Segmentation-Guided Homography Estimation for Long-Term Planar Tracking

Jonas Serych and Jiri Matas

Visual Recognition Group, Faculty of Electrical Engineering
Czech Technical University in Prague, Czech Republic
{serycjon,matas}@fel.cvut.cz

Abstract. Recent state-of-the-art visual trackers produce high quality and long-term-stable segmentation masks. We propose to leverage these strengths for planar object tracking, in which the goal is to estimate a precise 8-degrees-of-freedom homography pose, a geometric representation not estimated by segmentation trackers. We present SAM-H – a planar object tracker that estimates homographies from segmentation mask contours via a training-free pipeline. When SAM-H is applied to masks from SAM 2 [21], it sets a new state-of-the-art performance on the challenging PlanarTrack [15] benchmark by a large margin, +18.4pp on the p@5 metric. We further show that segmentation-based and correspondence-based homography estimation are complementary, and propose WOFT-SAM, which out-performs all prior methods on both PlanarTrack [15] and POT-210 [14]. We also provide precise re-annotations of PlanarTrack initial poses, enabling more accurate benchmarking in the high-precision p@5 metric. The code and the re-annotations are available at https://github.com/serycjon/WOFTSAM.

1 Introduction

Unlike generic visual object tracking, which estimates bounding boxes or segmentation masks, the planar object tracking task requires recovering the full 8-degrees-of-freedom homography transformation [10] relating the planar target's appearance across frames. This enables applications in, *e.g.*, augmented reality, movie post-production, and visual servoing, that demand geometric precision beyond mere localization.

Despite the long history of homography estimation [4, 17, 18, 22, 25, 27], planar tracking in general conditions is far from solved. Challenges include strong perspective distortion, texture-less surfaces, rotation, motion blur, highly reflective surfaces, transparent and virtual targets, and surfaces with dynamically changing appearance as illustrated in Fig. 1.

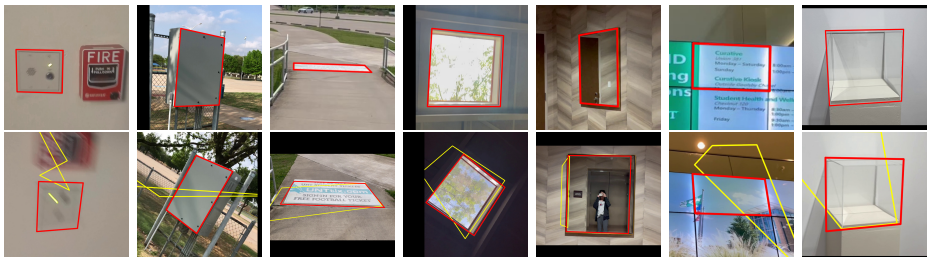

Fig. 1: Challenging scenarios from PlanarTrack [15]. Top: the initial frame, the tracked region in red. Bottom: the proposed **WOFTSAM** output in red, the best performing tracker on the dataset in prior art [15] WOFT [22] in yellow. WOFTSAM out-performs the prior state-of-the-art by introducing a segmentation-based robust re-detection mechanism. Examples include planar tracking challenges like perspective distortion, rotation, and scale change a-c as well as unconventional targets that change appearance in the video d-g. The last column depicts a partial failure.

While existing correspondence-based approaches perform reliably on well-textured objects and in short-term tracking scenarios, they require sufficiently matchable texture and struggle to recover when the target is temporarily lost due to occlusion, motion blur, or out-of-view motion. Meanwhile, general object tracking has been transformed by segmentation-based methods such as SAM 2 [21], which achieve robust long-term tracking through high-quality mask prediction even on low-textured objects and under strong appearance changes. However, these methods provide only region-level binary mask output, which lacks the geometric information needed for planar tracking.

This motivates the problem we study: can 8-DoF homographies be reliably extracted from segmentation masks? This is non-trivial and poses challenges such as occlusion handling and symmetry ambiguities resolution. For example in the case of quadrilateral targets (used in benchmarks, commonly occuring in reality), any cyclic permutation of corners yields the same quadrilateral, resulting in four homography candidates not distinguishable based on just the mask.

We propose SAM-H, a training-free geometric pipeline that decomposes the mask-to-homography tracking problem into mask boundary line extraction via the Hough transform, identification of the quadrilateral corners, and appearance-based symmetry disambiguation. While other approaches are conceivable, such as learning a direct masks-to-homography regression, our classical-geometry-based solution requires no task-specific training and serves as a strong baseline.

Applied to masks from SAM 2 [21], SAM-H outperforms all prior specialized planar trackers on the challenging PlanarTrack [15] benchmark, suggesting that robustness to target appearance changes and temporary tracking failures is the primary bottleneck in current planar tracking.

We further observe that segmentation-based and correspondence-based homography estimation have different strengths depending on the tracking situation: correspondences provide higher precision when texture is visible and tracking is

continuous, and during partial occlusions. Segmentation handles blur, re-detection after target loss and texture-less or appearance changing targets. These situations alternate within individual sequences. We combine both approaches in WOFTSAM, taking the best of both words (correspondence-based precision and segmentation-based re-detection ability).

