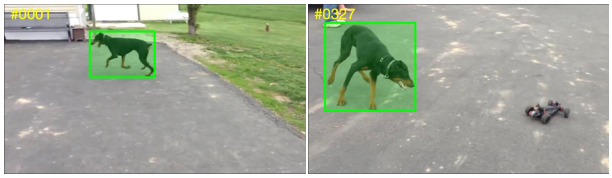
PlanarTrack: A Large-scale Challenging Benchmark for Planar Object Tracking

Xinran Liu[1,3*] Xiaoqiong Liu[2*] Ziruo Yi[2*] Xin Zhou[4*] Thanh Le[2] Libo Zhang[1,3†]
Yan Huang[2] Qing Yang[2] Heng Fan[2]
[1]Institute of Software, Chinese Academy of Sciences, Beijing, China
[2]Department of Computer Science and Engineering, University of North Texas, Denton, USA
[3]University of Chinese Academy of Sciences, Beijing, China
[4]Self-Employed

Abstract

Planar object tracking is a critical computer vision problem and has drawn increasing interest owing to its key roles in robotics, augmented reality, etc. Despite rapid progress, its further development, especially in the deep learning era, is largely hindered due to the lack of large-scale challenging benchmarks. Addressing this, we introduce ***PlanarTrack****, a large-scale challenging planar tracking benchmark. Specifically, PlanarTrack consists of 1,000 videos with more than 490K images. All these sequences are collected in complex unconstrained scenarios from the wild, which makes PlanarTrack, compared with existing benchmarks, more challenging but realistic for real-world applications. To ensure the high-quality annotation, each frame in PlanarTrack is manually labeled using four corners with multiple-round careful inspection and refinement. To our best knowledge, PlanarTrack, to date, is the largest and the most challenging dataset dedicated to planar object tracking. In order to analyze the proposed PlanarTrack, we evaluate 10 planar trackers and conduct comprehensive comparisons and in-depth analysis. Our results, not surprisingly, demonstrate that current top-performing planar trackers degenerate significantly on the challenging PlanarTrack and more efforts are needed to improve planar tracking in the future. In addition, we further derive a variant named* ***PlanarTrack***$_{\mathbf{BB}}$ *for generic object tracking from our PlanarTrack. Our evaluation of 10 excellent generic trackers on PlanarTrack*$_{\mathrm{BB}}$ *manifests that, surprisingly, PlanarTrack*$_{\mathrm{BB}}$ *is even more challenging than several popular generic tracking benchmarks and more attention should be paid to handle such planar objects, though they are rigid. All benchmarks and evaluations are released at* https://hengfan2010.github.io/projects/PlanarTrack/.



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

●, ●, ●, and ● represent the first, second, third and fourth corner point of the planar object

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

Figure 1. 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.*

1. Introduction

Planar tracking is an important problem in computer vision. Different than generic tracking which represents object with an axis-aligned bounding box, planar tracking represents target with four corners. Besides, the goal of generic tracking is to locate the target with axis-aligned rectangular bounding boxes [10, 38], while planar object tracking aims to estimate 2D transformations (*e.g.*, homograph) of the target and locate it with corner points (see Fig. 1). Owing to its importance in robotics and augmented reality (AR), planar object tracking has attracted increasing attentions in recent years. In particular, several benchmarks (*e.g.*, [20, 31, 19]) have been specially developed for evaluating and comparing different planar trackers, which greatly facilitates related research and progress on this topic. Despite this, they are limited in further pushing the frontier of planar tracking.

One of the major issues with existing benchmarks is their

*The authors make equal contributions and are co-first authors.
†Corresponding author: Libo Zhang (libo@iscas.ac.cn).

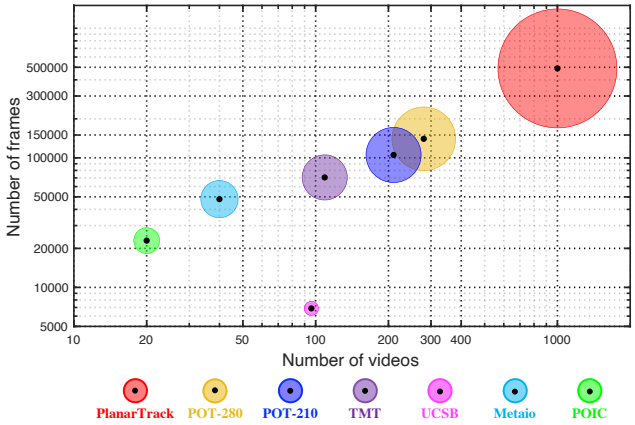


Figure 2. Summary of planar object tracking datasets, containing POT-280 [19], POT-210 [20], TMT [31], UCSB [13], Metiao [21], POIC [5], and PlanarTrack. The circle diameter is in proportion to the number of frames of a dataset. Our PlanarTrack is the *largest*.

relatively small scales. Especially, in the deep learning era, to unleash the potential of deep planar tracking, it is desired to have a large-scale platform. Nevertheless, as displayed in Fig. 2, currently all planar tracking benchmarks consist of *less than* 300 sequences, which is *insufficient* for large-scale learning of deep planar tracking. As a consequence, researchers are forced to leverage synthetic data generated from images (*e.g.*, [23]) for transformation learning in deep planar tracking, which may result in inferior performance due to domain gap between different tasks.

Besides the small-scale issue, another problem is the less challenging scenarios for planar object tracking. Early planar tracking datasets (*e.g.*, [21, 31, 13, 5]) are constructed from the indoor laboratories with simple background, which cannot reflect the diverse and complicated scenarios of real world in performance evaluation. To deal with this, recent datasets (*e.g.*, [20, 19]) directly collect videos in the wild. However, most of these videos are mainly involved with one challenge factor (or *attribute* in generic tracking), and very few (*e.g.*, 30 in [20] and 40 in [19]) contain multiple challenges (*i.e.*, the unconstrained condition). This may weaken the difficulties of planar tracking in the wild where arbitrary challenges could exist, and thus restricts datasets in assessing generalization of planar tracking in natural scenarios.

