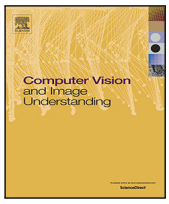
Computer Vision and Image Understanding 262 (2025) 104553

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu



PlanarTrack: A high-quality and challenging benchmark for large-scale planar object tracking

Yifan Jiao [a,b], Xinran Liu [a,b], Xiaoqiong Liu [c], Xiaohui Yuan [c], Heng Fan [c], Libo Zhang [a,b,*]

[a] Institute of Software, Chinese Academy of Sciences, Beijing, China
[b] University of Chinese Academy of Sciences, Beijing, China
[c] Department of Computer Science & Engineering, University of North Texas, Denton, United States of America

ARTICLE INFO

Communicated by Juergen Gall

Keywords:
Planar object tracking
Large-scale benchmark
High-quality annotation
Tracking evaluation

ABSTRACT

Planar tracking has drawn increasing interest owing to its key roles in robotics and augmented reality. Despite recent great advancement, further development of planar tracking, particularly in the deep learning era, is largely limited compared to generic tracking due to the lack of large-scale platforms. To mitigate this, we propose **PlanarTrack**, a large-scale high-quality and challenging benchmark for planar tracking. Specifically, PlanarTrack consists of 1150 sequences with over 733K frames, including 1000 short-term and 150 new long-term videos, which enables comprehensive evaluation of short- and long-term tracking performance. All videos in PlanarTrack are recorded in unconstrained conditions from the wild, which makes PlanarTrack challenging but more realistic for real-world applications. To ensure high-quality annotations, each video frame is manually annotated by four corner points with multi-round meticulous inspection and refinement. To enhance target diversity of PlanarTrack, we only capture a unique target in one sequence, which is different from existing benchmarks. To our best knowledge, PlanarTrack is by far the largest and most diverse and challenging dataset dedicated to planar tracking. To understand performance of existing methods on PlanarTrack and to provide a comparison for future research, we evaluate 10 representative planar trackers with extensive comparison and in-depth analysis. Our evaluation reveals that, unsurprisingly, the top planar trackers heavily degrade on the challenging PlanarTrack, which indicates more efforts are required for improving planar tracking. Moreover, we derive a variant named **PlanarTrack$_{BB}$** from PlanarTrack for generic tracking. Evaluation with 15 generic trackers shows that, surprisingly, our PlanarTrack$_{BB}$ is even more challenging than several popular generic tracking benchmarks, and more attention should be paid to dealing with planar targets, though they are rigid. Our data and results will be released at https://github.com/HengLan/PlanarTrack

1. Introduction

Planar object tracking is a fundamental problem in computer vision. Different from generic object tracking which aims at localizing the target with axis-aligned rectangle bounding boxes (Wu et al., 2013; Huang et al., 2019; Fan et al., 2019), the goal of planar object tracking is to predict the 2D transformations (*e.g.*, the homograph) of a target (*e.g.*, surface or plane of the object) and locate it with four corner points (see Fig. 1). Because of its important applications in augmented reality (AR) (*e.g.*, Comport et al., 2003; Wagner et al., 2009; Matveichev and Lin, 2021) and robotics (*e.g.*, Mondragón et al., 2010; Corso et al., 2003), planar object tracking has attracted increasing interest in recent years. Particularly, with the introduction of several benchmarks (*e.g.*, Liang et al., 2018, 2021; Roy et al., 2015), great progress has been seen in planar object tracking (*e.g.*, Zhan et al., 2022; Zhang and Ling, 2022; Šerỳch and Matas, 2023; Li et al., 2023). Despite this, these datasets are largely limited in further facilitating the development of planar object tracking, due to the following reasons:

Small-scale. One major issue with existing benchmarks is their relatively small scales. Especially, in the deep learning era, in order to unleash the potential of deep planar tracking, a large-scale platform with a great number of video sequences is highly desired for training. As demonstrated in Fig. 2, however, all existing datasets comprise *less than* 300 video sequences, which is far from being sufficient for training deep planar trackers. As a result, researchers in the community have to utilize synthetic data generated from images (*e.g.*, Lin et al., 2014) or videos from the generic bounding box-based tracking benchmark (*e.g.*, Huang et al., 2019) for deep planar tracking, which may result in suboptimal performance because of domain gap among different tasks. In addition to the training of deep planar trackers, a

* Corresponding author at: Institute of Software, Chinese Academy of Sciences, Beijing, China.
E-mail address: libo@iscas.ac.cn (L. Zhang).

https://doi.org/10.1016/j.cviu.2025.104553
Received 7 May 2025; Received in revised form 12 October 2025; Accepted 20 October 2025
Available online 24 October 2025

(a) Example of *generic object tracking* with rectangular bounding box

(b) Example of *planar object tracking* with corner points

Fig. 1. Comparison between generic object tracking (a) and planar object tracking (b). The former estimates axis-aligned rectangular bounding boxes for the target object, while the latter (our focus in this work) calculates 2D transformations of the target object to obtain the corresponding corner points for localization. All figures throughout this paper are best viewed in color and by zooming in. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

large-scale platform is necessary for reliable evaluation and comparison of different algorithms.

Less challenging scenario. Real-world scenarios are often challenging and complicated. Nevertheless, early planar tracking datasets (*e.g.*, Lieberknecht et al., 2009; Roy et al., 2015; Gauglitz et al., 2011; Chen et al., 2017) are developed from indoor laboratory environments with simple background, which cannot fully reflect the complicated and diverse scenarios in real applications while evaluating. To handle this, recent datasets (*e.g.*, Liang et al., 2018, 2021) directly collect videos in the wild. However, most sequences in these benchmarks are mainly involved with one challenge factor (or *attribute* in generic tracking), and very few (*e.g.*, 30 videos in Liang et al. (2018) and 40 videos in Liang et al. (2021)) contain multiple challenges (*i.e.*, the unconstrained condition). This may weaken the difficulties of planar tracking in the wild where arbitrary challenges could occur simultaneously, and thus restricts their usage in evaluating the generalization of planar tracking systems in the real world.

Limited diversity. The diversity of target objects is crucial for a tracking benchmark. In existing planar tracking datasets, the sample planar target is often utilized in multiple sequences, which largely reduces the diversity in target appearance and may lead to bias in performance assessment. For example, for the current largest planar tracking benchmark (Liang et al., 2021) (one target used in 7 videos), the number of planar targets does not exceed 40 (see Table 1). Such lack of diversity makes it difficult to use the current benchmarks for faithful evaluation of planar trackers in practice.

Lack of long-term tracking. The task of long-term tracking is more challenging and holds greater practical significance compared to short-term tracking. This is because long-term tracking requires algorithms capable of continuously capturing the target object over extended durations, while effectively handling scenarios wherein the target frequently disappears and reappears. This complexity makes long-term tracking tasks more reflective of real-world applications. In order to be deployed in real applications, a planar tracker is expected to perform well in not only short-term scenarios but also in long-term videos. Yet, existing benchmarks either contain only short-term videos (*e.g.*, Gauglitz et al., 2011; Liang et al., 2018, 2021) with an average length of less than 1000 frames or just a few long-term videos (*e.g.*, Roy et al., 2015; Chen et al., 2017). We note that the benchmark of Lieberknecht et al. (2009) could serve as a testbed for long-term planar tracking by containing 40 long sequences with an average length of 1200 frames. However, its diversity (with 5 targets) and scale (40 sequences in total) are significantly limited in further facilitating the development of planar tracking.

We notice that there exist several large-scale benchmarks (*e.g.*, Muller et al., 2018; Fan et al., 2019; Huang et al., 2019; Peng et al., 2024) for generic tracking. However, planar tracking differs fundamentally from generic tracking: instead of predicting bounding boxes, it requires estimating 2D homography via four corner points, which is crucial for applications such as augmented reality and robotics. Such geometric precision cannot be reliably achieved by post-processing generic trackers, as bounding boxes provide insufficient information and small errors are easily amplified. Owing to these different goals and settings (see Fig. 1), existing generic datasets are *not* suitable for planar tracking. In addition, a recent benchmark named MPOT-3K (Zhang et al., 2023) with 356 videos has been introduced for multi-planar tracking, which differs from the goal of single-planar tracking and is therefore not directly applicable. To further facilitate research on deep planar tracking, a dedicated large-scale benchmark is desired, which motivates our work.

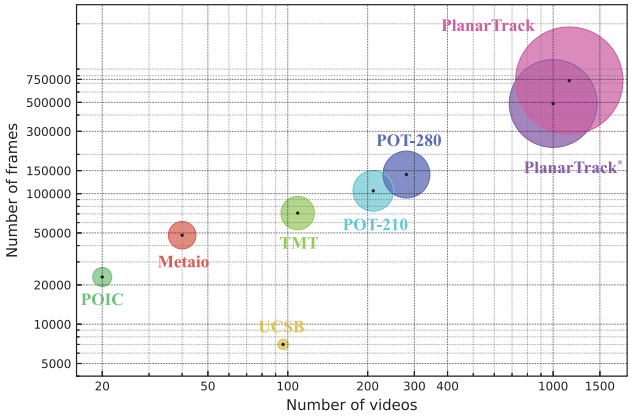


Fig. 2. Summary of planar object tracking datasets, containing POT-280 (Liang et al., 2021), POT-210 (Liang et al., 2018), TMT (Roy et al., 2015), UCSB (Gauglitz et al., 2011), Metiao (Lieberknecht et al., 2009), POIC (Chen et al., 2017), our PlanarTrack and PlanarTrack* from conference version (Liu et al., 2023). The circle diameter is in proportion to the number of frames of a dataset. Our PlanarTrack is the *largest* benchmark.

Unlike generic tracking that only predicts bounding boxes, planar tracking estimates 2D homography via four corner points, which is essential for applications such as augmented reality and robotics. Approximating this task by post-processing generic trackers is unreliable, as bounding boxes lack sufficient geometric information and small errors are easily amplified.

Table 1
Detailed comparison of the proposed PlanarTrack with other existing planar object tracking benchmarks. PlanarTrack* denotes for the conference version of PlanarTrack.

