similar Object Interference in Single Object Tracking. In *2023 7th International Conference on Computer Science and Artificial Intelligence (CSAI) (CSAI 2023), December 08–10, 2023, Beijing, China.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3638584.3638644

## 1 INTRODUCTION

Visual single object tracking is a highly complex research area in computer vision [24], with applications in autonomous driving, robotics, video surveillance, and unmanned aerial vehicles [20, 28]. Recently, the field has witnessed significant progress in tracking algorithm performance, mainly due to advancements in deep learning technology. Regrettably, the performance of trackers is still not at the point where it can be called perfect, and there is still a huge gap in performance on the SOT dataset compared to humans. Researchers subdivide the reasons for these gaps into different challenging factors: occlusion, fast motion, reappearance, *etc*., and have conducted extensive exploration on these challenge factors in terms of relevant benchmark establishment and algorithm improvement. However, one crucial challenging factor, similar object interference (SOI), which is evident in many failure cases (see Fig. 1), has been largely overlooked in research.

Possible limitations can be categorized into several aspects. Firstly, there is a lack of a precise definition of similar objects in the context of tracking. Typically, researchers approach similar objects from a human perspective, nobody really cares what the algorithms see them as. It is also not clear whether training a model to distinguish between similar objects based on human perceptions of them is effective. In addition, the majority of generic SOT datasets do not separate sequences that involve similar object interference challenge, and there have been no specific SOT datasets dedicated to addressing these challenges within the research community. Consequently, the lack of a comprehensive research environment has resulted in a limited focus on similar object interference within trackers. Mainstream algorithm designs primarily concentrate on enhancing the architecture to improve the algorithms' overall tracking capabilities. Only a few works [3, 22], have explored and specifically designed solutions to tackle the SOI challenge.

These issues have motivated us to research and explore the SOI challenge in SOT. To give a reasonable definition of similar objects based on the algorithm's perspective, we first try to find out what they see as similar objects. We conduct an analysis of failure cases in existing trackers using sequences from SOT dataset. Based on this analysis, we propose a new definition of similar objects, which is evident that the tracking algorithm's interpretation of similar objects deviates from traditional human perception.

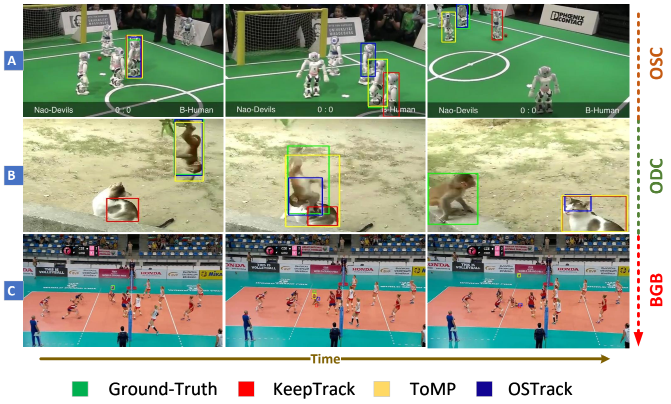

**Figure 1: The representative sequences of SOI2023. These sequences are mined from the generic SOT dataset represented by LaSOT [10]. To visually assess the tracking ability of current algorithms under SOI scenarios, we include the performances of state-of-the-art (SOTA) trackers. A, B, and C illustrate the instances of different classes of similar objects under the perspective of trackers, which are classified into three categories: objects with similar categories (OSC) to the target, objects with different categories (ODC) to the target, and background blocks (BGB) . The vertical arrow represents a decreasing level of semantic similarity, highlighting the disparity between trackers and humans in terms of their ability to recognize similar objects.**

Next, we propose a novel mining method capable of extracting SOI sub-datasets from any SOT datasets, effectively leveraging the existing data. Our mining method comprises two main components: determining the presence of the SOI challenge and screening sequences for SOI. Our method for determining the SOI challenge involves counting the number of target candidates using the confidence score map derived from the trackers' assessment of the search frame image. This approach provides determination results that are solely based on the algorithm's perception of similar objects and do not incorporate any human priors. Our strategy for screening SOI sequences is both simple and effective. Based on the SOI challenge determination results, we identify and compile the sequences that exhibit SOI challenge, creating the SOI sub-dataset of SOT datasets. Furthermore, we further categorize this dataset into subsets based on the frequency of occurrence of the SOI challenge. We firmly believe that this granular subdivision facilitates a deeper understanding of the capabilities and limitations of trackers in the face of the SOI challenge.

The method of SOI challenge determination can selecting any tracking algorithm as the determiner, and our SOI sequences screening strategy can be used on all SOT datasets. Utilizing the proposed mining method, we perform the establishment of the SOI dataset. The trackers selected as determiners include SuperDiMP [6] (based on CNN), OSTrack [31] (based on transformer), and ToMP [21] (based on CNN-transformer), which represent the highest level of cognitive ability of different architecture algorithms for similar objects. Specifically, tracking is a sequential decision-making process where longer sequences will contain richer challenges and set higher demands on algorithms, so we select LaSOT [10], a representative and the largest long-term tracking dataset, as the subject of mining, aiming that the constructed SOI dataset is sufficiently representative and highly challenging. Finally, we obtain a substantial SOI dataset named **SOI2023**, containing a total of 563 SOI sequences. Besides, we make a more detailed division according to the SOI challenge occurrence frequency yields three subsets that reflect the degrees of dominance for SOI challenge in sequences. Then we evaluate existing representative trackers on SOI2023 and analyze their performance for similar object interference scenes in several dimensions.