The contributions of the paper are: (i) We formulate the mask-to-homography problem for quadrilateral targets and propose SAM-H, a training-free method that extract homographies from segmentation masks using classical geometry and appearance-based symmetry disambiguation, setting a new state of the art on PlanarTrack [15], (ii) WOFTSAM: a planar tracker that combines correspondence-based precise homographies with robust target re-detection capability of SAM-H, achieving a new state-of-the-art on POT-210 [14] and outperforming all prior methods on PlanarTrack [15], (iii) Precise re-annotation of the PlanarTrack [15] benchmark initial poses, improving the PlanarTrack suitability for evaluation of high-precision methods. The effects of the initial frame ground-truth accuracy are significant, see Table 1.

2 Related Work

Planar object tracking via homography estimation is a classical computer vision problem. Early approaches align an image template to each video frame using intensity-based registration, *e.g.* Lucas–Kanade [1, 18]. Efficient Second-order Minimization (ESM) [3] and its extensions [5–7] accelerate and improve robustness of Lucas–Kanade by using second-order approximations and feature pyramids. However, direct alignment methods typically struggle under large blur, occlusion or out-of-view conditions common in real-world planar tracking.

A more robust class of trackers uses feature matching: detect and describe keypoints on the object and match them to each frame. Standard pipelines use SIFT [17], SURF [2] or other classical features and estimate a homography by RANSAC [9]. Modern variants replace hand-crafted descriptors with learned ones (e.g. GIFT [16], or LISRD [20]). Instead of matching keypoints independently, Gracker [25] and CGN [13] formulate the task as a graph matching problem. Current state-of-the-art, WOFT [22] estimates homographies from dense optical flow correspondences. It pre-warps the current frame using a previous-frame homography estimate so that the residual displacement between the initial frame and the current pre-warped frame is small. The homography is then estimated by a Weighted Flow Homography (WFH) module, which computes RAFT [23] optical flow between the template and the pre-warped current frame and fits a homography to the resulting correspondences via weighted least squares, where the per-correspondence weights are predicted by a learned module.

Another class of methods is based directly on deep learning. In HDN [26], the homography is estimated by regressing positions of four control points after a rough alignment via an estimated similarity transform. The HVC-Net [27] also regresses the four control points and on top of that models uncertainty and occlusions. PRTrack [28] tracks multiple planar targets simultaneously, estimating

control points (as peaks of per-point heatmaps) and target segmentation masks. These are then processed together in a multi-stage pipeline, modeling the target motion and occlusions.

The broader general object tracking field has recently shifted toward segmentation-based methods. Models like SAM 2 [21] or DAM4SAM [24] achieve robust long-term tracking with precise segmentation masks. However, their output (a binary segmentation mask) encodes only the region occupied by the target, without explicit geometric structure. We propose to utilize the long-term stable segmentation masks for 8-DoF homography estimation.

3 Method

Given a video sequence $(I_t)_{t=0}^{T}$ and a target X_0 in the first frame, the goal is to estimate a homography matrix $\mathbf{H}_t \in \mathbb{R}^{3\times3}$ in each frame I_t, such that warping the target with the estimated homography gives the target pose $X_t = \mathcal{W}(\mathbf{H}_t, X_0)$ in the current frame. The target is usually specified by four control points $X_0 = (x_1, x_2, x_3, x_4), x_i \in \mathbb{R}^2$, that form a quadrilateral.

Recent segmentation trackers [21,24] robustly track the target region throughout most videos in many benchmarks, producing a binary mask S_t on each frame. Our goal is to recover the homographies $\mathbf{H}_t$ from masks S_t. To this end, we propose **SAM-H** (Fig. 2), which uses Hough transform to find the edges of the target quadrilateral in the mask boundary, identifies the target corners as intersections of the edges, and resolves the 4-fold symmetry ambiguity (cyclic permutations of the four corners yields the same mask) using a motion model and appearance-based disambiguation. We then show in **WOFTSAM** (Fig. 3) that SAM-H homographies, while not pixel-precise, provide a robust re-initialization signal that can be combined with correspondence-based estimation [22] to achieve both high precision and long-term robustness.

SAM-H: From Segmentation Mask to Homography. We prompt the SAM 2 [21] segmentation tracker with a quadrilateral mask specified by the initial coordinates X_0 and let it track, providing a mask S_t on each frame.

Edge and corner extraction. We fit four lines to the contours of S_t via Hough transform [8,11] and find their intersections $\tilde{X}_t$, keeping only the ones close to the mask and discarding extra intersections. Detailed description of the Hough transform line search is in the supplementary Section D.