Benchmark	Year	Targets	Videos	Min frames	Mean frames	Max frames	Total frames	Annotated frames	Unconstrained videos	In the wild
Metaio (Lieberknecht et al., 2009)	2009	8	40	1200	1200	1200	48K	48K	n/a	✗
UCSB (Gauglitz et al., 2011)	2011	6	96	13	72	500	7K	7K	n/a	✗
TMT (Roy et al., 2015)	2015	12	109	191	648	2518	71K	71K	n/a	✗
POIC (Chen et al., 2017)	2017	20	20	283	1149	2666	23K	23K	n/a	✗
POT-210 (Liang et al., 2018)	2018	30	210	501	501	501	105K	53K	30	✓
POT-280 (Liang et al., 2021)	2021	40	280	501	501	501	140K	70K	40	✓
PlanarTrack* (Liu et al., 2023)	**2023**	**1000**	**1000**	**317**	**490**	**549**	**490K**	**490K**	**1000**	✓
PlanarTrack	**2024**	**1150**	**1150**	**317**	**638**	**3352**	**733K**	**733K**	**1150**	✓

1.1. Contribution

In this paper, we propose to develop a novel large-scale benchmark, named PlanarTrack, dedicated to planar object tracking. The contributions of PlanarTrack are summarized as follows:

(1) We present a dedicated large-scale benchmark, **PlanarTrack**, for planar object tracking. PlanarTrack contains 1150 sequences with more than 733K frames. All these videos are directly recorded in complicated *unconstrained* conditions from the wild scenarios. Compared to existing datasets (*e.g.*, Chen et al., 2017; Gauglitz et al., 2011; Liang et al., 2021, 2018; Lieberknecht et al., 2009; Roy et al., 2015), our PlanarTrack is much more challenging yet realistic in real applications. For each frame in PlanarTrack, we carefully inspected and manually annotated the coordinates of four corner points. To ensure annotation quality, each annotation is double-verified and corrected if necessary. As far as we know, PlanarTrack is so far the *largest* (in terms of the number of sequences and frames) and *most challenging* planar tracking dataset with high-quality dense annotations. By developing PlanarTrack, we aim to provide a dedicated large-scale platform for promoting the development and evaluations of deep-learning-based planar trackers.

(2) There is a huge increase in diversity of targets in PlanarTrack, compared to existing datasets. There are 1150 different targets while other datasets only contain 40 targets at most. The diversity of PlanarTrack makes a contribution to a more effective training and more equitable evaluations.

(3) PlanarTrack gives an opportunity for evaluation of long-term tracking. 150 out of 1150 sequences are produced as long sequences with an average length of 1622 frames. Further more, there are 4 *ultra-long* sequences longer than 3000 frames, enabling assessment of long-term trackers. Experiments on long-term and short-term sequences show that all planar trackers struggle to maintain target capture over extended periods, indicating the need for further research into long-term tracking.

(4) We offer more challenging information in PlanarTrack. Almost all sequences have multiple challenging factors (*i.e.*, *unconstrained conditions*) which are closer to the realistic scenarios, while existing benchmarks contain no or little unconstrained videos. Researchers can further understand planar trackers by carrying out experiments on different challenging factors.

(5) To analyze PlanarTrack and provide comparisons for future research, we evaluate 10 recent planar object tracking algorithms. Evaluation results show that all the trackers significantly decline on our more challenging PlanarTrack, which indicates that more efforts should be made for improvements. We further conduct an overall analysis of different challenging factors and long-term tracking with discussion to provide a guidance for future research. Besides, our re-training experiments show the usefulness and effectiveness of our benchmark in performance enhancement.

(6) To observe the performance of generic trackers in localizing planar-like targets, we develop **PlanarTrack$_{BB}$**, a by-product of PlanarTrack which is suitable for generic box tracking. We aim at *large-scale* learning and evaluation of generic trackers on tracking *rigid* targets, which is rarely investigated before. To this end, we select 15 top-performance transformer-based generic trackers for evaluation on PlanarTrack$_{BB}$. Results show that all trackers reveal heavy performance degeneration on PlanarTrack$_{BB}$ compared with existing large-scale generic tracking benchmarks (*e.g.*, LaSOT Fan et al., 2019 and TrackingNet Muller et al., 2018). More efforts should be made to handle planar objects though they are rigid.

This paper extends an early conference version in Liu et al. (2023). The main new contributions are as follows. **(i)** We expand the scale of PlanarTrack to be about 1.5 times larger in term of number of frames by introducing 243,326 new images with precise annotations. **(ii)** For long-term tracking, we introduce 150 long sequences with an average length of 1622 frames, among which 4 ultra-long sequences longer than 3000 frames are contained. Additional experiments have been conducted to highlight the significance of long-term planar object tracking. **(iii)** More details of PlanarTrack construction are provided. **(iv)** More thorough experiments and in-depth analysis are conducted on PlanarTrack for planar object tracking and PlanarTrack$_{BB}$ for generic tracking relatively, in order to show the advantages and necessity of dedicated large-scale benchmark.

The rest of this paper is organized as follows. Section 2 briefly introduces related tracking algorithms and benchmarks. In Section 3, we describe the construction of our PlanarTrack in detail with a comprehensive analysis of benchmark attributes. Experimental evaluation results and in-depth analysis are conducted in Section 4 for better understanding. Section 5 reports the construction of PlanarTrack$_{BB}$ and generic tracking experiments, followed by a conclusion in Section 6.

2. Related work

2.1. Planar tracking algorithms

Planar object tracking is a fundamental computer vision task, which aims at recovering the homography from the template to the current frame. Here we briefly review three mainstream trends including keypoint-based methods, region-based methods and deep-learning-based methods.

Keypoint-based methods Keypoint-based algorithms (Dick et al., 2013; Ozuysal et al., 2009; Wang and Ling, 2017; Hare et al., 2012; Zhao et al., 2015) typically represent an object with a set of points and their descriptors. Their tracking process is divided into two steps. Firstly, trackers detect the keypoints of objects (*e.g.*, SIFT Lowe, 2004, SURF Bay et al., 2008 and FAST Rosten et al., 2008). A pair of correspondences between object and image keypoints is established through descriptor matching. Then, a robust homography is estimated with geometric estimation algorithms (*e.g.*, RANSAC Fischler and Bolles, 1981 and its variants Torr and Zisserman, 2000; Chum and Matas, 2005). To deal with the huge per-frame motions, an approximate nearest neighbor search to estimate per-frame state updates is introduced in Dick et al. (2013). Authors in Ozuysal et al. (2009) propose to detect objects by leveraging hundreds of binary features and models class posterior probabilities in a naive Bayesian classification framework, making it

perform remarkably on datasets containing very significant perspective changes with less computational costs. A graph is applied in Wang and Ling (2017) to model a planar object and represent its structure, instead of a simple collection of keypoints.

Region-based methods Region-based methods (*e.g.*, Benhimane and Malis, 2004; Richa et al., 2011; Chen et al., 2017; Tan and Ilic, 2014) are sometimes called *direct methods*. These methods formulate the planar tracking task as an image registration problem. They directly estimate the homography by optimizing the alignment of the current frame with the object of the initial frame. The work of Benhimane and Malis (2004) presents a tracking algorithm based on minimizing the sum-of-squared-difference between a given template and the current image. The proposed minimization method is a second-order one, making it unnecessary to compute the Hessian and achieve the high convergence rate. To reduce the impact of non-linear illumination variations, the authors in Richa et al. (2011) introduced a direct tracking method based on an image similarity measure called the sum of conditional variance (SCV). The SCV requires less iterations to converge and has a significantly larger convergence radius, and achieves excellent performance under challenging illumination conditions and rapid motions. The work of Chen et al. (2017) also measures the similarity between two images through a second-order minimization method for planar object tracking. They suggested a denoising method based on the Perona–Malik function and a mask image to improve the robustness against image noise and low texture.

Deep-learning-based methods In addition to the above two types, another popular trend is to regress the homography with the deep neural networks (Zhan et al., 2022; Zhang and Ling, 2022; Li et al., 2023; Erlik Nowruzi et al., 2017; Wang et al., 2018; Liu et al., 2019; Sarlin et al., 2020; Šerỳch and Matas, 2023). A hierarchy of twin convolutional regression networks is introduced in Erlik Nowruzi et al. (2017) to estimate the homography between a pair of images. The framework achieves high performance with simple hierarchical arrangement of simple models due to the iterative nature. In Zhan et al. (2022), a novel homography decomposition approach is proposed to reduce and stabilize the condition number by decomposing the homography transformation into two groups and is trained in a semi-supervised fashion. Dense optical flow with weight is introduced in Šerỳch and Matas (2023) to estimate a homography by weighted least squares in a fully differentiable manner. HDN (Zhan et al., 2022) further improves robustness by introducing a homography decomposition network with semi-supervised learning, enabling stable estimation under challenging conditions. More recently, WOFT (Šerỳch and Matas, 2023) formulates planar tracking as weighted optical flow estimation, where homography is obtained via differentiable weighted least squares, achieving strong performance on multiple benchmarks. The above deep-learning-based planar trackers can not only avoid complicated keypoint feature extraction and be trained end to end, but also achieve outstanding performance. Thus, the deep regression-based methods have attracted increasing attention in planar tracking.

2.2. Planar tracking benchmarks

Datasets have played an important role in facilitating the development of planar object tracking. In recent years, there have been several planar tracking benchmarks, including Metaio (Lieberknecht et al., 2009), UCSB (Gauglitz et al., 2011), TMT (Roy et al., 2015), POIC (Chen et al., 2017), POT (POT-210 Liang et al., 2018, POT-280 Liang et al., 2021) and MPOT-3K (Zhang et al., 2023). Table 1 provides a detailed comparison between these benchmarks.

Metaio Metaio (Lieberknecht et al., 2009) is one of the earliest datasets for planar tracking. It consists of 40 videos with eight different textures using a camera mounted on the robotic measurement arm. The ratio of successfully tracked images is used for measuring the performance of the planar trackers.

UCSB UCSB (Gauglitz et al., 2011) has 96 sequences, containing six planar textures with 16 motion patterns each. The ground truth is semi-automatically annotated using four red markers fixed on a glass frame.

TMT TMT (Roy et al., 2015) comprises 109 sequences and each one is labeled with a challenging factor. Three trackers are used for ground truth annotations. The coordinates of four corners are determined when all three trackers are agree within a certain range. The goal of TMT is to evaluate different planar tracking algorithms for human and robot manipulation tasks.

POIC POIC (Chen et al., 2017) contains 10 sequences with total of 6663 frames. Objects with varying texture and lambertian/specular materials are provided to evaluate the performance of planar trackers in challenging complicated illumination environments.

POT Different from the above dataset collected from a simple laboratory environment, POT-210 (Liang et al., 2018) is the first one providing a dataset for planar object tracking in the wild, which contains 210 sequences of 30 planar objects. It is further extended to POT-280 in Liang et al. (2021) by introducing 70 more sequences of another 10 objects. Each planar object in POT (Liang et al., 2018, 2021) is captured in seven videos. However, six of these form one challenge, and only one contains multiple challenges in unconstrained conditions.

Previous algorithms have primarily relied on the POIC and POT datasets for experimentation and analysis. However, both datasets have significant limitations. On the one hand, POIC is small in scale and lacks sufficient category diversity, making it inadequate for fairly evaluating deep-based planar trackers, while deep-based algorithms are the current mainstream in this field. On the other hand, POT contains only seven sequences, six of which contains a single challenge factor, with only one sequence presenting multiple challenges under unconstrained conditions. This renders POT less representative of real-world scenarios. As a result, the field currently lacks a benchmark that addresses these shortcomings and provides a comprehensive evaluation framework for planar object tracking. To this end, we proposed PlanarTrack, the *largest* and most *challenging* and *diverse* benchmark with *high-quality* annotations for *long-term* planar object tracking. Table 1 displays a detailed comparison of our PlanarTrack with existing planar tracking benchmarks.