Contributions of this work can be summarized as follows: (1) We first investigate the challenge posed by similar objects from the perspective of trackers. Furthermore, we propose a novel and effective method for mining SOI sequences to construct the SOI dataset, which avoids the bias and workload caused by manual labeling and screening of sequences in the past, and makes full use of the existing enormous amount of SOT data. (2) We utilize the proposed mining method to construct the first SOI dataset, which contains SOI sequences are mined from the SOT dataset LaSOT [10], and the determiners are all representative trackers. We name it SOI2023. (3) We benchmark 20 recent state-of-the-art tracking approaches and their variants on SOI2023 and analyze their performance in this paper. We carry out extensive experiments to study the impact of SOI challenge on the performance of trackers.

This research aims to inspire the research community and foster the development of trackers that can gradually overcome this challenge in pursuit of real intelligence. We also believe that this generalizable construction of the challenge factor space can be migrated to the study and analysis of other visual tasks and other challenge factors, helping researchers to better utilize existing research resources to conduct efficient studies. We will release the code of our SOI mining method soon, as well as the SOI2023 dataset on https://github.com/updateforever/SOI2023.

## 2 RELATE WORK

### 2.1 SOT benchmarks

Since the first SOT benchmark VOT2013 [17] was proposed, researchers have progressively introduced more influential SOT benchmarks that include larger datasets and standardized metrics.

Initially, due to the limitations of early SOT tasks, the focus of SOT datasets was on short-term tracking. The renowned competition, VOT [16], provided specific keywords to define the SOT task: *single-target*, *model-free*, *causal trackers*, *single-camera*, and *short-term*. These keywords not only differentiate this task from other vision tasks conceptually, but also simplify the initial studies by imposing constraints. These constraints are gradually being lifted as the SOT task develops. Long-term tracking has emerged as the prevailing task in the field of SOT. They involves longer sequences that encompass a wider range of tracking scenes and more complex challenging factors [10, 23]. Additionally, the size of long-term tracking datasets has significantly expanded to accommodate the requirements of data-driven deep learning-based trackers. Recently,

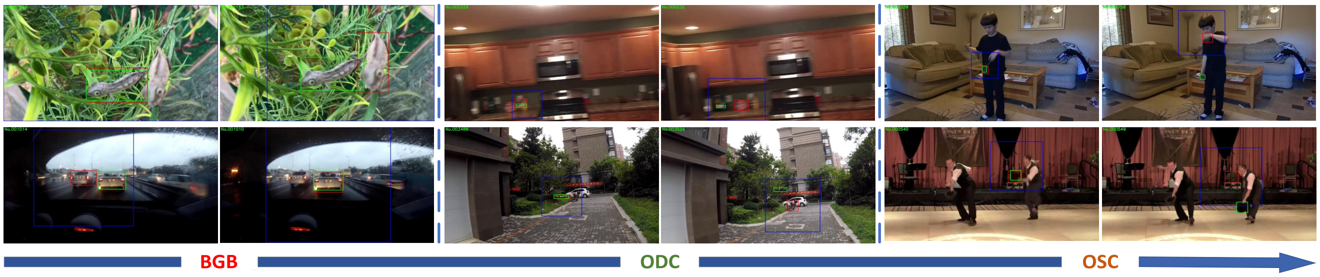

**Figure 2: Examples of possible SOI scenarios (based on OSTrack [31] with the LaSOT [10] benchmark). The 🟦 blue boxes indicate the search region, the 🟥 red boxes show the tracking result, and the 🟩 green boxes represent the ground-truth. The trend indicated by the downward arrow suggests the decreasing semantic level among similar objects in the three scenario groups. This gradual decline reflects the algorithm's diminishing cognitive ability.**

a novel task called global instance tracking (GIT) has been proposed [13]. This task focuses on locating user-specified instances in videos without any assumptions regarding camera or motion consistency. In contrast to traditional SOT methods, GIT eliminates the assumption of a single shot and utilizes datasets with longer sequences that include frequent shot switches.

However, the challenge of similar object interference in SOT has not received sufficient attention within benchmark datasets, as there are limited annotations or statistical results available. TrackingNet [23] introduces the similar object (SOB) attribute, which manually labels similar object information by visually analyzing a dataset comprising 511 videos. This attempt represents the first SOT benchmark for addressing the challenge of similar object interference, but the method used to determine SOB in TrackingNet relies heavily on human judgment, which may not be suitable for evaluating trackers.

## 2.2 SOT methods

The currently dominant SOT methods rely on siamese network architecture, which predicts the target by comparing the correlation between a template image and a search image. These methods have seen continuous development [1, 4, 5, 7]. In terms of model building, these methods can be further classified into CNN-based [1–3, 7, 9, 26], transformer-based [5, 31], and CNN-transformer based [4, 21] network structures. While these methods have demonstrated excellent performance, they struggle to effectively handle the interference caused by similar objects during tracking.

Several studies have proposed effective strategies to address the challenges posed by similar object interference (SOI). For instance, [27] utilizes hand-crafted association scores to link subsequent cross-frame detections and form short traces. Meanwhile, [3] maintains a learnable state that propagates scene information across frames, enabling the tracking of all regions within the scene. Additionally, [22] introduces a learnable network that explicitly and continuously tracks target candidates on a frame-by-frame basis. It is worth noting that these methods have been evaluated and proven effective in handling the SOI challenge, despite the absence of a dedicated SOI dataset. Instead, they have been directly tested on generic SOT datasets. Furthermore, the techniques developed to address similar object interference also offer valuable insights and assistance in the development of methods for generic scene tracking.