Furthermore, the diversity of current benchmarks is limited. In particular, the same planar target object is usually employed in multiple videos, which significantly decreases the diversity in target appearance. Even for current largest benchmark [19] (one target used in 7 videos), the number of planar targets does not exceed 40 (see Tab. 1). Such lack of diversity makes it difficult to use current datasets for faithful assessment of planar trackers in practice.

We are aware that there exist several large-scale datasets (*e.g.*, [27, 10, 16]) for generic tracking. Nevertheless, due to different setting and goal (see Fig. 1 again), these generic datasets are *not* suitable for planar tracking. To further facilitate research on deep planar tracking, a dedicated large-scale benchmark is desired, which motivates our work.

1.1. Contributions

In this paper, we propose a novel large-scale benchmark, dubbed **PlanarTrack**, dedicated for planar object tracking. Specifically, PlanarTrack consists of 1,000 video sequences. *All* these videos are directly collected in complicated *unconstrained* scenarios from the wild, which makes PlanarTrack, compared to existing datasets (*e.g.*, [13, 21, 5, 31, 20, 19]), much more challenging yet realistic for real applications. In order to diversify our PlanarTrack, each planar object appears exclusively in one video, which is different than other datasets. In total, there are over 490K frames in our PlanarTrack, and each one is manually labeled using four corner points[1] with cautious inspections and refinements to ensure high-qualify annotations. Besides, we offer challenge factor information for each video as in generic tracking [38] to enable in-depth analysis. To our best knowledge, PlanarTrack, to date, is the *largest* and *most challenging* planar tracking dataset. By releasing PlanarTrack, we aim to provide a dedicated platform for facilitating planar trackers.

In order to analyze PlanarTrack and provide comparisons for future research, we evaluate 10 representative planar object trackers. Our evaluation exhibits that, not surprisingly, existing top-performing planar trackers severely degrade on more challenging PlanarTrack. For example, the precision (PRE) score (as described later) of WOFT [32] on POT-210 is 0.805 but drops to 0.433 on PlanarTrack, and the score of HDN [41] drops from 0.612 on POT-210 to 0.263 on PlanarTrack. This consistently reveals the difficulties for planar tracking brought by realistic complicated scenes, and more efforts are required for improvements. To provide guidance for future research, we further conduct comprehensive analysis to analyze challenges in planar tracking and discuss potential directions to facilitate related research. Besides, our re-training experiments show the usefulness and effectiveness of our benchmark in performance enhancement.

Furthermore, as a by-product of PlanarTrack, we develop a new variant, **PlanarTrack**$_{\mathrm{BB}}$, which is suitable for generic box tracking. We aim at *large-scale* learning and evaluation of generic object trackers on localizing *rigid* targets, which is rarely investigated before. Our experiments on assessing 10 recent Transformer-based generic trackers reveals heavy performance degeneration on PlanarTrack$_{\mathrm{BB}}$ compared with their performance on large-scale generic tracking datasets (*e.g.*, LaSOT [10] and TrackingNet [27]) and more attention is needed in handling planar objects, though they are rigid.

In summary, our main contributions are as follows:

◇ *We introduce a novel benchmark termed PlanarTrack*

[1]Four points are the least number of points to determine the homograph of two planar objects, which is the reason to use four points for annotation.

Table 1. Detailed comparison of the proposed PlanarTrack with other existing planar object tracking benchmarks.

Benchmark	Year	Targets	Videos	Min frames	Mean frames	Max frames	Total frames	Annotated frames	Unconstrained Videos	In the wild
Metaio [21]	2009	8	40	1,200	1,200	1,200	48K	48K	n/a	✗
UCSB [13]	2011	6	96	13	72	500	7K	7K	n/a	✗
TMT [31]	2015	12	109	191	648	2,518	71K	71K	n/a	✗
POIC [5]	2017	20	20	283	1,149	2,666	23K	23K	n/a	✗
POT-210 [20]	2018	30	210	501	501	501	105K	53K	30	✓
POT-280 [19]	2021	40	280	501	501	501	140K	70K	40	✓
PlanarTrack (ours)	**2023**	**1,000**	**1,000**	**317**	**490**	**549**	**490K**	**490K**	**1,000**	✓

for planar tracking. To the best of our knowledge, PlanarTrack is to date the largest as well as the most challenging planar tracking benchmark in the wild.

- ◇ *We conduct comprehensive evaluations to analyze PlanarTrack and provide comparison for future research.*
- ◇ *We conduct retraining experiments to validate the effectiveness of the proposed PlanarTrack in improving deep planar tracking performance.*
- ◇ *Based on PlanarTrack, we develop PlanarTrack$_{\mathrm{BB}}$ for generic tracking on planar-like targets and conduct extensive evaluation and analysis.*

2. Related Work

2.1. Planar Tracking Benchmarks

Datasets have played an important role in facilitating the development of planar object tracking. **Metaio** [21] is one of the earliest datasets for planar tracking. It comprises 40 videos with eight different textures using a camera mounted on the robotic measurement arm. **UCSB** [13] contains 96 videos for investigating interest point detectors and feature descriptors for planar object tracking. **TMT** [31] consists of 109 videos and each one is labeled with a challenging factor. The goal is to evaluate different planar tracking algorithms for human and robot manipulation tasks. **POIC** [5] provides 20 sequences and mainly focuses on evaluating the performance of planar trackers in complicated illumination environments. In order to assess the planar tracking performance in the wild, **POT-210** [20] collects 210 videos of 30 planar objects from natural scenarios. Later in [19], POT-210 is further extended to **POT-280** by introducing 70 extra videos of 10 planar targets. For each planar object in POT [20, 19], seven videos are captured, however, six of them simply comprise one challenge and only one contains multiple challenges in unconstrained conditions.

Despite the above benchmarks, the further development of planar object tracking, especially in the deep learning, is limited due to lacking a large-scale, challenging and diverse platform, which motivates our PlanarTrack, the *largest* and most *challenging* and *diverse* planar tracking benchmark to date. Tab. 1 displays a detailed comparison of PlanarTrack with existing planar tracking benchmarks.