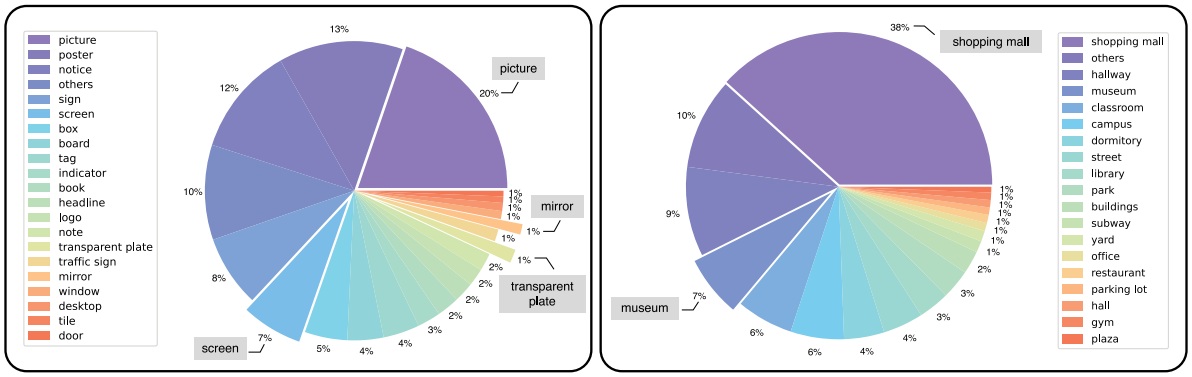
2.3. Large-scale generic tracking benchmarks

Large-scale benchmarks make it possible for efficient training and reliable evaluation, which have greatly facilitated the development of tracking in recent years. Examples of large-scale benchmarks include GOT-10k (Huang et al., 2019), LaSOT (Fan et al., 2019, 2021), TrackingNet (Muller et al., 2018), OxUvA (Valmadre et al., 2018), TNL2K (Wang et al., 2021b), and VastTrack (Peng et al., 2024).

GOT-10k GOT-10k (Huang et al., 2019) consists of 10K videos, aiming to provide rich motion trajectories for short-term tracking. It is the first one to propose a novel one-shot evaluation for assessing tracking performance.

LaSOT LaSOT (Fan et al., 2019) is a high-quality large-scale benchmark for single object tracking with 1400 sequences and more than 3.5M frames. The average sequence length is more than 2500 frames and each sequence has various challenges deriving from the wild. It is later extended in Fan et al. (2021) by providing 150 extra sequences.

TrackingNet TrackingNet (Muller et al., 2018) is the first large-scale dataset and benchmark for object tracking in the wild, which contains more than 30K videos with more than 14 million dense annotations.



(a) Distribution of classes in all sequences (b) Distribution of scenarios in all sequences

Fig. 3. Distribution of classes and scenarios in all sequences. (a): Planar targets can be divided into 21 classes. Four representative classes are highlighted. (b): Videos are all collected in these 19 scenarios.

The goal of TrackingNet is to further improve and generalize deep trackers.

OxUvA OxUvA (Valmadre et al., 2018) consists of 366 sequences spanning 14 h, which is designed for long-term tracking. It is more challenging due to the frequent target disappearance.

TNL2K TNL2K (Wang et al., 2021b) comprises 2K sequences with 124K frames and 663 words, aiming to evaluate trackers specifically for vision-language tracking.

VastTrack VastTrack (Peng et al., 2024) is a recently proposed large-scale generic tracking benchmark. It comprises over 50K video sequences with more than 2K categories, aiming to facilitate the exploration of more general and universal tracking.

Different from the aforementioned benchmarks, PlanarTrack is specifically designed for planar object tracking. Rather than using axis-aligned rectangular bounding boxes for targets, PlanarTrack utilizes corner point annotations for improved precision.

3. The proposed PlanarTrack benchmark

3.1. Design principle

Our goal is to establish a dedicated benchmark, PlanarTrack, for training and evaluating planar object trackers. To this end, we follow five principles in establishing PlanarTrack, aiming at addressing all the issues of existing planar tracking benchmarks mentioned in previous sections:

Dedicated large-scale benchmark An important motivation for our work is to train and fairly evaluate the deep-learning-based planar trackers by providing a large-scale benchmark. For this purpose, we capture 1150 sequences with over 733K frames in the proposed benchmark, which is four times larger than the scale of POT-280 (Liang et al., 2021).

Challenging realistic objects in the wild To preserve tracking challenges in complicated realistic scenarios and faithfully reflect the performance of planar trackers in practice, videos of PlanarTrack are collected from natural scenarios with multiple challenge factors (*i.e.* unconstrained condition).

Long-term tracking sequences Frequent disappear and reenter is a common situation in long-term tracking. As a result, some long sequences should be included in the benchmark for evaluating long-term tracking algorithms.

Diverse planar objects The diversity of objects is crucial for the generalization of planar trackers. Considering this, the planar target in each sequence of our PlanarTrack should be unique, which is different from the existing benchmarks (*e.g.*, POT-210/280 Liang et al., 2018, 2021).

High-quality dense annotations Accurate annotations are indispensable for effective training and fair evaluation. Therefore, each frame in PlanarTrack is manually labeled with careful refinement by well-trained annotators, in order to ensure the high-quality annotations.

3.2. Data collection

Different from existing generic object tracking benchmarks (Fan et al., 2019; Huang et al., 2019; Muller et al., 2018; Peng et al., 2024) that source videos from YouTube (https://www.youtube.com/), we construct our PlanarTrack by recording videos from reality. We record sequences from natural scenarios using mobile phone because we find that there are few videos focused on planar objects on YouTube. Specifically, we invite many volunteers who are familiar with planar tracking to capture videos using various phones with different resolutions, in order to diversify the video sources. Following the principles mentioned above, we select various categories of planar objects, including *box*, *poster*, *tag*, *picture*, *mirror*, *screen*, *traffic sign*, *tile*, *board*, *transparent plate* and so on. Each sequence has a unique target and is captured in unconstrained conditions from various natural scenes (*e.g. shopping mall*, *restaurant*, *library*, *dormitory*, *museum* for indoor scenarios, *campus*, *street*, *playground*, *park*, *plaza* for outdoor scenarios). We demonstrate the distribution of scenarios and classes in Fig. 3. From 3 we can see that, our PlanarTrack is highly diverse in both scenarios and classes. All sequences are collected in 19 scenarios, while the shopping mall occupies the highest percentage. For the diversity of objects, all planar targets are divided into 21 classes, in which the picture has the greatest number. We purposely capture some targets with unconventional appearance changes (*e.g.*, *screen*, *transparent plate* and *mirror*) to enhance the challenge of our dataset.

In total, PlanarTrack is divided into two parts. The first part (*part-1* for short) contains 1000 sequences with an average length of 490 frames. Initially, we collected over 2500 videos for *part-1*. After a careful inspection, we choose 1000 sequences which best meet the principles mentioned above. For these 1000 videos, we further verify their contents and remove inappropriate parts to ensure that they are suitable for planar tracking. Although the sequence length of *part-1* can reach the level of the existing benchmark, *part-1* does not address the issue of long-term tracking. To this end, we introduce another part (*part-2* for short), which comprises 150 *long* sequences with an average length of 1622 frames, which contains 4 *ultra-long* sequences of more than 3000 frames. We at first recorded more than 300 sequences in other

Fig. 4. Examples of annotated sequences in the proposed PlanarTrack. Each video is annotated with four corner points.

places different from *part-1*. In these long sequences, we capture objects that frequently enter and leave the view to reflect the real-world scenarios. After carrying through the same selecting and preprocessing flow, we provide 150 sequences with the best quality in *part-2*. Eventually, we compile our PlanarTrack, a large-scale challenging benchmark dedicated to planar tracking by including 1150 unconstrained sequences with more than 733K frames from 1150 unique planar objects. Table 1 provides a detailed summary of PlanarTrack and its comparison with existing planar tracking benchmarks.

3.3. Annotation

PlanarTrack is annotated by several well-trained annotators and experts. We manually label each frame to provide a high-quality dense annotation. We employed a customized annotation tool developed in MATLAB, which allowed annotators to mark the four corner points with zoom-in support under challenging conditions. Before annotation, annotators were trained with clear guidelines covering common cases, missing corners, and heavy occlusion or blur. Specifically, we annotate four corner points for the planar target of each frame in the given order if all its four corner points or four edges are clearly visible. When the four corner points and four edges are both hard to recognize due to the occlusion, out-of-view or heavy blur, we will assign an absent flag to this frame.

With the above strategy, we carry out the annotation by the following workflow. Firstly, each sequence is annotated by an annotator. The annotation result is then distributed to two experts for double verification. If the annotation is not unanimously approved by the experts, it will be returned to the original annotator for careful refinement. Such a verification-refinement process will last for multiple rounds until the annotation finally receives unanimous approval in order to ensure the high annotation quality. Fig. 4 shows some annotation examples of PlanarTrack.

In order to better understand our PlanarTrack, we show four representative statistics of the annotations in Fig. 5, compared with POT-210/280. Specifically, we present the distributions of target motion, target size (area of target), target scaling (relative area to the initial target) and Intersection over Union (IoU) between targets in adjacent frames. From Fig. 5, we find that the planar targets in PlanarTrack have rapid size changes and speed of movement. Compared to POT-210/280 (Liang et al., 2018, 2021), PlanarTrack has relatively smaller target sizes and faster motions, while most target of POT-210/280 scale around 1 relative to the initial target and only moves a few pixels. Therefore, our PlanarTrack provides new challenges for planar tracking in the wild.

Notice that, since POT-210/280 labels every two frames, we perform linear interpolation on their annotations for statistics comparison.

3.4. Analysis of ground truth quality

Since the ground truth (GT) for each frame in our PlanarTrack dataset is manually annotated, some errors are inevitably introduced. To select appropriate evaluation metric thresholds and prevent researchers from overfitting to GT errors, we conducted an analysis of the GT quality in PlanarTrack.