## 3 SIMILAR OBJECT INTERFERENCE UNDER TRACKERS' COGNITIVE LEVEL

This section describes our study of the definition of similar objects.

Firstly, we aim to investigate the robustness of current trackers regarding the interference from similar objects, given the continuous advancements in research. Among the transformer-based algorithms, OSTrack [31] demonstrates the best performance on generic tracking benchmarks, making it an ideal choice to represent the upper bound of tracking algorithm performance. We execute OSTrack on the widely recognized LaSOT dataset [10], saving and visually presenting its tracking results. Here, we adhere to the widely accepted assumption that target discrimination relies on the confidence score map generated by the tracker's final output. According to this assumption, we consider the presence of the target in a frame whenever confidence scores surpass a certain threshold (typically set at 0.25), selecting the position with the highest confidence score to predict target information [1]. Consequently, we analyze the failure cases with challenges posed by similar object interference in OSTrack, as depicted in Fig. 2.

Our initial conclusion reveals that appearance-based trackers have different interpretations of similar objects when compared to humans. Our findings demonstrate that even the most effective algorithms are prone to tracking drift caused by interference from similar objects, whereas such errors are nearly nonexistent in human perception. Additionally, similar objects often yield high confidence scores, indicating that the algorithm fails to recognize when the target disappears and lacks the capability to distinguish between similar objects. These instances of failure illustrate that current trackers lack advanced cognitive abilities. They rely solely on appearance features for matching templates and search images to predict tracking targets.

Based on the actual performance of trackers, we provide a definition of similar objects that aligns with the cognitive capabilities of the algorithms. In the current search region, similar objects to

[1] Without special notes, all of our subsequent studies are based on this assumption

the target can be categorized into three groups based on their semantic characteristics and appearance feature information: (1) the other objects of the same category as target (OSC, such as in Fig.2-Left); (2) the objects of different categories with similar appearance features to target (ODC, such as in Fig.1-Middle); and (3) the background blocks with similar appearance features (BGB, such as in Fig.1-Right), which have no specific category and even do not meet the definition of objects. Among these categories, objects belonging to OSC have a higher semantic level and align with human cognition. Objects in the ODC and BGB categories differentiate the algorithm's performance from human perception. The former has a lower semantic level compared to OSC, while the latter lacks semantic information and relies solely on apparent features.

As soon as the cognitive level of algorithms for similar objects is clarified, it is a crucial step to build a SOI dataset that can satisfy the relevant algorithm's evaluation review. So we propose a novel approach for mining SOI datasets which will elaborate in next section. As we observe this is the first attempt at building the SOI benchmark, and our hope is to lead the community to better assess the cognitive capabilities of current trackers and to construct relevant SOI datasets based on their existing cognitive stages - in order to contribute to the evaluation and improvement of trackers' performance.

# 4 SOI SUB-DATASET MINING METHOD

Based on the cognitive level of trackers, we design a novel mining method for SOI sub-dataset, and the specific process of the mining method is shown in Fig. 3. It consists of two main components: the SOI challenge determination method and the SOI sequence screening strategy, which we describe in detail below. We also apply our SOI mining approach on the representative dataset LaSOT [10] to construct a appropriate dataset known as SOI2023.

## 4.1 SOI determination based on the confidence score map of trackers

We propose a novel method for determining SOI based on the confidence score map generated by the trackers. This method enables us to capture the trackers' perception of similar objects.

The confidence scores predicted by the siamese network tracking method provide a reliable reference for assessing appearance similarity. Higher confidence scores indicate a stronger resemblance between the tracked objects and the target. In general, trackers select the index of the highest score point from the score map of the current frame to map the center coordinate of target at a coarse-grained level, and predict the final target state via bounding box regression branch.

As one of the requirements, the confidence score map $\hat{S}^t \in \mathbb{R}^{h \times w}$ of the determiner tracker on frame $I_t$ is given, where $(h, w)$ denote the size of confidence score matrix. We use maximum pooling operation to obtain local maxima score on confidence score map:

$$\tilde{S}_c^t = MaxPool(\hat{S}^t) \tag{1}$$

where $MaxPool$ is a maximal pooling whose kernel size is set to $(5, 5)$ and the padding is set to 2. This operation extracts the local maximum scores from the confidence score map while simultaneously applying an approximate non-maximum suppression (NMS) technique to address the issue of duplicate candidate determination arising from large target sizes. $\tilde{S}_c^t = \{\tilde{S}_{c1}^t, \ldots, \tilde{S}_{cm}^t\}$ is the set of target candidates for $I_t$. Next we filter out the target candidates with insufficient confidence scores:

$$S_c^t = \{\tilde{S}_{ci}^t \mid \tilde{S}_{ci}^t \geq \alpha \cdot \tilde{S}_{max}^t and \tilde{S}_{ci}^t \geq \eta\} \tag{2}$$

where $\tilde{S}_{max}^t = max(\tilde{S}_c^t)$, $\alpha$ is a hyperparameter for the confidence score threshold and $\eta$ is the threshold hyperparameter for the target not found.