2.2. Planar Tracking Algorithms

The goal of planar tracking is to estimate the homograph. Current approaches can be roughly divided into three types: keypoint methods, direct method and deep regression methods. Keypoint-based planar trackers (*e.g.*, [8, 28, 35]) first detect the keypoints (*e.g.*, SIFT [25] or SURF [2]) of objects and then estimate homograph using these interesting points. Direct methods [3, 30, 5] aim to directly calculate the homograph by optimizing the alignment of current frame with object of initial frame. In addition to the above two types, another recent trend is to employ the deep neural networks to regress the homograph. These deep regression-based planar trackers [41, 42, 32] avoid complex keypoint feature extraction and can be trained in an end-to-end fashion. Due to outstanding performance, the deep regression-based methods have attracted increasing attentions in planar tracking.

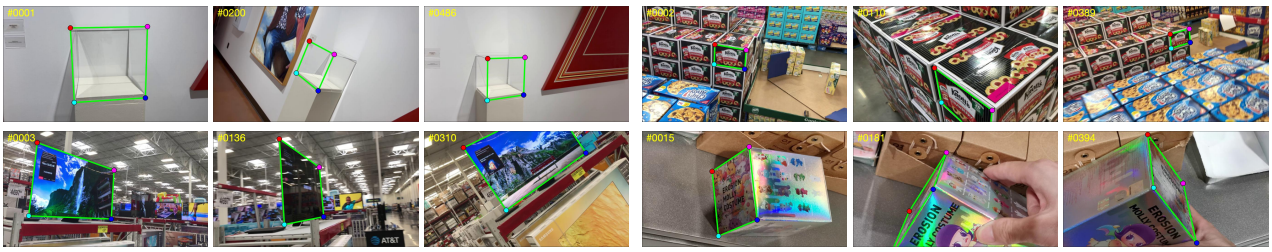
2.3. General Tracking Benchmarks

There are many benchmarks developed for generic tracking. Some early representatives include OTB [38, 37], TC-128 [18], VOT challenge [17], etc. However, these datasets are usually small in scale. To further facilitate development of tracking, large-scale tracking benchmarks have recently been introduced, including GOT-10k [16], LaSOT [10, 9], TrackingNet [27], OxUvA [33], and TNL2K [36]. These large-scale benchmarks greatly push the start-of-the-arts in visual tracking, which to some degree motives the development of PlanarTrack. But, different from existing general tracking benchmarks, the proposed PlanarTrack is specially developed for planar object tracking. For this goal, we provide annotations of corner points in PlanarTrack for targets instead of bounding boxes in aforementioned datasets.

3. The Proposed PlanarTrack Benchmark

3.1. Design Principle

PlanarTrack in this work expects to provide a large-scale platform for developing deep planar tracking and to offer a more challenging and faithful testbed for evaluating planar



●, ●, ●, and ● represent the first, second, third and fourth corner point of the planar object

Figure 3. Examples of annotated sequences in the proposed PlanarTrack. Each video is annotated with four corner points.

trackers in practice. To meet these requirements, we follow four rules in constructing our PlanarTrack:

- *Dedicated large-scale benchmark.* One important motivation for our work is to facilitate deep planar tracking with a large-scale dedicate benchmark. To this end, we hope to collect 1,000 videos with over 450K frames in the new benchmark.
- *Realistic challenge in the wild.* To faithfully reflect the performance of planar trackers in practice, it is crucial to collect videos with realistic challenges. For this purpose, we require all videos in the benchmark captured from natural scenarios in unconstrained conditions.
- *Diverse planar objects.* The diversity of targets is beneficial for assessing the generalization of planar trackers. Considering this, the planar targets in the videos should be unique, which differs from current datasets.
- *High-quality dense annotation.* The annotation is crucial for both training and evaluation. For this, we manually label every frame in PlanarTrack with careful refinement to ensure its high-quality annotations.

3.2. Video Collection

We construct PlanarTrack starting by collecting videos. Different from generic tracking benchmarks (*e.g.*, [10, 16, 27]) sourcing videos from YouTube, we collect sequences from natural scenarios using smart phones as we observe the videos from YouTube seldom focus on the motion of planar objects. To diversify the video sources, we invite volunteers who are familiar with this task to record the sequences using different phones with different resolutions. With the above principles in mind, we include a wide selection of the planar targets (*e.g.*, *box*, *poster*, *picture*, *board*, *logo*, *door*, *mirror*, *book*, *traffic sign*, *tile*, *wall*, *tile*, *screen*, and *table*) for video recording, and each sequence is captured in unconstrained conditions from various natural scenes (*e.g.*, *shopping mall*, *street*, *library*, *restaurant*, *supermarket*, *playground*, *park*, *museum*, *apartment*, *hall*, and *classroom*).

Initially, we collected over 2,500 videos. After a careful inspection conducted by a few experts (PhD students working on related topics), we choose 1,000 available videos for developing PlanarTrack. It is worth noticing that, for these 1,000 videos, we further verify their contents and remove inappropriate parts to make sure they are suitable for planar tracking. Eventually, we compile a dataset dedicated for planar tracking by including 1,000 unconstrained sequences with more than 490K frames from 1,000 unique planar objects. Tab. 1 provides a detailed summary of PlanarTrack and its comparison with other planar tracking benchmarks.

3.3. Annotation

To offer high-quality annotation in PlanarTrack, we manually label each frame. Specifically, for each image, we annotate four corner points for the planar target if all its four corner points or four edges are clearly visible to. Otherwise, if the four corner points and four edges are both not available due to occlusion or out-of-view, or, the planar target is severely blurred, we will assign an absent flag to this frame.

With the above strategy, we assemble a team with several experts and volunteers for annotation. Each sequence is first annotated by a volunteer. Then, the annotation result will be sent to two experts for verification. If the annotation is not unanimously agreed by the experts, it will be returned back the original annotator for careful refinement. To ensure the high annotation quality, the verification-refinement process may last for multiple rounds until the final annotation result passes the inspection. We demonstrate some annotation examples of PlanarTrack in Fig. 3.