Specifically, following WOFT (Šerỳch and Matas, 2023), we randomly selected a small subset from PlanarTrack, consisting of 10,920 frames, which was meticulously annotated by two experts highly familiar with planar object tracking, obtaining a refined GT. Subsequently, we computed the root of the mean square distances between the GT and the refined GT (*i.e.*, the alignment error). Given four GT points $\mathbf{x}_i \in \mathbf{X}$

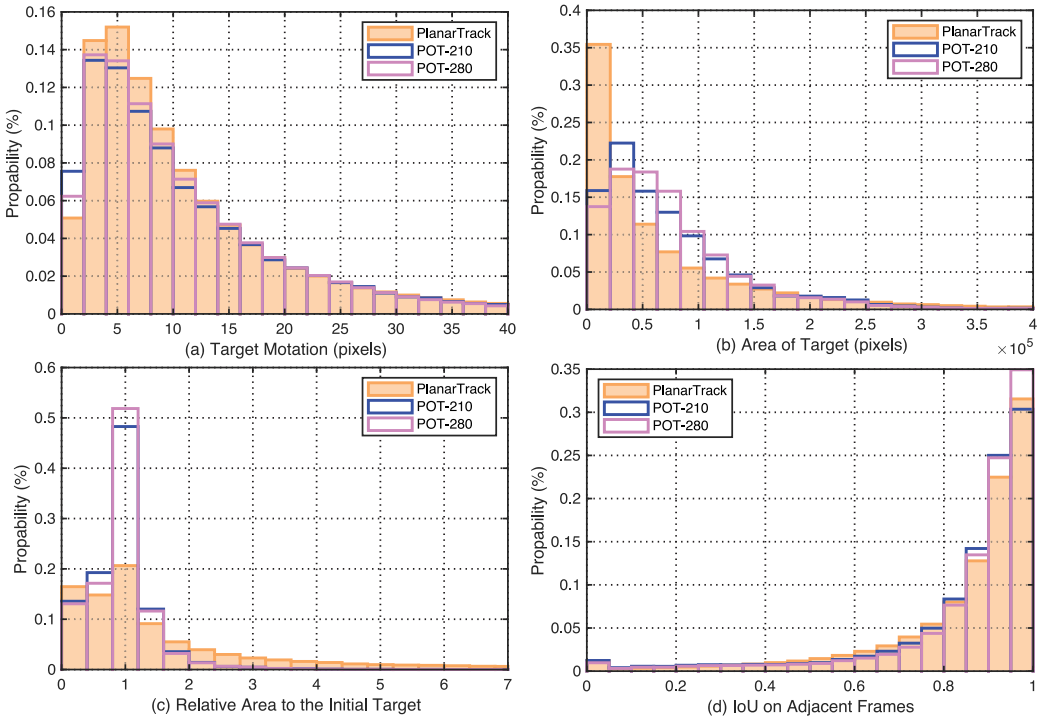


Fig. 5. Statistics of planar target motion, size, relative area compared to initial object and IoU of targets in adjacent frames in PlanarTrack and comparison with the recent POT-210/280 (Liang et al., 2018, 2021). We can see the targets in our dataset have smaller sizes and faster and more challenging motions.

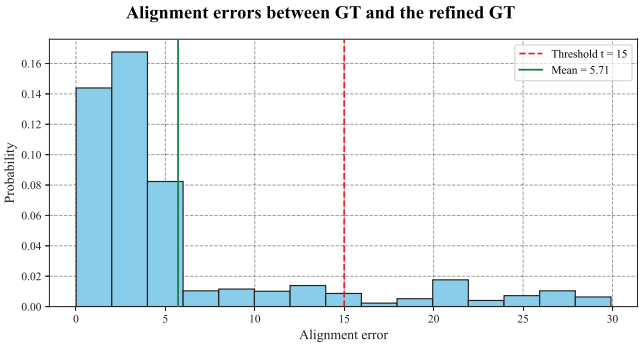


Fig. 6. Distribution of alignment error between the GT and the refined GT of PlanarTrack.

and four refined GT points $\mathbf{x}_i^* \in \mathbf{X}^*$, the alignment error e_{AL} can be calculated as

$$e_{\mathrm{AL}}(\mathbf{X}, \mathbf{X}^*) = \sqrt{\frac{1}{4}\sum_{i=1}^{4}\left(\mathbf{x}_i - \mathbf{x}_i^*\right)^2}. \tag{1}$$

The results indicate that the mean alignment error between the GT and the refined GT on the refined-annotated subset of PlanarTrack is *5.71* pixels. Fig. 6 illustrates the distribution of alignment errors, with *10.71%* of annotations exhibiting errors exceeding 15 pixels. Please note that, our PlanarTrack includes a greater number of challenging scenarios, such as *heavier blur*, *more extreme illumination changes*, and *faster motion*. These factors make our precise annotation more difficult. Consequently, compared to the GT quality of POT-210 reported in WOFT (Šerỳch and Matas, 2023), our PlanarTrack exhibits slightly higher errors.

3.5. Challenging factors

Following other tracking benchmarks (Liang et al., 2018; Fan et al., 2021), we label each sequence with several challenging factors in PlanarTrack to further analyze planar tracking algorithms in different challenging conditions. Specifically, we define eight challenging factors that widely exist for planar tracking. The challenging factors are listed below:

Occlusion (OCC) Object is occluded by itself or other objects in the background. To increase the difficulty, we also manually occlude the object while moving the camera.

Motion Blur (MB) Motion blur caused by fast camera movement at low frame rates can generate the fuzzy corner points, making it difficult to track a planar object robustly.

Rotation (ROT) Rotation describes a common situation that an object's direction is changed relative to the camera.

Scale Variation (SV) Scale variation is assigned when the ratio of planar annotation is outside the range $[0.5, 2]$.

Perspective Distortion (PD) Perspective distortion is assigned when the perspective between the object and camera is changed.

Out-of-view (OV) Out-of-view is assigned when part or all of the object leaves the image, which makes some sides or corners of the target invisible.

Low Resolution (LR) Low resolution is assigned when the region of the target in any frame of a sequence is less than 1000 pixels.

Background Clutter (BC) Background clutter is assigned when the background region looks visually similar to the target, including similar colors, multiple similar targets, etc.

Light Interactive Surface (LIS) Light Interactive Surface is assigned when significant appearance changes of the planar object occur due to light phenomena such as reflection and refraction, *e.g.*, *mirrors* and

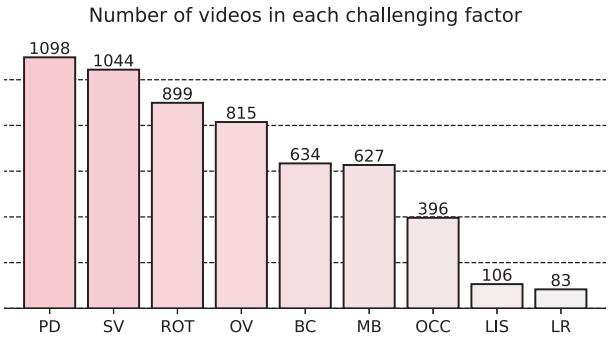


Fig. 7. Distribution of sequences on each challenging factor. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

transparent plates. *Screens* are also classified under this category, as the videos displayed on them can cause significant appearance changes.

It is worth mentioning that, some common challenging factors used in generic object tracking are not suitable for planar objects. Thus, we exclude a few of them, such as deformation and illumination change. The vast majority of sequences (1135 out of 1150) in PlanarTrack simultaneously contain multiple challenging factors (*i.e.*, recorded in *unconstrained conditions*). Therefore, our PlanarTrack is much more challenging and practical for real applications, compared to POT-210/280.

The distribution of the above challenging factors on PlanarTrack is presented in Fig. 7. We notice that perspective distortion is the most common challenging factor in PlanarTrack, which may lead to serious misalignment problems for planar tracking. In addition, scale variation and rotation frequently exist in PlanarTrack.

3.6. Dataset split and evaluation metric

Training/Test Set Split PlanarTrack contains 1150 sequences. We use 805 sequences for training ($\text{PlanarTrack}_{\text{Tra}}$) and 345 for evaluation ($\text{PlanarTrack}_{\text{Tst}}$). We try our best to keep the distributions of training and test sets close to each other. As for the four ultra-long sequences, we put two of them into a training set and the other two into a test set for long-term tracking and evaluation. Table 2 shows a comparison of these two sets.

For further comparison between training and test sets of PlanarTrack, we present the ratios of sequences in these two sets on eight different challenging factors in Fig. 8. From Fig. 8 we can see that, our split makes the training and test sets closing to each other, which ensures the consistency of training/test split in PlanarTrack. Notice that, the number of test sequences is significantly higher than training sequences on OV factor. This is because frequent disappearance may lead to a decrease of training data but make it more challenging for evaluation. Detailed split files will be released on our project website.

Evaluation Metric For the evaluation, we adopt the *precision* (PRE) metric following Liang et al. (2021). Please note here, we do *not* utilize the SUC metric as in previous studies for evaluation, because the SUC, that represents the percentage of successful frames in which the error between estimated and real homography is less than or equal to a certain threshold, depends heavily on the position of the target in the image. When the target is located in the bottom-right corner of the image, a very small tracker imprecision can lead to a huge re-projection error. This makes the SUC metric *cannot* access the true accuracy of tracking results.

However, there are some differences between our PRE and that used for generic tracking (Wu et al., 2013). For planar tracking, PRE

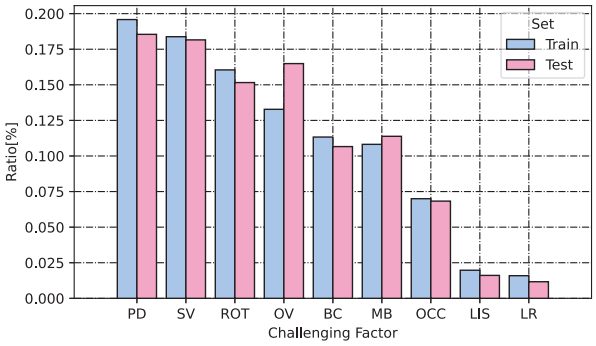


Fig. 8. Distribution of challenging factor on training and testing sets.

Table 2
Comparison of *training* and *test* sets.

	Videos	Min frames	Mean frames	Max frames	Total frames
$\text{PlanarTrack}_{\text{Tra}}$	805	317	636	3352	512K
$\text{PlanarTrack}_{\text{Tst}}$	345	362	641	3150	221K

is defined as the percentage of frames in which the alignment error between corner points of predicted result and groundtruth is within a given threshold. Based on the quality analysis of GT in Section 3.4, we selected 15 pixels as the primary threshold for the PRE metric. Additionally, since 75.98% of cases exhibit errors below 5 pixels, we retained the 5 px threshold for the PRE metric as used in POT-210. In summary, we adopted 5 px and 15 px thresholds for the PRE metrics to enable a more comprehensive evaluation, denoted as P@5 and P@15, respectively.

4. Evaluation

4.1. Evaluated planar object tracking algorithms

We do several evaluations of planar object trackers on PlanarTrack to demonstrate its reliability and novelty. As there are not many planar object trackers compared to generic tracking (actually, this is the biggest motivation for us to introduce PlanarTrack for promoting research on planar object tracking), we select 10 representative algorithms about planar tracking with accessible source codes. Specifically, these trackers are WOFT (Šerỳch and Matas, 2023), HDN (Zhan et al., 2022), GIFT (Liu et al., 2019), LISRD (Pautrat et al., 2020), SIFT (Lowe, 2004), Gracker (Wang and Ling, 2017), SOL (Hare et al., 2012), SCV (Richa et al., 2011), ESM (Benhimane and Malis, 2004) and IC (Baker and Matthews, 2004). Particularly, WOFT (Šerỳch and Matas, 2023) and HDN (Zhan et al., 2022) are two recent planar trackers using deep learning. All other algorithms can be used for homography estimation. We modify them to the corresponding planar object trackers. It is worth mentioning that, we are not able to evaluate generic trackers on PlanarTrack because of the incompatible inputs and results. For this, we construct a new $\text{PlanarTrack}_{\text{BB}}$ for generic tracking evaluation, as described later.