The set $S_c^t = \{S_{c1}^t, \ldots, S_{cn}^t\}$ represents the presence of similar objects, denoting that they may cause interference on the image $I_t$. Clearly, if the size of $S_c^t$ exceeds 1, it indicates the existence of a challenge posed by similar object interference.

## 4.2 Screening strategy for SOI sequences

We propose a systematic and stringent sequence screening strategy for the generic SOT datasets, which relies on the outcome of the SOI determination method detailed in Section 4.1. This strategy can be summarized as follows:

First, we execute $m$ determiners on the SOT dataset to generate predicted confidence score maps for each frame in all sequences. Then, we employ the proposed SOI determination method to obtain the result $S_{cj}^t = \{S_{c1}^t, \ldots, S_{cm}^t\}$. Our sequence screening process consists of two distinct stages of judgment:

**Whether the sequence has SOI challenges.** If the length of $S_{cj}^t$ exceeds 1, it indicates that the associated determiner acknowledges the presence of a SOI in frame $I_t$. If more than $\beta = \lfloor (2/3)m \rfloor$ determiners affirm this, frame $I_t$ is designated as an SOI frame and added to the set of SOI frames for this sequence, denoted as $I_{soi}$.

$$I_{soi} = \{\{I_{ti}\}_1^k \mid I_{ti} \notin I_{absent}\} \tag{3}$$

where we filter out the SOI determination results of the frames with target absence $I_{absent}$ , which adheres to the current tracking benchmark's evaluation principles. $I_{soi} = \{I_{t1}, \ldots, I_{tk}\}$ is obtained after judging all frames. If $k > 0$, the sequence exists SOI challenges and is divided into the SOI sub-dataset $L_{Total}$. On the contrary it will be removed.

**Further division based on the SOI challenge occurrence frequency.** In order to extract more comprehensive information regarding SOI challenges and gain insights into their impact on tracking, we conduct a more detailed division based on the occurrence frequency of SOI challenge in sequences. First, we read the SOI frame set, denoted as $I_{soi}$, for the sequence within the SOI sub-dataset. We calculate the number of occurrences of SOI challenges by using a standard interval of $T = 10$. A new occurrence of a SOI challenge is identified if the interval between the previous and current SOI frames exceeds 10. Consequently, we obtain the count of SOI challenge occurrences $T_f$ for the sequence, which we refer to as the SOI challenge occurrence frequency. We observe that the occurrence frequency of SOI challenges in large-scale sequences follows a long-tailed distribution trend (see Fig. 4). To better analyze the impact of the dominance degree of SOI challenges on tracking, we divide the SOI sequences into three sets $L_{Once}$, $L_{More}$, and $L_{Most}$, which represent sequences with only one occurrence of SOI challenge, multiple occurrences of SOI challenge, and a dominance

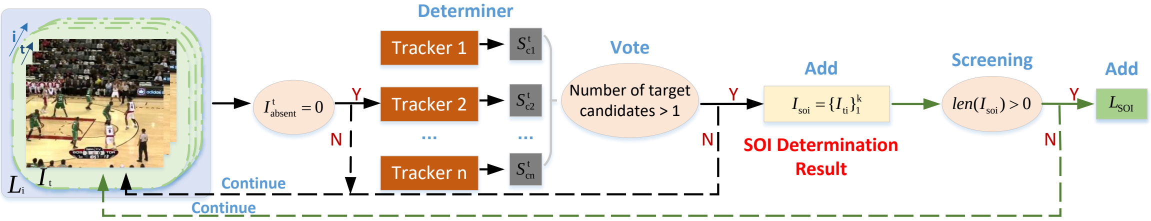

**Figure 3: The overall process of mining SOI sub-dataset from the SOT dataset, where the black arrow indicates the process of SOI challenge determination and the green arrow indicate the process of SOI sequence screening. Our mining method involves traversing all frames of the SOT dataset and employing multiple trackers to vote on the existence of the SOI challenge in each frame. Frames identified as containing the SOI challenge are saved in the set $I_{soi}$. Based on the results of the SOI determination, the corresponding SOI sequences are selected and included in the SOI sub-dataset set $L_{SOI}$.**

of SOI challenge, respectively:

$$\begin{cases} L_{Once} = \{L_i \mid T_{f_i} = 1\} \\ L_{More} = \{L_i \mid T_{f_i} \in (1, 10]\} \\ L_{Most} = \{L_i \mid T_{f_i} > 10\} \end{cases} \tag{4}$$

Obviously, $L_{Total} = L_{Once} + L_{More} + L_{Most}$.

We posit that the SOI sub-datasets derived from mining the generic SOT datasets hold value for the advancement of the SOI challenge. To better spearhead the relevant research, we construct a first SOI dataset using the proposed mining method, which is detailed in section 4.3.

## 4.3 Construction of the SOI dataset

*4.3.1 Selection of determiners and generic SOT dataset.* Our SOI challenge determination method can be automatically executed by any tracker, while our SOI sequences screening strategy is applicable to a wide range of generic SOT datasets.