To verify the annotation quality, we randomly select 20 frames in each labeled video and ask different annotators to improve annotations. The standard deviation of the corner distances using existing and newly improved annotations is 0.28 (pretty small) with an almost zero mean, verifying the high quality of our annotations.

Statistics of annotations. In order to better understand the planar targets in PlanarTrack, we show representative statistics of the annotations in Fig. 4. In particular, we display the distributions of target motion, target size, relative area to the initial object and Intersection over Union (IoU) between targets in adjacent frames. From Fig. 4, we see that the planar targets vary rapidly in size and temporal motions. Besides, Fig. 4 also compares our PlanarTrack and the re-

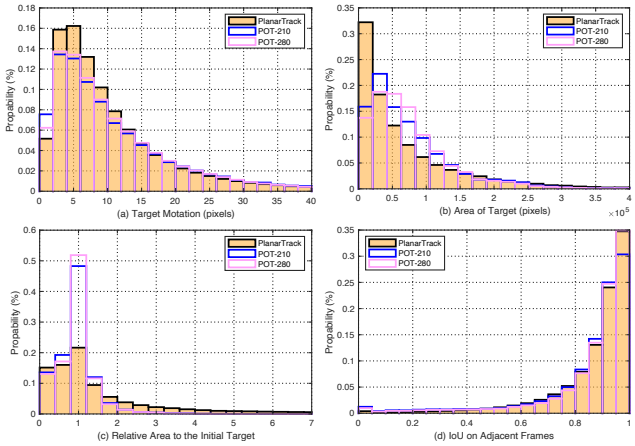


Figure 4. Statistics of planar target motion, size, relative area compared to initial object and IoU of targets in adjacent frames in PlanarTrack and comparison with the recent POT-210/280 [20, 19]. We can see the targets in our dataset have smaller sizes and faster and more challenging motions.

cent POT-210/280 [20, 19]. Notice that, since POT-210/280 are labeled every two frames, we perform linear interpolation on their annotation for the comparison purpose. From Fig. 4, we can see that the targets in PlanarTrack are relatively smaller and moving faster, which consequently leads to new challenges for planar tracking in the wild.

3.4. Challenging Factors

Following other tracking datasets [38, 20, 11], we provide challenging factors (also called *attributes* in other datasets) for each sequence in PlanarTrack to enable further in-depth analysis of different algorithms. In specific, we define eight challenging factors that widely exist for planar tracking and annotate each sequence with these factors, including (1) occlusion (OCC), which is assigned when any part of the target object is occluded, (2) motion blur (MB), (3) rotation (ROT), (4) scale variation (SV), which is assigned when the ratio of planar annotation is outside the range [0.5, 2], (5) perspective distortion (PD), which is assigned when the perspective between the object and camera is changed, (6) out-of-view (OV), (7) low resolution (LR), which is assigned when the region of the target planar is less than 1,000 pixels, and (8) background clutter (BC), which is assigned when the background region looks visually similar to the target. It is worthy to note that, we exclude a few common challenging factors used in generic object tracking such as deformation and illumination change because they are not suitable for planar targets. Particularly for illumination variation, it happens usually due to the severely varied lighting conditions, which are hard in the natural environment. Each video in PlanarTrack may simultaneously contain multiple challenging factors (*i.e.*, recorded in *unconstrained condition*), which is, compared to POT-210/280, more practical for real applications.

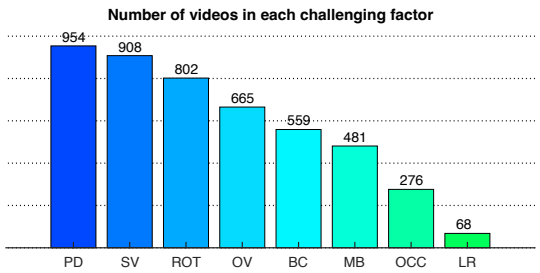


Figure 5. Distribution of sequences on each challenging factor.

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

	Videos	Min frames	Mean frames	Max frames	Total frames
PlanarTrack$_{\mathrm{Tst}}$	300	346	493	534	148K
PlanarTrack$_{\mathrm{Tra}}$	700	317	489	549	342K

The distribution of the aforementioned challenging factors on PlanarTrack is presented in Fig. 5. We observe that the most common challenging factor in PlanarTrack is perspective distortion, which may cause serious misalignment problem for planar tracking. In addition, scale variation and rotation frequently happen in the sequences.

3.5. Dataset Split and Evaluation Metric

Training/Test Split. PlanarTrack consists of 1,000 videos. We use 700 sequences for training (PlanarTrack$_{\mathrm{Tra}}$) and the rest 300 for evaluation (PlanarTrack$_{\mathrm{Tst}}$). We try our best to keep the distributions of training and test sets close to each other. Tab. 2 shows the comparison of these two sets, and please see *supplementary material* for challenge-wise comparisons. The split will be released at our project website.

Evaluation Metric. For the evaluation, we follow [20] and adopt the *precision* (PRE) and *success* (SUC) metrics. It is worthy to notice, the PRE and SUC differ from those used for generic tracking [38]. Specifically, for planar tracking, the PRE is defined as the percentage of frames where alignment error between the corner points of tracking result and groundtruth is within a given threshold (*e.g.*, typically 5 pixels). The SUC is calculated by the percentage of successful frames in which the discrepancy between estimated and real homography is smaller than or equal to a certain threshold. We set the threshold to 30 in our evaluation as the threshold of 10 in [20] is too tight. For more details of PRE and SUC for planar tracking evaluation, please kindly refer to [20].