4.2. Evaluation results

4.2.1. Overall performance

Totally, we evaluate 10 representative planar object trackers on $\text{PlanarTrack}_{\text{Tst}}$, among which WOFT and HDN are utilized without modifications as they are specifically developed for the planar tracking task. For the remaining methods, we modify them so that they can be used for planar object tracking. Their implementations except GIFT and LISRD are borrowed from Liang et al. (2018). We adapt GIFT and LISRD to planar object tracking due to some setting problems

Table 3
Summary of evaluated planar trackers. Representation: "Deep" for deep-learning-based Method, "Keypoint" for Keypoint-based Method, and "Direct" for Direct Method.

Method	Backbone	Representation		
		Deep	Keypoint	Direct
WOFT (Šerých and Matas, 2023)	RAFT	✓		
HDN (Zhan et al., 2022)	ResNet-50	✓		
GIFT (Liu et al., 2019)	CNN	✓		
LISRD (Pautrat et al., 2020)	VGG16	✓		
SIFT (Lowe, 2004)	–		✓	
Gracker (Wang and Ling, 2017)	–		✓	
SOL (Hare et al., 2012)	–		✓	
SCV (Richa et al., 2011)	–			✓
ESM (Benhimane and Malis, 2004)	–			✓
IC (Baker and Matthews, 2004)	–			✓

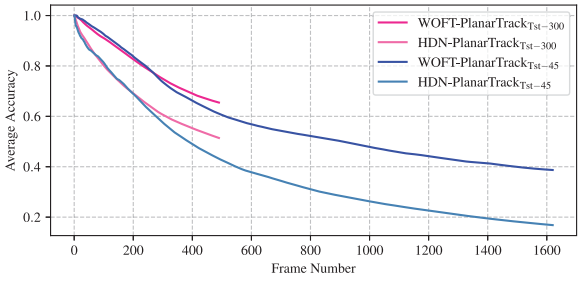


Fig. 9. Accuracy changes of two planar trackers WOFT and HDN with respect to frame number.

in Liang et al. (2018). Fig. 10 shows the evaluation results of the above approaches in P@5 and P@15. From Fig. 10 we can see that, WOFT achieves the best P@5 score of 0.402 and P@15 score of 0.607. GIFT applies transformation-invariant deep visual descriptors for planar object tracking, which demonstrates the second best P@5 score of 0.221 and P@15 score of 0.402. Notice that, all the top four approaches leverage deep neural networks for planar target localization, which shows the great potential of deep-learning-based planar tracking in the future.

Short-term Tracking analysis Our PlanarTrack consists of 1000 sequences with an average length of 490 frames, which is suitable for short-term tracking. To evaluate the performance of deep-learning-based planar trackers, we perform regular experiments on PlanarTrack$_{Tst-300}$, the test set for short-term tracking. Evaluation results are shown in Table 4. WOFT achieves the highest P@15 score of 0.641, which is obviously better than HDN.

Long-term Tracking analysis To analyze the performance of the top four methods in long-term planar object tracking, we demonstrate the tracking results on PlanarTrack$_{Tst-300}$, PlanarTrack$_{Tst-45}$ and PlanarTrack$_{Tst-345}$ in Table 4. Notice that, PlanarTrack$_{Tst-45}$ is the test set consisting entirely of long sequences, while PlanarTrack$_{Tst-345}$ is the test set of the whole PlanarTrack. From Table 4, we can observe that both WOFT and HDN show performance degradation while HDN has the most significant decline in the long-term tracking scenario. Additionally, we plot the accuracy of these tow planar trackers as a function of frame number, as shown in Fig. 9. From Fig. 9, it can be observed that, the accuracy trends for the same tracker in the short-term intervals of PlanarTrack$_{Tst-300}$ and PlanarTrack$_{Tst-45}$ are relatively similar. However, during long-term tracking on PlanarTrack$_{Tst-45}$, the accuracy consistently declines, suggesting that current trackers struggle to maintain target capture over extended periods. Several factors may contribute to this issue. For example, frequent disappearances and reappearances of the target over time can cause significant spatial shifts relative to the last successfully tracked frame, which is particularly

Table 4
Comparison and analysis of two planar trackers in short-term tracking and long-term tracking.

		WOFT	HDN
PlanarTrack$_{Tst-300}$	P@5	0.433	0.263
	P@15	0.641	0.499
PlanarTrack$_{Tst-45}$	P@5	0.253	0.085
	P@15	0.379	0.164
PlanarTrack$_{Tst-345}$	P@5	0.402	0.211
	P@15	0.607	0.455

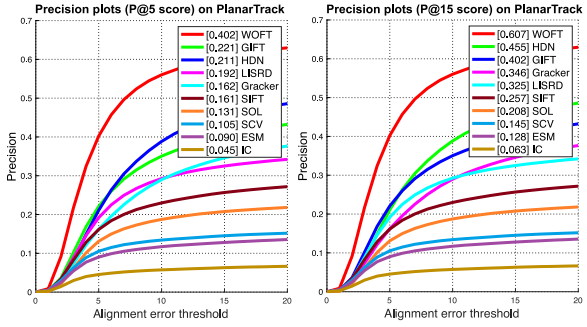


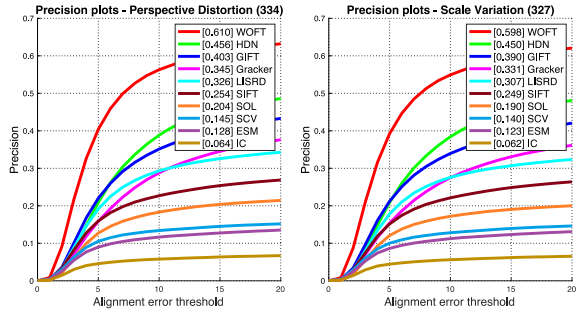
Fig. 10. Precision plots of all planar trackers on PlanarTrack$_{Tst}$ using P@5 score and P@15 score, respectively.

detrimental to trackers relying on displacement prediction. Additionally, repeated appearance changes of the target over a long duration may exceed the trackers' ability to manage long-term associations.

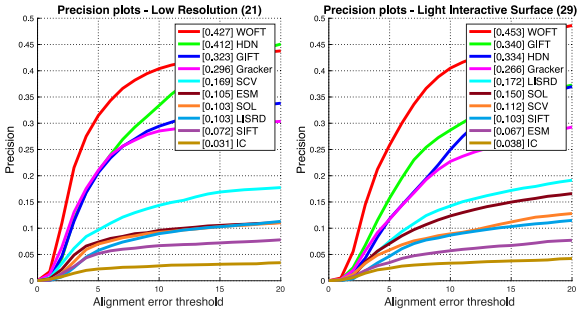
This highlights the need for a dedicated platform dedicated for long-term planar object tracking, which could drive the development of advanced long-term tracking algorithms.

4.2.2. Challenging factor-based evaluation

For better analysis of different planar trackers, we further evaluate the above trackers on the eight challenging factors. Fig. 11 displays the tracking results on the two most common challenging factors (*perspective distortion* (PD) and *scale variation* (SV)) and on the two most difficult challenging factors (*low resolution* (LR) and *light interactive surface* (LIS)). From Fig. 11 we can see that, WOFT achieves the best performance on both the commonest and most difficult scenarios. Specifically, WOFT achieves the best P@15 scores of 0.610, 0.598, 0.427 and 0.453 on PD, SV, LR and LIS, which again shows the importance of temporal information for planar tracking. Besides, the tracking performances severely decrease on LR and LIV. A reasonable explanation is that these two challenges may be harmful to the feature extraction of points or targets, leading to tracking drifts or failures.



(a) Evaluation on the two most common challenging factors using precision.



(b) Evaluation on the two most difficult challenging factors using precision.

Fig. 11. Precision plots of trackers on the two most common challenging factors including *perspective distortion* and *scale variation* and on the two most difficult challenging factors including *low resolution* and *light interactive surface* using P@15.

From our perspective, research should be devoted to improvements in these two situations.

Fig. 12 shows the whole results on all 9 challenging factors with P@15 score. From Fig. 12 we observe that WOFT achieves the best performance on all 9 challenging factors with P@15 scores. HDN obtains the second best results on 8 out of 9 factors with P@15 score. Among the four deep-learning-based tracking methods, WOFT is far ahead of the rest three approaches due to the introduction of temporal information. An interesting observation is that LISRD performs extremely poorly on LR. A potential reason is that the small target information is buried in background when extract features by its CNN-based backbone.

4.2.3. Qualitative evaluation

To better understand the above planar trackers, we demonstrate sampled tracking results of them in different challenging factors such as *background clutter*, *scale variation*, *perspective distortion*, *motion blur*, *rotation*, *out-of-view*, *low resolution* and *ultra-long-term tracking* in Fig. 13. From Fig. 13 we observe that, although some trackers can deal with certain challenging factors, they may drift to the background region or fail to localize the planar target when multiple challenging factors occur simultaneously. For Fig. 13-(a), trackers except WOFT can only roughly localize the target with large alignment error because of the varying reflection and large scale variation. A possible solution to handle this issue is to use some temporal information with the last and current frames (like optical flow in WOFT). We also evaluate the trackers on our proposed *ultra-long* sequences (see Fig. 13-(f)). WOFT can localize the planar target in most frames benefit from its motion clues. However, it may misidentify when there are similar targets (Fig. 13-(g)).

4.3. Comparison with POT-210

POT-210 (Liang et al., 2018) is currently one of the most popular benchmarks for planar object tracking. However, there remain some issues that limit the development of deep-learning-based planar object tracking algorithms. Firstly, most videos of POT-210 contain mainly one challenging factor and very few (*i.e.* 30 in POT-210 and 40 in POT-280) are involved in unconstrained conditions. This could not faithfully reflect the difficulties and complexities in reality for evaluation. Besides, the lack of planar target diversity also limits its usage. In addition, the biggest drawback is that POT-210 only contains 53K annotated frames (70K in POT-280), which is far from enough for training and fair evaluation. To address these issues, we first construct PlanarTrack with 1150 sequences and totally 733K frames, making it a large-scale benchmark for planar object tracking. For each sequence, we freely capture a unique target for diversity with multiple challenging factors. Therefore, our PlanarTrack is more challenging and realistic in practical applications.

To verify the above, we compare existing planar trackers on POT-210 and PlanarTrack$_{\text{Tst}}$. Please note that, among the ten selected trackers, only four trackers are deep-based (*i.e.*, WOFT, HDN, GIFT and LISRD) that require training before inference, as shown in Table 3. The remaining six trackers are training-free and can directly track planar objects. Therefore, in Table 5, we evaluate the performance of the six training-free trackers by directly performing inference on POT-210, POT210$_{\text{UC}}$, and PlanarTrack$_{\text{Tst}}$. For the four deep-based trackers, we first train them on POT-210 and then perform inference on POT-210, POT210$_{\text{UC}}$, and PlanarTrack$_{\text{Tst}}$ to obtain the evaluation results.