The former relies on the confidence score maps of trackers, allowing for flexible determination of the number of determiners. This approach aligns with the conventional manual attribute determination standard, where multiple professionals annotate challenge attributes to ensure accuracy [13, 15]. Striking a balance between normality and efficiency, we opt for three determiners to identify the SOI challenge: SuperDiMP [6] is a CNN and correlation filter based tracker, ToMP [21] introduces the transformer into the CNN network to construct a better model predictor, and for the transformer-based model we select OSTrack [31], which is currently the most popular, to be representative. These trackers capture various architecture-based trackers in dealing with similar object interference and provide evidential support for the accuracy and validity of SOI challenge determinations. More details about them will be supplemented in Section 5.1.1

Given the variations in application scenarios and focus factors across SOT benchmarks, we specifically choose to utilize LaSOT [10] as the representative SOT dataset. As one of the most influential long-term tracking benchmarks in the tracking community, it provides a large-scale new benchmark for SOT. It encompasses 1400 sequences, averaging 2500 frames per sequence. It incorporates over 3 million manually labeled bounding box annotations, meticulously inspected, and covers a diverse range of 70 categories.

**Table 1: The results of SOI sequences screening in LaSOT**

| | soi sequences | original sequences |
|---|---|---|
| train | 433 | 1120 |
| test | 130 | 280 |

*4.3.2 A representative SOI dataset: SOI2023.* We perform SOI determination using the selected trackers on LaSOT [10], and then perform the SOI sequences screening work [2], and the SOI sequence screening results shown in Tab. 1. LaSOT covers all scenarios of generic target tracking and represents a majority of SOT benchmarks, which partitions 1400 sequences into train and test sets. Following its established division standard we conduct our mining work and get the SOI2023, which also includes both train and test sets. It is evident that the long sequences in the long-term tracking dataset encompass more complex tracking scenarios, resulting in a high number of SOI sequences. Evidently, the occurrence of SOI challenges also relates to the degree of difficulty in tracking.

Additionally, we calculated the occurrence frequency of SOI challenges in these sequences (refer to Fig. 4). Evidently, the division results show a Long-Tail distribution trend when considering the occurrence frequency attribute of SOI challenges. To assess the performance of trackers under different levels of dominant SOI challenge, we also categorize the SOI sequences into more detailed groups ($L_{Once}$, $L_{More}$, and $L_{Most}$).

Overall, SOI2023 consists of 433 sequences for the train set and 130 sequences for the test set. In Section 5, we will evaluate the performance of representative trackers on the SOI2023 dataset. Additionally, we will discuss and analyze the algorithms' performance from multiple perspectives.

[2] Trackers are all run according to the settings and parameters provided in the original paper and strictly follow the original description and requirements of dataset.

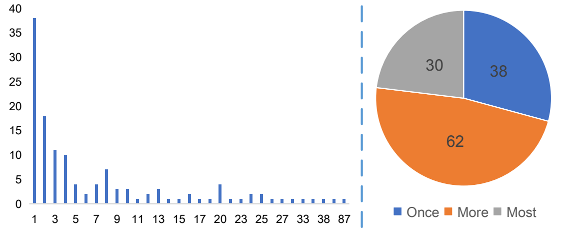

**Figure 4: The comparison results of the SOI dominance degree of sequences in test set. The left is the statistical results of SOI challenge occurrence frequency, and the right is the results of sequence division based on the attributes of the SOI challenge occurrence frequency.**

# 5 EXPERIMENT

## 5.1 Evaluation and Analysis on SOI2023

In this section, we conduct comprehensive experiments using established trackers on SOI2023. These tracking results not only validate the rationale behind our proposed SOI challenge determination method and SOI sequence screening strategy, but also offer valuable guidance for future studies in this domain.

**Table 2: Overall tracking results of representative trackers on SOI2023 and LaSOT [10]. The trackers are ranked by their accuracy ( AUC ) scores on SOI2023, with precision (PRE) and normalized precision (NPRE) scores presented. The Properties column denotes the feature representation of the different trackers ( CNN - Convolutional Neural Network, HOG - Histogram of Gradients, TRANS - Transform Network).**

| Tracker | LaSOT | SOI2023 | | | Venue | Properties |
|---|---|---|---|---|---|---|
| | AUC | AUC | PRE | NPRE | | |
| KCF [12] | 0.211 | 0.139 | 0.146 | 0.153 | TPAMI'15 | HOG |
| SiamFC [1] | 0.336 | 0.232 | 0.259 | 0.275 | ECCV'16 | CNN |
| SiamDW [32] | 0.356 | 0.273 | 0.285 | 0.323 | CVPR'19 | CNN |
| ECO [8] | 0.324 | 0.283 | 0.296 | 0.313 | CVPR'17 | HOG,CNN |
| SiamFC++ [30] | 0.544 | 0.352 | 0.349 | 0.365 | AAAI'20 | CNN |
| SiamRPN [19] | 0.475 | 0.355 | 0.352 | 0.382 | CVPR'18 | CNN |
| DaSiamRPN [34] | 0.415 | 0.360 | 0.361 | 0.387 | CVPR'18 | CNN |
| ATOM [7] | 0.515 | 0.380 | 0.383 | 0.394 | CVPR'19 | CNN |
| Ocean [33] | 0.560 | 0.399 | 0.395 | 0.420 | ECCV'20 | CNN |
| KYS [3] | 0.554 | 0.418 | 0.414 | 0.427 | ECCV'20 | CNN |
| GlobalTrack [14] | 0.521 | 0.420 | 0.428 | 0.440 | AAAI'20 | CNN |
| DiMP [2] | 0.569 | 0.443 | 0.447 | 0.456 | ICCV'19 | CNN |
| PrDiMP [9] | 0.598 | 0.481 | 0.490 | 0.502 | CVPR'20 | CNN |
| SiamRCNN [26] | 0.648 | 0.506 | 0.523 | 0.520 | CVPR'20 | CNN |
| SuperDiMP [6] | 0.631 | 0.522 | 0.543 | 0.542 | CVPR'20 | CNN |
| TransT [4] | 0.649 | 0.529 | 0.568 | 0.552 | CVPR'21 | TRANS |
| MixFormer [5] | 0.692 | 0.570 | 0.606 | 0.590 | CVPR'22 | TRANS |
| OSTrack [31] | 0.691 | 0.572 | 0.611 | 0.591 | ECCV'22 | TRANS |
| ToMP [21] | 0.676 | 0.572 | 0.62 | 0.600 | CVPR'22 | CNN,TRANS |
| KeepTrack [22] | 0.671 | 0.594 | 0.637 | 0.624 | ICCV'21 | CNN |