4. Experiments on PlanarTrack

4.1. Evaluated Planar Trackers

Since there are not many planar object trackers compared to generic tracking (in fact, it motivates us to introduce PlanarTrack for fostering research on planar object tracking),

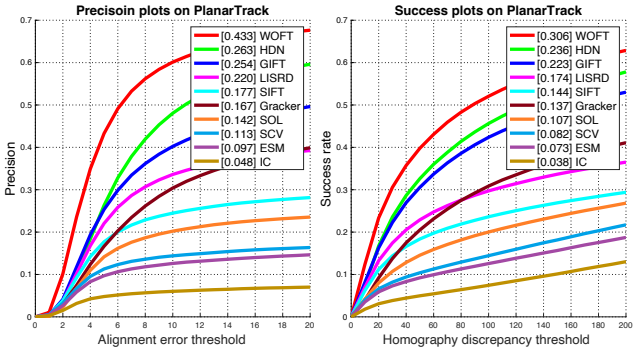
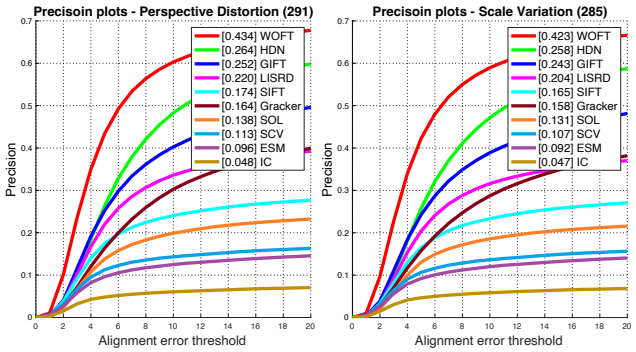


Figure 6. Overall performance on PlanarTrack$_{\mathbf{Tst}}$ in terms of precision (left) and success (right).

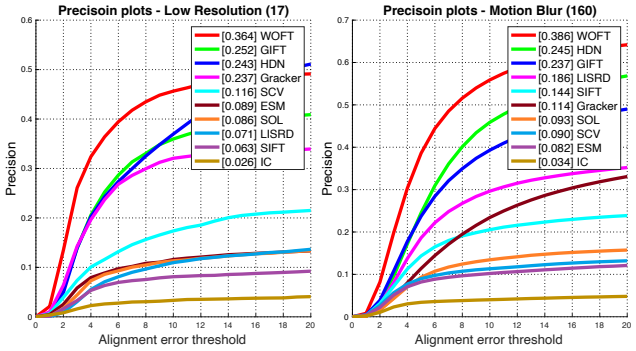
we select 10 representative algorithms with available source codes consisting of two very recent ones. Specifically, these trackers are Gracker [35], GIFT [24], ESM [3], LISRD [29], SOL [15], SIFT [25], IC [1], SCV [30], HDN [41], and WOFT [32]. Particularly, the HDN [41] and WOFT [32] are two recently specially developed planar trackers using deep learning. Notice that, we do not evaluate generic trackers on our PlanarTrack due to incompatible inputs and tracking results. Instead, we will create a new PlanarTrack$_{\mathrm{BB}}$ suitable for generic tracking evaluation, as described later.

4.2. Evaluation Results

Overall Performance. We evaluate 10 typical planar object trackers on the test set of PlanarTrack. Please note that, the methods of HDN and WOFT are utilized without modifications in our evaluation as they are specifically developed for the planar tracking task. For all other approaches, they are customized to achieve the planar tracking. Their implementations except for LISRD and GIFT are borrowed from [20], and we adapt LISRD and GIFT to planar tracking because of some setting problems provided by [20]. The evaluation results of these approaches are reported in Fig. 6 using precision (PRE) and success (SUC). From Fig. 6, we can observe that WOFT demonstrates the best PRE score of 0.433 and SUC score of 0.306, and HDN shows the second best PRE score of 0.263 and SUC score of 0.236. Both WOFT and HDN are recent planar trackers which formulate planar tracking as a deep homography estimation problem. Compared with HDN, WOFT introduces the optical flow into homography estimation and effectively boosts the robustness of tracking, which exhibits the importance of video temporal information for tracking. The method of GIFT applies transformation-invariant deep visual descriptors for planar tracking and achieves the third best of PRE score of 0.254 and SUC score of 0.233. It is worth mentioning that, all the top four trackers leverage deep neural networks for planar target localization, which demonstrates the great potential of deep planar tracking in the future. This is also the motivation of our work to offer a dedicated large-scale platform for developing deep planar trackers.



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

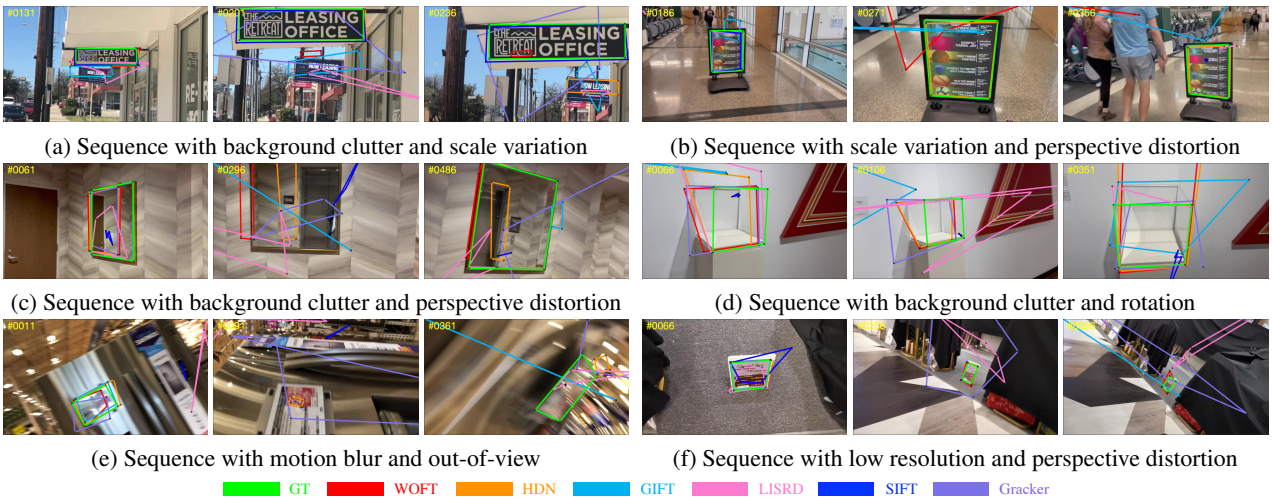


(b) Evaluation on the two most difficult challenging factors using precision

Figure 7. Performance evaluation 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 *motion blur* using precision (Please refer to *supplementary material* for full results and comparisons).