Table 5 shows the tracking results. From Table 5 we observe that, WOFT achieves the best P@5 score of 0.805 and 0.768 on POT-210 and POT210$_{\text{UC}}$. However, when used for tracking planar targets on PlanarTrack$_{\text{Tst}}$, its performance is significantly degenerated. GIFT with the second best performance also absolutely declines from POT-210 to PlanarTrack$_{\text{Tst}}$. Other trackers are declined more or less on PlanarTrack$_{\text{Tst}}$.

In addition to POT-210, we further compare POT-210$_{\text{UC}}$, a small subset of POT-210 with all videos captured in unconstrained conditions, with PlanarTrack$_{\text{Tst}}$ in Table 5, as they are both have multiple challenging factors in a sequence. As in Table 5, tracking performances on POT-210$_{\text{UC}}$ are significantly worse than those on POT-210, which means that POT-210$_{\text{UC}}$ is more challenging than POT-210. Compared to POT-210$_{\text{UC}}$, all trackers achieve the worst P@5 score on PlanarTrack$_{\text{Tst}}$, which implies that our PlanarTrack is challenging. The best tracker WOFT on POT-210$_{\text{UC}}$ shows P@5 score of 0.768, while it degrades to 0.402 on PlanarTrack$_{\text{Tst}}$ with an absolute drop of 36.6%.

Table 5
Comparison of PlanarTrack$_{\text{Tst}}$ to POT-210 (Liang et al., 2018) and its subset POT-210$_{\text{UC}}$ in unconstrained condition using P@5 score. We also compare the P@5 score and P@15 score on our PlanarTrack$_{\text{Tst}}$.

Method	POT-210	POT-210$_{\text{UC}}$	PlanarTrack$_{\text{Tst}}$	
	P@5	P@5	P@5	P@15
WOFT (Šerých and Matas, 2023)	0.805	0.768	0.402	0.607
HDN (Zhan et al., 2022)	0.612	0.567	0.211	0.455
GIFT (Liu et al., 2019)	0.553	0.528	0.221	0.402
LISRD (Pautrat et al., 2020)	0.617	0.581	0.192	0.325
SIFT (Lowe, 2004)	0.692	0.578	0.161	0.257
Gracker (Wang and Ling, 2017)	0.392	0.185	0.162	0.346
SOL (Hare et al., 2012)	0.417	0.289	0.131	0.208
SCV (Richa et al., 2011)	0.228	0.105	0.105	0.145
ESM (Benhimane and Malis, 2004)	0.204	0.100	0.090	0.128
IC (Baker and Matthews, 2004)	0.121	0.053	0.045	0.063

Table 6
Retraining of HDN (Zhan et al., 2022) using PlanarTrack$_{\text{Tra}}$.

		Original HDN	Retrained HDN
POT-210 (Liang et al., 2018)	P@5	0.612	0.637 (+2.5%)
PlanarTrack$_{\text{Tst}}$	P@5	0.211	0.281 (+7.0%)
	P@15	0.455	0.520 (+6.5%)

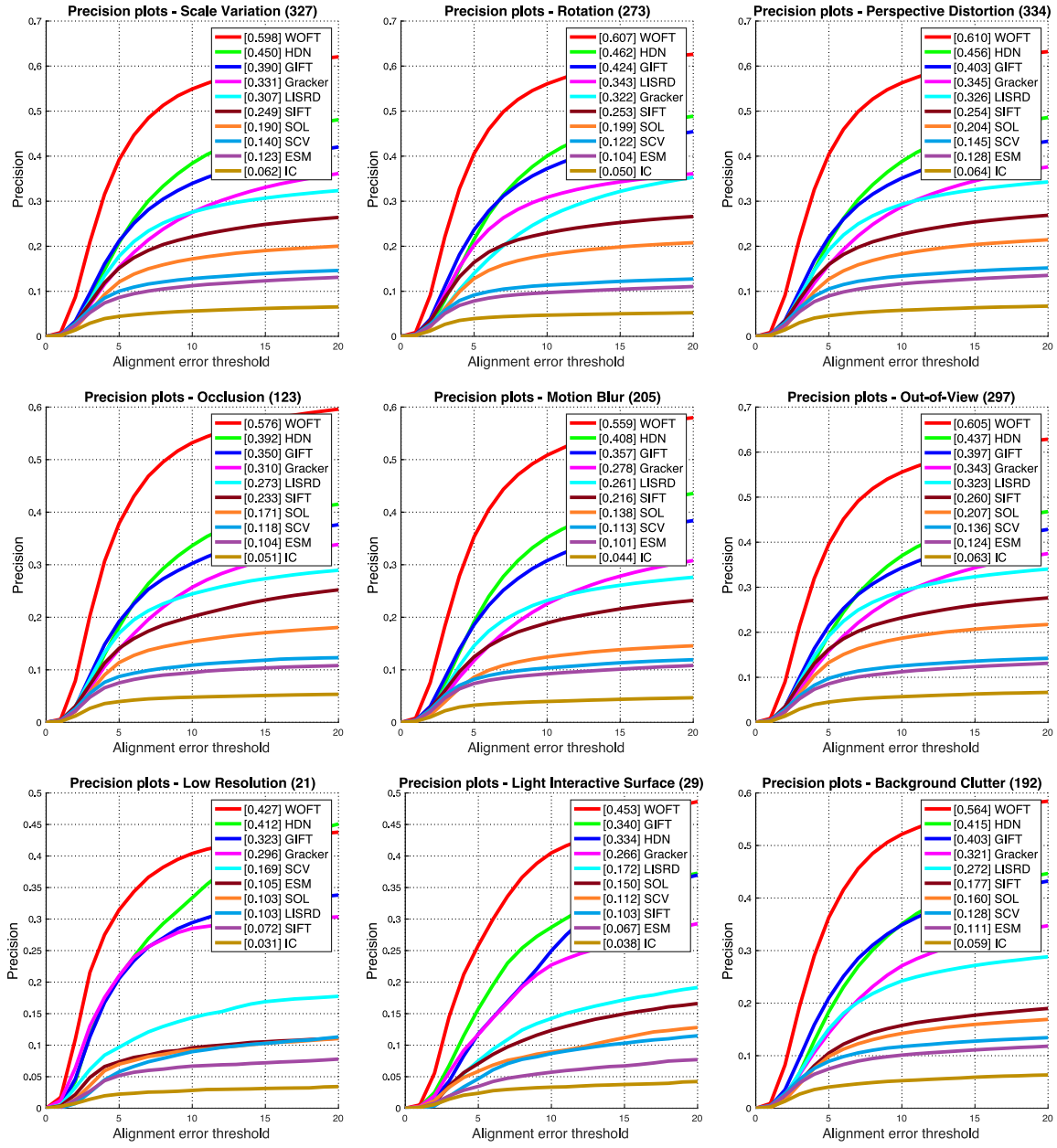


Fig. 12. Precision plots of trackers on each challenging factor using P@15 score. Best viewed in color.

From the above comparisons and analysis, we clearly see that POT-210 is a little simple for existing deep-learning-based planar trackers, which limits the development of planar object tracking algorithms. By contrast, our PlanarTrack is more challenging, complicated and large enough for planar object tracking. There is still a big room for improving tracking performance on PlanarTrack.

4.4. Retraining on PlanarTrack

Deep-based algorithms often face the challenge of data hungry, where increasing the dataset size can significantly enhance generalization performance. As one of our central aspirations is to provide a large-scale platform for promoting the development of deep-learning-based planar trackers, we conduct retraining experiments on PlanarTrack. Please note that, among the four deep-based algorithms, GIFT and LISRD are not end-to-end trackers and are *not* well-suited for retraining. WOFT released code but did *not* provide a training script. As a result, we performed the retraining experiments solely on HDN. Specifically, we retrain the recent HDN using $\text{PlanarTrack}_{\text{Tra}}$, instead of the synthetic data. While retraining, all the parameters and settings are kept the same as in the original method. After retraining, we demonstrate the results of HDN on POT-210 and $\text{PlanarTrack}_{\text{Tst}}$ in Table 6. From Table 6, we observe consistent performance gains on the two benchmarks. In other words, leveraging enough task-specific data in training can obviously improve the tracking performance. In specific, after retraining and testing on POT-210 by a fixed training/test split, the P@5 scores on POT-210 are increased from 0.612 to 0.637, with an absolute improvement of 2.5%. On $\text{PlanarTrack}_{\text{Tst}}$, the P@5 and P@15 scores have a more significant rise of 7.0%/6.5%, from 0.211/0.455 to 0.281/0.520. These improvements show that a large-scale training set is effective and necessary for improving planar object tracking performance.