*5.1.1 Single Object Tracking Methods.* In this work, we select recent transformer and convolutional neural networks (CNN) based trackers that have demonstrated outstanding performance across various benchmarks and challenges. Additionally, to ensure comprehensiveness, we have also included correlation filter based trackers in our experiments. Tab. 2 shows 20 representing SOT algorithms covering both classic and SOTA methods. Brief descriptions of these methods are provided below.

(1) KCF [12] is a classical Correlation Filter-based method that achieves a balance between tracking accuracy and high speed. ECO [8] is the first attempt to fuse Convolutional Neural Networks (CNN) with Correlation Filter (CF) methods. SiamFC [1] pioneered the concept of the Siamese Neural Network (SNN) based tracker, which delivers satisfactory tracking performance via a straightforward network structure for feature matching between the template region and the search region. (2) Subsequently, SiamRPN [19] utilizes the region proposal network [25] to attain precise target regression. DaSiamRPN [34] employs data augmentation techniques to enhance discrimination. On the other hand, SiamRPN++ [18] and SiamDW [32] introduce deeper and wider backbone networks based on ResNet [11] for enhanced feature extraction. SiamFC++ [30] and Ocean [33] adopt an anchor-free architecture to mitigate the complexity associated with anchors. (3) GlobalTrack [14] assumes the absence of motion consistency and conducts a comprehensive image search to mitigate cumulative errors. Additionally, SiamRCNN [26] develops a robust re-detection mechanism utilizing FasterRCNN [25]. (4) ATOM [7] combines CF and SNN to design a new tracking framework. Building upon this framework, DiMP [2] enhances discriminative power by optimizing the loss function. Additionally, PrDiMP [9] and SuperDiMP [6] employ probabilistic regression techniques to further enhance tracking accuracy. KYS [3] incorporates scene information and merges it with the appearance model to localize objects. KeepTrack [22] establishes a candidate association network to handle similar object interference, which is trained through the mining of challenging sequences from LaSOT [10]. (5) Recently, transformer-based trackers have developed quickly. TransT [4] leverages the global attention mechanism to reconstruct features extracted from the backbone network and enhance tracking performance. MixFormer [5] introduces an end-to-end converter-based framework that facilitates parallel feature extraction and integration of target information. ToMP [21] incorporates transformers into correlation filtering operations to generate more robust model weights, resulting in improved accuracy of tracking results. OSTrack [31] devises a novel one-stream tracking pipeline that conducts feature extraction and association modeling simultaneously. Additionally, it introduces an early candidate elimination method to remove irrelevant attention information, thereby enhancing the model's speed.

**Notes.** Due to the extensive workload, we conduct experiments using the public code of these trackers without any modifications. It is important to note that the validation results may contain some inaccuracies due to variations in code environments, as well as differences in hardware and software configurations on different machines. We have conducted a comprehensive evaluation using one pass evaluation (OPE) and measured the precision, normalized precision, and success ratio of diverse trackers using well-established tracking protocols [10, 29]. For all experiments, we employed a server equipped with a 56 core Intel(R) Xeon(R) 2.0GHz CPU and 4 GeForce GTX TITAN X graphic cards.

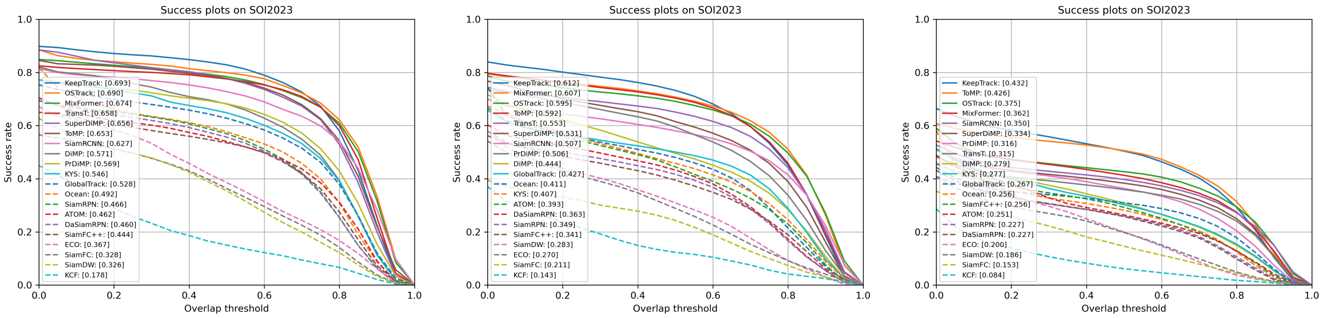

**Figure 5: The performance of trackers under different SOI challenge dominance degrees, represented by success rate. (a) to (c) contain sequences obtained by dividing them according to the occurrence frequency of SOI challenges.**