Challenging Factor-based Evaluation. For in-depth analysis of different planar trackers, we further conduct evaluation on the eight challenging factors. Due to limited space, we display the results on the two most common challenging factors including *perspective distortion* (PD) and *scale variation* (SV) and on the two most difficult challenging factors including *low resolution* (LR) and *motion blur* (MB) in Fig. 7, and refer reader to *supplementary material* for more results. From Fig. 7, we can observe that WOFT shows the best performance on both the commonest and most difficult challenges. In specific, it achieves the PRE scores of 0.434, 0.423, 0.364 and 0.386 on PD, SV, LR and MB, which outperform HDN, the second best on PD, SV and MB with PRE scores of 0.264, 0.258 and 0.252, and GIFT, the second best on LR with 0.252 PRE score. This again demonstrates the importance of temporal information for planar tracking. In addition, the tracking performance severely degrades on LR and MB. We argue that these two challenges may result in ineffective feature extraction of points or targets, causing tracking drifts or failures. Future research can be devoted to improvements in these two situations.

Qualitative Results. To better understand the planar track-



(a) Sequence with background clutter and scale variation

(b) Sequence with scale variation and perspective distortion

(c) Sequence with background clutter and perspective distortion

(d) Sequence with background clutter and rotation

(e) Sequence with motion blur and out-of-view

(f) Sequence with low resolution and perspective distortion

Figure 8. 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.

Table 3. Comparison of PlanarTrack$_{\mathbf{Tst}}$ to POT-210 [20] and its subset POT-210$_{\mathrm{UC}}$ in unconstrained condition on PRE and SUC. Note that, the threshold for SUC is set to the same 30 for all experiments for fair comparison.

		WOFT [32]	HDN [41]	GIFT [24]	LISRD [29]	SIFT [25]	Gracker [35]	SOL [15]	SCV [30]	ESM [3]	IC [1]
POT-210 [20]	PRE	0.805	0.612	0.553	0.617	0.692	0.392	0.417	0.228	0.204	0.121
	SUC	0.572	0.484	0.404	0.463	0.445	0.331	0.312	0.200	0.183	0.114
POT-210$_{\mathrm{UC}}$ [20]	PRE	0.768	0.567	0.528	0.581	0.578	0.185	0.289	0.105	0.100	0.053
	SUC	0.536	0.442	0.379	0.419	0.378	0.195	0.224	0.092	0.086	0.050
PlanarTrack$_{\mathbf{Tst}}$	PRE	0.433	0.263	0.254	0.167	0.142	0.121	0.113	0.097	0.064	0.048
	SUC	0.306	0.236	0.223	0.137	0.107	0.098	0.082	0.073	0.147	0.038

ing algorithms, we show qualitative results of top six trackers with the highest precision scores, consisting of WOFT, HDN, GIFT, LISRD, SIFT, and Gracker, in different challenges, including *background clutter*, *scale variation*, *perspective distortion*, *motion blur*, *rotation*, *out-of-view* and *low resolution* in Fig. 8. As in Fig. 8, we can see that although some trackers can deal with certain challenging factor. However, when multiple challenging factors occur simultaneously, the trackers may drift or even lose the target.

4.3. Comparison with POT-210.

POT-210 [20] is currently one of the most popular benchmarks for planar object tracking. However, most sequences in POT-210 contain mainly one challenging factors and very few (*i.e.*, 30) are involved with different challenges, which may not faithfully reflect the evaluation in real scenarios. In addition, the lack of diversity in planar targets also limits its usage. To mitigate these, all the videos in PlanarTrack are freely recorded in unconstrained conditions and the targets are unique in each sequence for diversity. Consequently, our PlanarTrack is more challenging and realistic in practice.

To verify the above, we compare existing planar trackers on POT-210 and PlanarTrack$_{\mathbf{Tst}}$. Tab. 3 shows the comparison results. From Tab. 3, we can see that the best performing tracker on POT-210 is WOFT that achieves 0.805/0.572 PRE/SUC scores. Nevertheless, when utilized for tracking planar targets on PlanarTrack$_{\mathbf{Tst}}$, its performance is severely degenerated. In specific, the PRE/SUC scores are decreased from 0.805/0.572 to 0.433/0.306, showing absolute perform drop of 37.2%/26.6% in PRE/SUC. Besides, SIFT with the second best PRE score of 0.692 on POT-210 heavily degrades to 0.142 on PlanarTrack$_{\mathbf{Tst}}$, and HDN with the second best SUC score of 0.484 to 0.236. Furthermore, other trackers are degenerated as well on PlanarTrack$_{\mathbf{Tst}}$.

In addition to POT-210, we further compare POT-210$_{\mathrm{UC}}$, a subset of POT-210 with all videos in unconstrained conditions, with PlanarTrack$_{\mathbf{Tst}}$ as they are both unconstrained. The comparisons are shown in Tab. 3. As in Tab. 3, we can see that POT-210$_{\mathrm{UC}}$ is more challenging than POT-210, yet less difficult than PlanarTrack. The best tracker WOFT on POT-210$_{\mathrm{UC}}$ demonstrates PRE/SUC scores of 0.786/0.536, while it degrades to 0.433/0.306 on PlanarTrack$_{\mathbf{Tst}}$ with performance drop of 35.3% and 23.0%.

Through the above comparisons and analysis, we clearly see that PlanarTrack is more challenging and complex, and there is still a big room for improvements.

Table 4. Retraining of HDN [41] using PlanarTrack$_{\mathbf{Tra}}$.