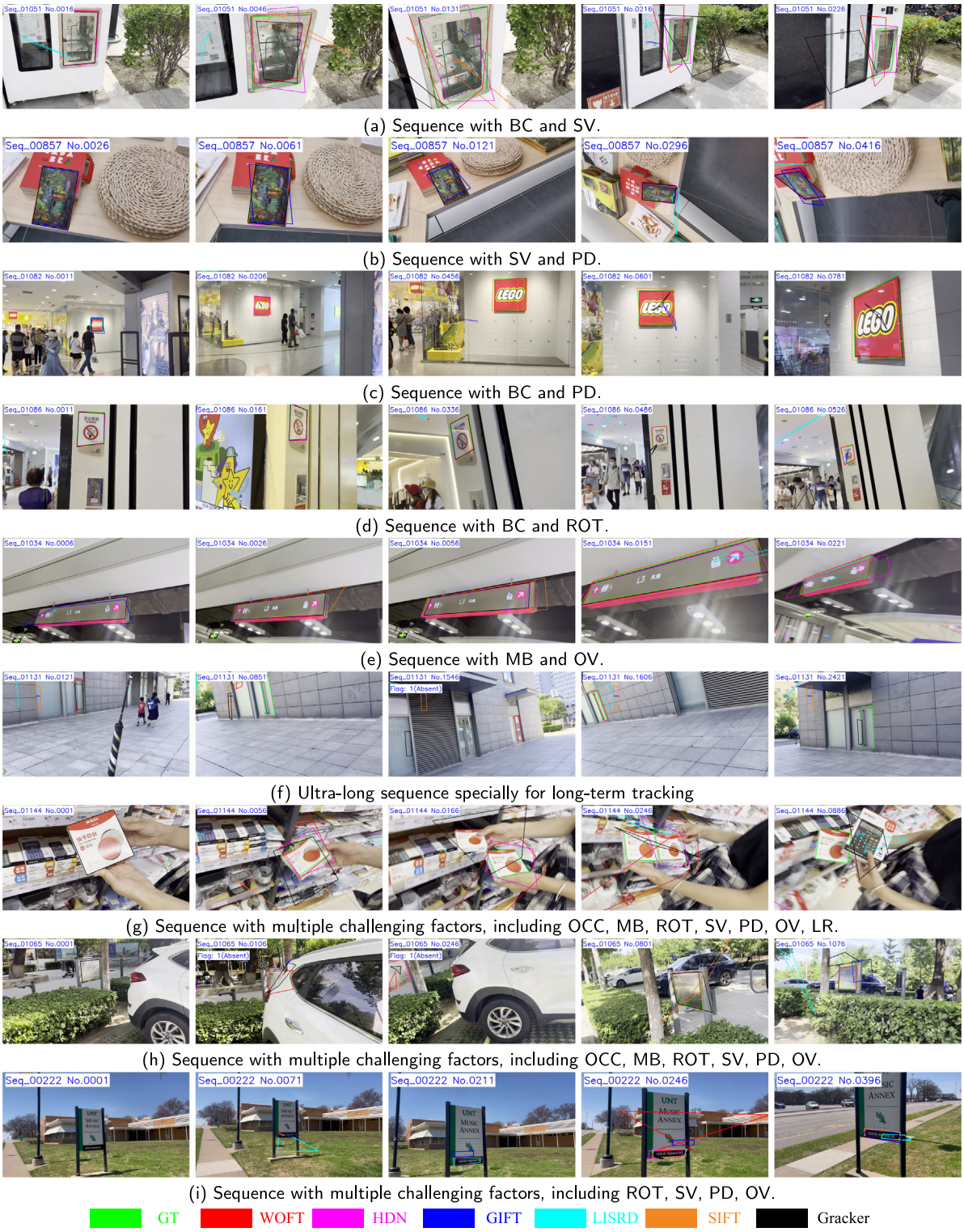


Fig. 13. Qualitative results of six trackers with the highest precision scores on different sequences. We observe that these planar trackers drift to the background region or even lose the target object due to different challenging factors in the videos such as background clutter, scale variation, perspective distortion, motion blur, rotation, out-of-view and low resolution.

5. PlanarTrack$_{BB}$ and experiments

A certified generic tracker should be able to locate the targets robustly without prior knowledge of their categories. Planar objects (*e.g.* posters, screen, board) are very common things in our daily life. Surprisingly, there is little study on localization of planar targets with *generic visual trackers*at large scale, even in the existing large-scale generic tracking benchmarks (*e.g.* Fan et al., 2021; Huang et al., 2019; Muller et al., 2018).

In order to figure out the capacities of these generic trackers in tracking planar targets, we further develop a new benchmark named PlanarTrack$_{BB}$ based on our PlanarTrack. To be specific, PlanarTrack$_{BB}$

Table 7
Evaluation of generic trackers on PlanarTrack$_{BB}$ and comparison with other popular generic benchmarks using SUC$_{BB}$.

	TrackingNet (Muller et al., 2018)	LaSOT (Fan et al., 2019)	PlanarTrack$_{BB}$ (ours)
SeqTrack (Chen et al., 2023)	0.855	0.725	0.670
ROMTrack (Cai et al., 2023)	0.841	0.714	0.667
DropTrack (Wu et al., 2023)	0.841	0.718	0.665
MixFormerV2 (Cui et al., 2024)	0.834	0.706	0.648
MixFormer (Cui et al., 2022)	0.839	0.701	0.647
OStrack (Ye et al., 2022)	0.839	0.711	0.642
SwinTrack (Lin et al., 2022)	0.840	0.713	0.638
ARTrack (Wei et al., 2023)	0.856	0.731	0.633
TransInMo (Guo et al., 2022)	0.817	0.657	0.620
STARK (Yan et al., 2021)	0.820	0.671	0.615
AiATrack (Gao et al., 2022)	0.827	0.690	0.613
TransT (Chen et al., 2021)	0.814	0.649	0.603
SimTrack (Chen et al., 2022)	0.834	0.705	0.601
ToMP (Mayer et al., 2022)	0.815	0.685	0.597
TrDiMP (Wang et al., 2021a)	0.784	0.639	0.589

shares the same images and training/test split as PlanarTrack. The only difference between PlanarTrack$_{BB}$ and PlanarTrack is that we convert annotations from four annotated corner points to an axis-aligned bounding box in PlanarTrack$_{BB}$, especially used for large-scale evaluation of generic trackers. Specifically, we calculate the axis-aligned bounding box based on the four annotated corner points and adjust it to ensure it completely fits within the image boundaries. Notice that, in PlanarTrack$_{BB}$ we actually represent the coordinates of the axis-aligned bounding box in XYWH format (*i.e.* $[x_{\min}, y_{\min}, width, height]$) like LaSOT (Fan et al., 2019) and GOT-10k (Huang et al., 2019). The difference and some examples of PlanarTrack and PlanarTrack$_{BB}$ are demonstrated in Fig. 14.

To further understand PlanarTrack$_{BB}$, we select 15 recent state-of-the-art generic trackers for evaluation. All the trackers are transformer-based, including SeqTrack (Chen et al., 2023), ROMTrack (Cai et al., 2023), DropTrack (Wu et al., 2023), MixFormerV2 (Cui et al., 2024), MixFormer (Cui et al., 2022), OStrack (Ye et al., 2022), SwinTrack (Lin et al., 2022), ARTrack (Wei et al., 2023), TransInMo (Guo et al., 2022), STARK (Yan et al., 2021), AiATrack (Gao et al., 2022), TransT (Chen et al., 2021), SimTrack (Chen et al., 2022), ToMP (Mayer et al., 2022), TrDiMP (Wang et al., 2021a). We employ the best version of each generic tracker for evaluation except SimTrack and ARTrack. Sim-L/14 performs best but only Sim-B/16 is released in Simtrack, while ARTrack-L$_{384}$ achieves the best performance but only ARTrack-B$_{384}$ is given. For metrics, we use the success score for bounding box-based tracking (Wu et al., 2013), named SUC$_{BB}$.

Table 7 shows the evaluation results of the above generic trackers and comparisons with existing large-scale generic tracking benchmarks including LaSOT Fan et al., 2019 and TrackingNet Muller et al., 2018. Due to the different evaluation metrics, we do not compare our PlanarTrack$_{BB}$ with GOT-10k (Huang et al., 2019). From Table 7 we observe that, although existing generic trackers can achieve remarkable performance on LaSOT and TrackingNet, they are significantly degraded when handling planar-like targets on PlanarTrack$_{BB}$. For instance, the best generic tracker SeqTrack obtains 0.855/0.725 SUC scores on LaSOT/TrackingNet, but obviously declines to 0.670 on PlanarTrack$_{BB}$, with an absolute drop of 18.5%/5.5%. The second best ROMTrack is also decreased from 0.841/0.714 to 0.667. This may indicate that more attention should be paid to improve such planar trackers, though they are rigid.

For in-depth analysis of generic tracking performances on PlanarTrack$_{BB}$, we further demonstrate the evaluation results of the above generic trackers in Fig. 15 by using a modified LaSOT (Fan et al., 2019) evaluation toolkit. Under One Pass Evaluation (OPE) protocol, we utilize bounding box-based precision and success plots as in generic tracking (Wu et al., 2013) for assessment. From Fig. 15 we can see that, the top two generic trackers SeqTrack and ROMTrack achieve 0.684/0.670 and 0.674/0.667 relatively on PlanarTrack$_{BB}$.

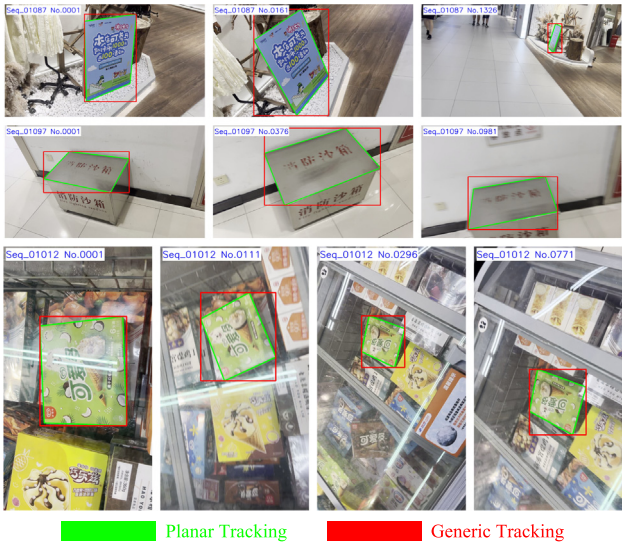


Fig. 14. Examples from PlanarTrack$_{BB}$. The targets are annotated by white axis-align bounding boxes for genetic visual tracking. Best viewed in color.

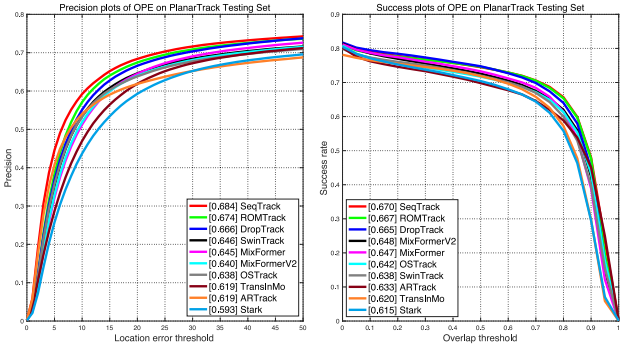


Fig. 15. Performance of evaluated generic visual trackers on PlanarTrack$_{BB}$ using bounding box-based precision and success plots. To facilitate clearer analysis, we exclusively present the top 10 trackers. Best viewed in color.

6. Conclusion

In this paper, we introduced a brand new benchmark named PlanarTrack. PlanarTrack consists of 1150 videos recorded in unconstrained

conditions from realistic scenarios, and has more than 733K annotated image frames in total. High-quality dense annotations are provided and great diversity of targets is ensured in PlanarTrack. To the best of our knowledge, PlanarTrack is the *first* challenging large-scale dataset dedicated to planar object tracking. To further understand existing approaches and provide a comparison for further research, we perform experiments by evaluating ten recent planar trackers and carry out a detailed analysis of PlanarTrack. By releasing PlanarTrack, we sincerely hope that we can offer the community a dedicated platform for research and applications of planar tracking. In addition, we provide $\text{PlanarTrack}_{\text{BB}}$, a by-product dataset based on PlanarTrack, for studying generic trackers on tracking planar-like target objects. Evaluation results indicate that there is still huge room for future improvement on PlanarTrack and $\text{PlanarTrack}_{\text{BB}}$. For future research, we see several promising directions: (i) robust feature learning under low resolution and light-interactive surfaces, (ii) better temporal modeling for long-term tracking, (iii) integration of multi-modal cues such as depth or inertial data, and (iv) effective re-detection strategies for disappeared objects.

CRediT authorship contribution statement

Yifan Jiao: Writing – review & editing, Writing – original draft, Validation, Software, Data curation. **Xinran Liu:** Software. **Xiaoqiong Liu:** Supervision. **Xiaohui Yuan:** Supervision. **Heng Fan:** Writing – review & editing, Supervision, Project administration. **Libo Zhang:** Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

Libo Zhang was supported by National Natural Science Foundation of China (No. 62476266). Xiaohui Yuan and Heng Fan were not supported by any fund for this work.