*5.1.2 Overall Performance.* Tab. 2 present the overall performance of trackers in OPE mechanism. It can be seen that all these trackers perform worse on SOI2023 than on the original generic SOT dataset LaSOT [10]. Due to the presence of numerous SOI challenges in SOI2023, the algorithm scores tend to be low. Most algorithms heavily depend on target appearance information and typically utilize the template frame and the previous frame to aid in target localization. Unfortunately, this approach often leads to poor performance when similar objects are present. Meanwhile, KeepTrack [22] demonstrates superior performance compared to recent trackers such as MixFormer [5], ToMP [21], and OSTrack [31], establishing a state-of-the-art (SOTA) result on SOI2023. It proves that the ability of KeepTrack to cope with the SOI challenge is very powerful, and on the other hand, it indicates that our SOI mining approach is reasonable and effective.

*5.1.3 Attribute Performance.* The SOI2023 dataset categorizes sequences into subsets $L_{Once}$, $L_{More}$, and $L_{Most}$ based on three attributes that signify the level of dominance of the SOI challenge. Fig. 5 presents the detailed results of the trackers on these subsets, we discuss here the success rate metric as a proxy for the fact that all algorithms have the same experimental performance in terms of precision and normalized precision. When the SOI challenge occurs only once, all trackers perform close to the performance gained on the generic SOT task. With the increase in the number of SOI challenge occurrence, the performance of all algorithms deteriorates. Notably, KeepTrack [22], which is specifically designed for handling the SOI challenge, achieves the most favorable results.

Clearly, as the occurrence frequency of SOI challenge in sequences increases, trackers face a higher demand to distinguish similar objects and encounter more formidable challenges. Additionally, the outcomes on the subsets, obtained through attribute division, also provide a more straightforward demonstration of the efficacy and rationality of our proposed methods for determining SOI and screening sequences.

## 5.2 Visual Analysis

We present here some examples of tracking result visualisations (Fig. 6). Unlike humans, the mainstream trackers rely solely on the appearance features of the target during tracking, without employing advanced semantic inference to assess if the tracked object is still the original target.

Also, we have observed that cases of trackers mistakenly following the wrong target solely due to the presence of similar objects (Fig.6-A) are rare. Instead, tracking failures often occur as a result of combined challenges and the presence of SOI. When only similar objects appear during tracking, trackers can typically maintain consistent tracking performance. Despite high confidence scores for both the target and the distractor, the tracking results do not deteriorate. However, the presence of other challenging factors, such as low resolution (Fig.6-B), occlusion (Fig.6-C), fast motion (Fig.6-D), *etc.*, significantly impacts tracking performance. These challenges cause a rapid decrease in the confidence score of the target, leading to the identification of the distractor with high confidence as the target. Consequently, tracking drift or failure occurs. It is noteworthy that trackers tend to continue learning information about distractors even after tracking the wrong target. This makes it challenging for trackers to correct their mistakes, resulting in consecutive tracking failures.

Currently, numerous researchers focus on enhancing the feature extraction capability of trackers to extract more robust target appearance features. However, this approach offers minimal improvement to the cognitive capabilities of the tracker. We advocate that as the challenge of tracking SOI with low-level semantic information has been progressively addressed, future research should focus on enhancing the cognitive capabilities of trackers.

# 6 CONCLUSION

This paper presents a novel approach to addressing the challenge of similar object interference in SOT. We redefine similar objects based on the cognitive capabilities of current algorithms, and then develop a novelty SOI mining method to construct a representative SOI dataset called SOI2023. Finally, we evaluate 20 representative methods using comprehensive evaluation mechanisms and metrics specific to SOI2023.

Notably, our definition of similar objects and the establishment of the SOI dataset SOI2023, based on our proposed SOI mining method, are designed to be dynamically updated. This allows them to be adjusted as trackers continue to improve in performance and

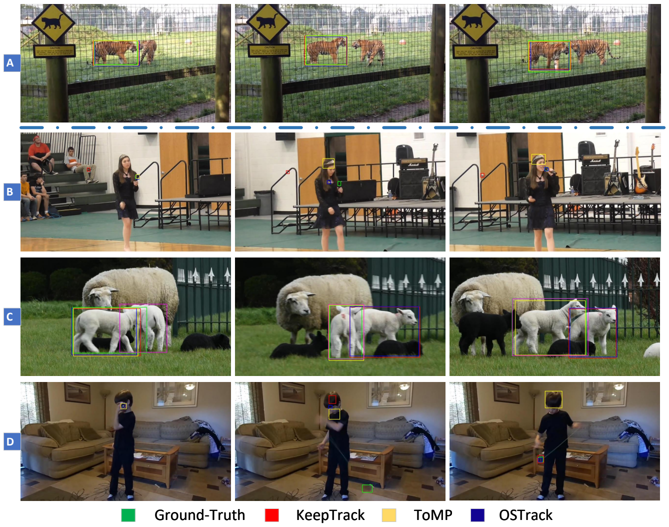

**Figure 6: Visualization results of representative trackers on SOI2023, including both good case (A) and bad cases (B-D).**

cognitive abilities. Furthermore, these updates serve to support the ongoing research and development of trackers.