		Original HDN [41]	Retrained HDN
POT-210 [20]	PRE	0.612	0.637 (+2.5%)
	SUC	0.484	0.497 (+1.3%)
PlanarTrack$_{\mathbf{Tst}}$	PRE	0.263	0.294 (+3.1%)
	SUC	0.236	0.260 (+2.4%)

4.4. Retraining on PlanarTrack

One of the major goals for our PlanarTrack is to provide a dedicated platform for developing deep planar trackers. To validate its effectiveness, we conduct retraining experiments using PlanarTrack$_{\text{Tra}}$ instead of the synthetic data on the recent HDN. Please notice that, we do not perform retraining on WOFT because it does not provide the training implementation. In the retraining, the parameters and settings are kept the same as in the original approach. Tab. 4 demonstrates the results of the retraining experiment. From Tab. 4, we can observe clearly that, when leveraging task-specific data for training, the performance of planar tracker is significantly increased. In specific, the PRE/SUC scores are increased from 0.612/0.484 to 0.637/0.495 on POT-210 and from 0.263/0.236 to 0.294/0.260 on our PlanarTrack$_{\mathbf{Tst}}$, which demonstrates the effectiveness of PlanarTrack.

5. PlanarTrack$_{\text{BB}}$ and Experiments

Planar objects are common to see in our daily life. However, localization of planar targets with *generic visual trackers* has rarely been studied at large scale, even in the existing large-scale generic tracking benchmarks (*e.g.*, [10, 16, 27]). For generic trackers, they should locate the targets regardless of their categories. To discover the capacities of these generic trackers in handling planar-like targets, we design PlanarTrack$_{\text{BB}}$, a by-product of PlanarTrack. Specifically, PlanarTrack$_{\text{BB}}$ shares the same sequences and dataset split from PlanarTrack but converts four annotated corner points to an axis-aligned bounding box in each frame, and it is specially used for evaluation of generic trackers in locating planar-like targets. Please refer to *supplementary material* for detailed construction of PlanarTrack$_{\text{BB}}$ and examples.

We select ten state-of-the-art generic trackers for evaluation. Notice that, these trackers are all Transformer-based, consisting of SwinTrack [22], OStrack [40], SimTrack [4], MixFormer [7], AiATrack [12], ToMP [26], STARK [39], TransInMo [14], TransT [6] and TrDiMP [34], and the best version of each visual tracker is employed for evaluation with SUC$_{\text{BB}}$ which is success score for bounding box-based tracking [38] (please note, besides SUC$_{\text{BB}}$, other metrics such as EAO in VOT [17] can be adopted for evaluation, and we will consider this in the future due to limited time). Tab. 5 reports the evaluation results and comparisons with other large-scale generic tracking benchmarks including LaSOT [10] and TrackingNet [27]. Notice, GOT-10k [16] is not included for comparison because it adopts a different evaluation metric. From Tab. 5, we can observe that although existing generic trackers achieve outstanding performance, they are heavily degraded when dealing with planar-like target objects. For example, the top-performing generic trackers SwinTrack ans OStrack obtain 0.713/0.840 and 0.701/0.839 SUC scores on LaSOT/TrackingNet, while degrade 0.663 and 0.648, respectively, on PlanarTrack$_{\text{BB}}$, which indicates that more attention should be paid to handle such planar trackers, though they are rigid. Due to limited space, please see *supplementary material* for more results.

Table 5. Evaluation of generic trackers on PlanarTrack$_{\text{BB}}$ and comparison with other popular generic benchmarks using SUC$_{\text{BB}}$.

	TrackingNet [27]	**LaSOT** [10]	**PlanarTrack$_{\text{BB}}$** (ours)
SwinTrack [22]	0.840	0.713	0.663
MixFormer [7]	0.839	0.701	0.657
OStrack [40]	0.839	0.711	0.648
TransInMo [14]	0.817	0.657	0.636
AiATrack [12]	0.827	0.690	0.624
STARK [39]	0.820	0.671	0.618
TransT [6]	0.814	0.649	0.608
SimTrack [4]	0.834	0.705	0.606
ToMP [26]	0.815	0.685	0.605
TrDiMP [34]	0.784	0.639	0.584

6. Conclusion and Limitation

In this work, we introduce a new benchmark named PlanarTrack. PlanarTrack consists of 1,000 videos collected in unconstrained conditions from natural scenes, and has more than 490K image frames. To our best knowledge, PlanarTrack is, to date, the first large-scale challenging dataset for planar tracking. To understand existing methods on PlanarTrack and provide comparison for future research, we perform experiments by evaluating ten representative planar trackers and conduct in-depth analysis. By releasing PlanarTrack, we expect to facilitate research and applications of planar tracking. Furthermore, we develop a by-produce dataset, dubbed PlanarTrack$_{\text{BB}}$, based on PlanarTrack for studying generic trackers on localizing planar-like targets.

Despite contributions, there are limitations of this work. First, given the propose large-scale PlanarTrack, a baseline that outperforms other planar trackers is not provided. Second, since videos in PlanarTrack are relatively short, they may not be suitable for long-term tracking. Considering our aim is to make the first attempt for large-scale planar tracking, we keep these as open questions for future research.

Acknowledgement. Libo Zhang was supported by Youth Innovation Promotion Association, CAS (2020111). Heng Fan and his employer received no financial support for this work. We thank Helen Li for collecting part of videos.

References

[1] Simon Baker and Iain Matthews. Lucas-kanade 20 years on: A unifying framework. *IJCV*, 56:221–255, 2004. 6, 7

[2] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *ECCV*, 2006. 3

[3] Selim Benhimane and Ezio Malis. Real-time image-based tracking of planes using efficient second-order minimization. In *IROS*, 2004. 3, 6, 7

[4] Boyu Chen, Peixia Li, Lei Bai, Lei Qiao, Qiuhong Shen, Bo Li, Weihao Gan, Wei Wu, and Wanli Ouyang. Backbone is all your need: a simplified architecture for visual object tracking. In *ECCV*, 2022. 8

[5] Lin Chen, Fan Zhou, Yu Shen, Xiang Tian, Haibin Ling, and Yaowu Chen. Illumination insensitive efficient second-order minimization for planar object tracking. In *ICRA*, 2017. 2, 3

[6] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *CVPR*, 2021. 8

[7] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. Mixformer: End-to-end tracking with iterative mixed attention. In *CVPR*, 2022. 8

[8] Travis Dick, Camilo Perez Quintero, Martin Jägersand, and Azad Shademan. Realtime registration-based tracking via approximate nearest neighbour search. In *RSS*, 2013. 3