Data availability

Data will be made available on request.

References

Baker, S., Matthews, I., 2004. Lucas-Kanade 20 years on: A unifying framework. Int. J. Comput. Vis. 56, 221–255.

Bay, H., Ess, A., Tuytelaars, T., Van Gool, L., 2008. Speeded-up robust features (SURF). Comput. Vis. Image Underst. 110 (3), 346–359.

Benhimane, S., Malis, E., 2004. Real-time image-based tracking of planes using efficient second-order minimization. In: 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566). Vol. 1, IEEE, pp. 943–948.

Cai, Y., Liu, J., Tang, J., Wu, G., 2023. Robust object modeling for visual tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9589–9600.

Chen, B., Li, P., Bai, L., Qiao, L., Shen, Q., Li, B., Gan, W., Wu, W., Ouyang, W., 2022. Backbone is all your need: A simplified architecture for visual object tracking. In: European Conference on Computer Vision. Springer, pp. 375–392.

Chen, X., Peng, H., Wang, D., Lu, H., Hu, H., 2023. Seqtrack: Sequence to sequence learning for visual object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14572–14581.

Chen, X., Yan, B., Zhu, J., Wang, D., Yang, X., Lu, H., 2021. Transformer tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8126–8135.

Chen, L., Zhou, F., Shen, Y., Tian, X., Ling, H., Chen, Y., 2017. Illumination insensitive efficient second-order minimization for planar object tracking. In: 2017 IEEE International Conference on Robotics and Automation. ICRA, IEEE, pp. 4429–4436.

Chum, O., Matas, J., 2005. Matching with PROSAC-progressive sample consensus. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR'05, Vol. 1, IEEE, pp. 220–226.

Comport, A.I., Marchand, É., Chaumette, F., 2003. A real-time tracker for markerless augmented reality. In: The Second IEEE and ACM International Symposium on Mixed and Augmented Reality, 2003. Proceedings. IEEE, pp. 36–45.

Corso, J., Burschka, D., Hager, G., 2003. Direct plane tracking in stereo images for mobile navigation. In: 2003 IEEE International Conference on Robotics and Automation. Vol. 1, IEEE, pp. 875–880.

Cui, Y., Jiang, C., Wang, L., Wu, G., 2022. Mixformer: End-to-end tracking with iterative mixed attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13608–13618.

Cui, Y., Song, T., Wu, G., Wang, L., 2024. Mixformerv2: Efficient fully transformer tracking. Adv. Neural Inf. Process. Syst. 36.

Dick, T., Quintero, C.P., Jägersand, M., Shademan, A., 2013. Realtime registration-based tracking via approximate nearest neighbour search. In: Robotics: Science and Systems.

Erlik Nowruzi, F., Laganiere, R., Japkowicz, N., 2017. Homography estimation from image pairs with hierarchical convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 913–920.

Fan, H., Bai, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Harshit, Huang, M., Liu, J., et al., 2021. Lasot: A high-quality large-scale single object tracking benchmark. Int. J. Comput. Vis. 129, 439–461.

Fan, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Bai, H., Xu, Y., Liao, C., Ling, H., 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. pp. 5374–5383.

Fischler, M.A., Bolles, R.C., 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM 24 (6), 381–395.

Gao, S., Zhou, C., Ma, C., Wang, X., Yuan, J., 2022. Aiatrack: Attention in attention for transformer visual tracking. In: European Conference on Computer Vision. Springer, pp. 146–164.

Gauglitz, S., Höllerer, T., Turk, M., 2011. Evaluation of interest point detectors and feature descriptors for visual tracking. Int. J. Comput. Vis. 94, 335–360.

Guo, M., Zhang, Z., Fan, H., Jing, L., Lyu, Y., Li, B., Hu, W., 2022. Learning target-aware representation for visual tracking via informative interactions. arXiv preprint arXiv:2201.02526.

Hare, S., Saffari, A., Torr, P.H., 2012. Efficient online structured output learning for keypoint-based object tracking. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1894–1901.

Huang, L., Zhao, X., Huang, K., 2019. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. IEEE Trans. Pattern Anal. Mach. Intell. 43 (5), 1562–1577.

Li, K., Liu, H., Wang, T., 2023. Centroid-based graph matching networks for planar object tracking. Mach. Vis. Appl. 34 (2), 31.

Liang, P., Ji, H., Wu, Y., Chai, Y., Wang, L., Liao, C., Ling, H., 2021. Planar object tracking benchmark in the wild. Neurocomputing 454, 254–267.

Liang, P., Wu, Y., Lu, H., Wang, L., Liao, C., Ling, H., 2018. Planar object tracking in the wild: A benchmark. In: 2018 IEEE International Conference on Robotics and Automation. ICRA, IEEE, pp. 651–658.

Lieberknecht, S., Benhimane, S., Meier, P., Navab, N., 2009. A dataset and evaluation methodology for template-based tracking algorithms. In: 2009 8th IEEE International Symposium on Mixed and Augmented Reality. IEEE, pp. 145–151.

Lin, L., Fan, H., Zhang, Z., Xu, Y., Ling, H., 2022. Swintrack: A simple and strong baseline for transformer tracking. Adv. Neural Inf. Process. Syst. 35, 16743–16754.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer, pp. 740–755.

Liu, X., Liu, X., Yi, Z., Zhou, X., Le, T., Zhang, L., Huang, Y., Yang, Q., Fan, H., 2023. PlanarTrack: A large-scale challenging benchmark for planar object tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20449–20458.

Liu, Y., Shen, Z., Lin, Z., Peng, S., Bao, H., Zhou, X., 2019. Gift: Learning transformation-invariant dense visual descriptors via group cnns. Adv. Neural Inf. Process. Syst. 32.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. 60, 91–110.

Matveichev, D., Lin, D.-T., 2021. Mobile augmented reality: Fast, precise, and smooth planar object tracking. In: 2020 25th International Conference on Pattern Recognition. ICPR, IEEE, pp. 6406–6412.

Mayer, C., Danelljan, M., Bhat, G., Paul, M., Paudel, D.P., Yu, F., Van Gool, L., 2022. Transforming model prediction for tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8731–8740.

Mondragón, I.F., Campoy, P., Martinez, C., Olivares-Méndez, M.A., 2010. 3D pose estimation based on planar object tracking for UAVs control. In: 2010 IEEE International Conference on Robotics and Automation. IEEE, pp. 35–41.

Muller, M., Bibi, A., Giancola, S., Alsubaihi, S., Ghanem, B., 2018. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 300–317.

Ozuysal, M., Calonder, M., Lepetit, V., Fua, P., 2009. Fast keypoint recognition using random ferns. IEEE Trans. Pattern Anal. Mach. Intell. 32 (3), 448–461.

Pautrat, R., Larsson, V., Oswald, M.R., Pollefeys, M., 2020. Online invariance selection for local feature descriptors. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. Springer, pp. 707–724.

Peng, L., Gao, J., Liu, X., Li, W., Dong, S., Zhang, Z., Fan, H., Zhang, L., 2024. Vasttrack: Vast category visual object tracking. Adv. Neural Inf. Process. Syst. 37, 130797–130818.

Richa, R., Sznitman, R., Taylor, R., Hager, G., 2011. Visual tracking using the sum of conditional variance. In: 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, pp. 2953–2958.

Rosten, E., Porter, R., Drummond, T., 2008. Faster and better: A machine learning approach to corner detection. IEEE Trans. Pattern Anal. Mach. Intell. 32 (1), 105–119.

Roy, A., Zhang, X., Wolleb, N., Quintero, C.P., Jägersand, M., 2015. Tracking benchmark and evaluation for manipulation tasks. In: 2015 IEEE International Conference on Robotics and Automation. ICRA, IEEE, pp. 2448–2453.

Sarlin, P.-E., DeTone, D., Malisiewicz, T., Rabinovich, A., 2020. Superglue: Learning feature matching with graph neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4938–4947.

Šerỳch, J., Matas, J., 2023. Planar object tracking via weighted optical flow. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1593–1602.

Tan, D.J., Ilic, S., 2014. Multi-forest tracker: A chameleon in tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1202–1209.

Torr, P.H., Zisserman, A., 2000. MLESAC: A new robust estimator with application to estimating image geometry. Comput. Vis. Image Underst. 78 (1), 138–156.

Valmadre, J., Bertinetto, L., Henriques, J.F., Tao, R., Vedaldi, A., Smeulders, A.W., Torr, P.H., Gavves, E., 2018. Long-term tracking in the wild: A benchmark. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 670–685.

Wagner, D., Reitmayr, G., Mulloni, A., Drummond, T., Schmalstieg, D., 2009. Real-time detection and tracking for augmented reality on mobile phones. IEEE Trans. Vis. Comput. Graphics 16 (3), 355–368.

Wang, T., Ling, H., 2017. Gracker: A graph-based planar object tracker. IEEE Trans. Pattern Anal. Mach. Intell. 40 (6), 1494–1501.

Wang, X., Shu, X., Zhang, Z., Jiang, B., Wang, Y., Tian, Y., Wu, F., 2021b. Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13763–13773.

Wang, X., Wang, C., Bai, X., Liu, Y., Zhou, J., 2018. Deep homography estimation with pairwise invertibility constraint. In: Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshop, S+ SSPR 2018, Beijing, China, August 17–19, 2018, Proceedings 9. Springer, pp. 204–214.

Wang, N., Zhou, W., Wang, J., Li, H., 2021a. Transformer meets tracker: Exploiting temporal context for robust visual tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1571–1580.

Wei, X., Bai, Y., Zheng, Y., Shi, D., Gong, Y., 2023. Autoregressive visual tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9697–9706.

Wu, Y., Lim, J., Yang, M.-H., 2013. Online object tracking: A benchmark. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2411–2418.

Wu, Q., Yang, T., Liu, Z., Wu, B., Shan, Y., Chan, A.B., 2023. Dropmae: Masked autoencoders with spatial-attention dropout for tracking tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14561–14571.

Yan, B., Peng, H., Fu, J., Wang, D., Lu, H., 2021. Learning spatio-temporal transformer for visual tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10448–10457.

Ye, B., Chang, H., Ma, B., Shan, S., Chen, X., 2022. Joint feature learning and relation modeling for tracking: A one-stream framework. In: European Conference on Computer Vision. Springer, pp. 341–357.

Zhan, X., Liu, Y., Zhu, J., Li, Y., 2022. Homography decomposition networks for planar object tracking. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36, pp. 3234–3242.

Zhang, H., Ling, Y., 2022. Hvc-net: Unifying homography, visibility, and confidence learning for planar object tracking. In: European Conference on Computer Vision. Springer, pp. 701–718.

Zhang, Z., Liu, S., Yang, J., 2023. Multiple planar object tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23460–23470.

Zhao, L., Li, X., Xiao, J., Wu, F., Zhuang, Y., 2015. Metric learning driven multi-task structured output optimization for robust keypoint tracking. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 29.