We have found that the cognitive level of existing trackers is limited, resulting in poor tracking performance on SOI sequences. Going forward, we anticipate that our proposed SOI mining method and the SOI2023 dataset will serve as a valuable resource for future research. Additionally, our analysis of the cognitive level of algorithms and the principles of algorithmic intelligence development is not only applicable to the SOI challenge in tracking, but can also be relevant and valid in other research domains.

## REFERENCES

[1] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. 2016. Fully-convolutional siamese networks for object tracking. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II 14*. Springer, 850–865.

[2] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. 2019. Learning discriminative model prediction for tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6182–6191.

[3] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. 2020. Know your surroundings: Exploiting scene information for object tracking. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*. Springer, 205–221.

[4] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. 2021. Transformer tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8126–8135.

[5] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. 2022. Mixformer: End-to-end tracking with iterative mixed attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13608–13618.

[6] Martin Danelljan and Goutam Bhat. 2019. PyTracking: Visual tracking library based on PyTorch.

[7] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. 2019. Atom: Accurate tracking by overlap maximization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4660–4669.

[8] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. 2017. Eco: Efficient convolution operators for tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6638–6646.

[9] Martin Danelljan, Luc Van Gool, and Radu Timofte. 2020. Probabilistic regression for visual tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7183–7192.

[10] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. 2019. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5374–5383.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[12] João F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. 2014. High-speed tracking with kernelized correlation filters. *IEEE transactions on pattern analysis and machine intelligence* 37, 3 (2014), 583–596.

[13] Shiyu Hu, Xin Zhao, Lianghua Huang, and Kaiqi Huang. 2022. Global Instance Tracking: Locating Target More Like Humans. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 1 (2022), 576–592.

[14] Lianghua Huang, Xin Zhao, and Kaiqi Huang. 2019. GlobalTrack: A Simple and Strong Baseline for Long-term Tracking. *arXiv: Computer Vision and Pattern Recognition* (2019).

[15] Lianghua Huang, Xin Zhao, and Kaiqi Huang. 2019. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE transactions on pattern analysis and machine intelligence* 43, 5 (2019), 1562–1577.

[16] Matej Kristan, Jiri Matas, Aleš Leonardis, Tomáš Vojíř, Roman Pflugfelder, Gustavo Fernandez, Georg Nebehay, Fatih Porikli, and Luka Čehovin. 2016. A novel performance evaluation methodology for single-target trackers. *IEEE transactions on pattern analysis and machine intelligence* 38, 11 (2016), 2137–2155.

[17] Matej Kristan, Roman Pflugfelder, Ales Leonardis, Jiri Matas, Fatih Porikli, Luka Cehovin, Georg Nebehay, Gustavo Fernandez, Tomas Vojir, et al. 2014. The vot2013 challenge: overview and additional results. (2014).

[18] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. 2018. SiamRPN++: Evolution of Siamese Visual Tracking with Very Deep Networks. *CoRR* abs/1812.11703 (2018). arXiv:1812.11703 http://arxiv.org/abs/1812.11703

[19] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. 2018. High Performance Visual Tracking With Siamese Region Proposal Network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[20] Seyed Mojtaba Marvasti-Zadeh, Li Cheng, Hossein Ghanei-Yakhdan, and Shohreh Kasaei. 2021. Deep learning for visual tracking: A comprehensive survey. *IEEE Transactions on Intelligent Transportation Systems* (2021).

[21] Christoph Mayer, Martin Danelljan, Goutam Bhat, Matthieu Paul, Danda Pani Paudel, Fisher Yu, and Luc Van Gool. 2022. Transforming model prediction for tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8731–8740.

[22] Christoph Mayer, Martin Danelljan, Danda Pani Paudel, and Luc Van Gool. 2021. Learning target candidate association to keep track of what not to track. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13444–13454.

[23] Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. 2018. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European conference on computer vision (ECCV)*. 300–317.

[24] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[25] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015).

[26] Paul Voigtlaender, Jonathon Luiten, Philip H.S. Torr, and Bastian Leibe. 2020. Siam R-CNN: Visual Tracking by Re-Detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6578–6588.

[27] Paul Voigtlaender, Jonathon Luiten, Philip HS Torr, and Bastian Leibe. 2020. Siam r-cnn: Visual tracking by re-detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6578–6588.

[28] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. 2013. Online object tracking: A benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2411–2418.

[29] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. 2013. Online Object Tracking: A Benchmark. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*. 2411–2418.

[30] Yinda Xu, Zeyu Wang, Zuoxin Li, Ye Yuan, and Gang Yu. 2020. SiamFC++: Towards Robust and Accurate Visual Tracking with Target Estimation Guidelines. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 7 (2020), 12549–12556.

[31] Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. 2022. Joint feature learning and relation modeling for tracking: A one-stream framework. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*. Springer, 341–357.

[32] Zhipeng Zhang and Houwen Peng. 2019. Deeper and Wider Siamese Networks for Real-Time Visual Tracking. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4591–4600.

[33] Zhipeng Zhang and Houwen Peng. 2020. Ocean: Object-aware Anchor-free Tracking. In *ECCV (21)*. 771–787.

[34] Zheng Zhu, Qiang Wang, Bo Li, Wei Wu, Junjie Yan, and Weiming Hu. 2018. Distractor-aware Siamese Networks for Visual Object Tracking. *CoRR* abs/1808.06048 (2018). arXiv:1808.06048 http://arxiv.org/abs/1808.06048