[9] Heng Fan, Hexin Bai, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Mingzhen Huang, Juehuan Liu, Yong Xu, et al. Lasot: A high-quality large-scale single object tracking benchmark. *IJCV*, 129:439–461, 2021. 3

[10] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *CVPR*, 2019. 1, 2, 3, 4, 8

[11] Heng Fan, Halady Akhilesha Miththanthaya, Siranjiv Ramana Rajan, Xiaoqiong Liu, Zhilin Zou, Yuewei Lin, Haibin Ling, et al. Transparent object tracking benchmark. In *ICCV*, 2021. 5

[12] Shenyuan Gao, Chunluan Zhou, Chao Ma, Xinggang Wang, and Junsong Yuan. Aiatrack: Attention in attention for transformer visual tracking. In *ECCV*, 2022. 8

[13] Steffen Gauglitz, Tobias Höllerer, and Matthew Turk. Evaluation of interest point detectors and feature descriptors for visual tracking. *IJCV*, 94:335–360, 2011. 2, 3

[14] Mingzhe Guo, Zhipeng Zhang, Heng Fan, Liping Jing, Yilin Lyu, Bing Li, and Weiming Hu. Learning target-aware representation for visual tracking via informative interactions. In *IJCAI*, 2022. 8

[15] Sam Hare, Amir Saffari, and Philip HS Torr. Efficient online structured output learning for keypoint-based object tracking. In *CVPR*, 2012. 6, 7

[16] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *TPAMI*, 43(5):1562–1577, 2021. 2, 3, 4, 8

[17] Matej Kristan, Jiri Matas, Aleš Leonardis, Tomáš Vojíř, Roman Pflugfelder, Gustavo Fernandez, Georg Nebehay, Fatih Porikli, and Luka Čehovin. A novel performance evaluation methodology for single-target trackers. *TPAMI*, 38(11):2137–2155, 2016. 3, 8

[18] Pengpeng Liang, Erik Blasch, and Haibin Ling. Encoding color information for visual tracking: Algorithms and benchmark. *TIP*, 24(12):5630–5644, 2015. 3

[19] Pengpeng Liang, Haoxuanye Ji, Yifan Wu, Yumei Chai, Liming Wang, Chunyuan Liao, and Haibin Ling. Planar object tracking benchmark in the wild. *Neurocomputing*, 454:254–267, 2021. 1, 2, 3, 5

[20] Pengpeng Liang, Yifan Wu, Hu Lu, Liming Wang, Chunyuan Liao, and Haibin Ling. Planar object tracking in the wild: A benchmark. In *ICRA*, 2018. 1, 2, 3, 5, 6, 7, 8

[21] Sebastian Lieberknecht, Selim Benhimane, Peter Meier, and Nassir Navab. A dataset and evaluation methodology for template-based tracking algorithms. In *ISMAR*, 2009. 2, 3

[22] Liting Lin, Heng Fan, Zhipeng Zhang, Yong Xu, and Haibin Ling. Swintrack: A simple and strong baseline for transformer tracking. In *NeurIPS*, 2022. 8

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2

[24] Yuan Liu, Zehong Shen, Zhixuan Lin, Sida Peng, Hujun Bao, and Xiaowei Zhou. Gift: Learning transformation-invariant dense visual descriptors via group cnns. *NeurIPS*, 2019. 6, 7

[25] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60:91–110, 2004. 3, 6, 7

[26] Christoph Mayer, Martin Danelljan, Goutam Bhat, Matthieu Paul, Danda Pani Paudel, Fisher Yu, and Luc Van Gool. Transforming model prediction for tracking. In *CVPR*, 2022. 8

[27] Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *ECCV*, 2018. 2, 3, 4, 8

[28] Mustafa Ozuysal, Michael Calonder, Vincent Lepetit, and Pascal Fua. Fast keypoint recognition using random ferns. *TPAMI*, 32(3):448–461, 2009. 3

[29] Rémi Pautrat, Viktor Larsson, Martin R Oswald, and Marc Pollefeys. Online invariance selection for local feature descriptors. In *ECCV*, 2020. 6, 7

[30] Rogério Richa, Raphael Sznitman, Russell Taylor, and Gregory Hager. Visual tracking using the sum of conditional variance. In *IROS*, 2011. 3, 6, 7

[31] Ankush Roy, Xi Zhang, Nina Wolleb, Camilo Perez Quintero, and Martin Jägersand. Tracking benchmark and evaluation for manipulation tasks. In *ICRA*, 2015. 1, 2, 3

[32] Jonáš Šerỳch and Jiří Matas. Planar object tracking via weighted optical flow. In *WACV*, 2023. 2, 3, 6, 7

[33] Jack Valmadre, Luca Bertinetto, Joao F Henriques, Ran Tao, Andrea Vedaldi, Arnold WM Smeulders, Philip HS Torr, and Efstratios Gavves. Long-term tracking in the wild: A benchmark. In *ECCV*, 2018. 3

[34] Ning Wang, Wengang Zhou, Jie Wang, and Houqiang Li. Transformer meets tracker: Exploiting temporal context for robust visual tracking. In *CVPR*, 2021. 8

[35] Tao Wang and Haibin Ling. Gracker: A graph-based planar object tracker. *TPAMI*, 40(6):1494–1501, 2017. 3, 6, 7

[36] Xiao Wang, Xiujun Shu, Zhipeng Zhang, Bo Jiang, Yaowei Wang, Yonghong Tian, and Feng Wu. Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark. In *CVPR*, 2021. 3

[37] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Object tracking benchmark. *TPAMI*, 37(9):1834–1848, 2015. 3

[38] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *CVPR*, 2013. 1, 2, 3, 5, 8

[39] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. In *ICCV*, 2021. 8

[40] Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *ECCV*, 2022. 8

[41] Xinrui Zhan, Yueran Liu, Jianke Zhu, and Yang Li. Homography decomposition networks for planar object tracking. In *AAAI*, 2022. 2, 3, 6, 7, 8

[42] Haoxian Zhang and Yonggen Ling. Hvc-net: Unifying homography, visibility, and confidence learning for planar object tracking. In *ECCV*, 2022. 3