Symmetry disambiguation. The resulting points cannot be used directly to estimate the homography, since the quadrilateral is invariant to cyclical shifts of the corners due to its symmetrical shape. The goal of the *symmetry disambiguation* step is to order the detected Hough lines intersections $\tilde{X}_t$ to form X_t which matches to the initial corners X_0. For frame-to-frame corner ordering, we select the cyclic permutation minimizing displacement from the previous frame corner positions X_{t-1}. At typical video frame rates ($\geq$25 fps), inter-frame motion is small relative to the target size, making this a reliable default.

After tracking loss (due to occlusion, out-of-view, or tracking failure), the previous-frame pose is unreliable for disambiguation. In these cases we use the

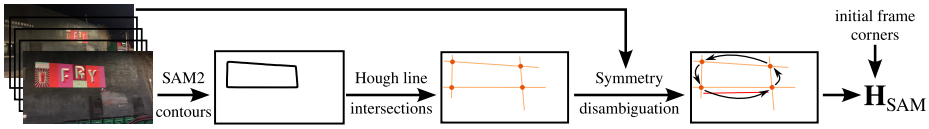

Fig. 2: Overview of the SAM-H procedure for estimating homography from a segmentation tracker output. First, corners of the SAM 2 [21] mask are robustly extracted via intersection of Hough lines. Next, a symmetry disambiguation process based on a motion model and the target appearance matches the corners with the initial frame and a $\mathbf{H}_{\mathrm{SAM}}$ homography is estimated.

target appearance to find the best cyclical corner shift. In particular we compare DINOv2 [19] features between the current frame and the initial view warped under each candidate corner ordering. DINOv2's self-supervised features capture semantic structure robust to geometric imprecisions of the alignment and to significant appearance changes due to reflections, blur, or change of resolution, making them suitable to the task. We require consistent agreement over $\Theta_r = 5$ frames to prevent re-detection on frames where target is partially visible or where the appearance is ambiguous. After the symmetry disambiguation, a homography is estimated from the matched corners.

Partial visibility. In case all four points are visible and detected, we directly compute the homography $\mathbf{H}_t$ between the X_0 and X_t. In case less than four points are detected, the homography $\mathbf{H}_t$ is composed of the homography to the previous frame $\mathbf{H}_{t-1}$ and a residual transformation $\mathbf{\Delta H}_{t-1\to t}$. When at least two points are visible, $\mathbf{\Delta H}$ is a 4-DoF similarity transform (translation + rotation + single scale) estimated from the motion of the visible points since the previous frame. When only a single point is visible, $\mathbf{\Delta H}$ is a pure 2-DoF translation of the visible point with respect to the previous frame.

WOFTSAM: Precise Tracking With Robust Re-Detection. SAM-H homographies $\mathbf{H}_{\mathrm{SAM}}$ are robust, but not pixel-precise, since they are derived solely from the mask boundaries. Correspondence-based methods achieve higher precision when the target is textured and continuously visible, but cannot recover after tracking loss. WOFTSAM combines both strengths in a simple cascade (Fig. 3). Given frame I_t, WOFTSAM first attempts correspondence-based estimation: the current frame is pre-warped ($\mathcal{W}(\cdot,\cdot)$) with the previous-frame homography $\mathbf{H}_{t-1}$ and a refined homography is estimated via the Weighted Flow Homography (WFH) module from [22]: $\mathbf{H}^{(1)} = \mathrm{WFH}\left(I_0, \mathcal{W}(\mathbf{H}_{t-1}^{-1}, I_t)\right)$.

The reliability of this estimate is assessed by the inlier ratio, *i.e.*, the fraction of flow correspondences consistent with the fitted homography. If the inlier ratio falls below $\Theta_i = 20\%$, indicating tracking difficulty or loss, WOFTSAM attempts *re-detection*. The current frame is instead pre-warped with the SAM-H homography $\mathbf{H}_{\mathrm{SAM}}$, providing a rough but robust alignment, and WFH is applied again: $\mathbf{H}^{(2)} = \mathrm{WFH}\left(I_0, \mathcal{W}(\mathbf{H}_{\mathrm{SAM}_t}^{-1}, I_t)\right)$. This leverages SAM-H's re-detection ability while recovering correspondence-level precision through the flow refinement. If even this second attempt fails the inlier check, WOFTSAM falls back to the

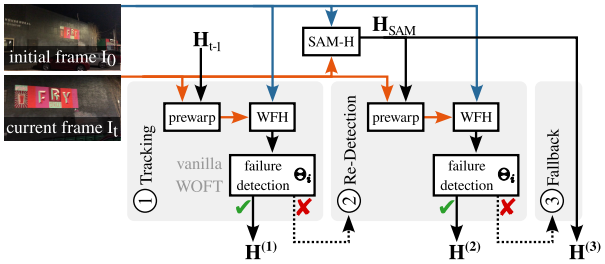

Fig. 3: Overview of the proposed WOFTSAM planar tracker. Like in WOFT [22], tracking ① consists of image pre-warping with the previous frame homography $\mathbf{H}_{t-1}$ and homography estimation via the Weighted Flow Homography (WFH) [22] module. If the homography estimation fails (detected by correspondence support set being small), WOFTSAM does a re-detection step ②, in which the SAM-H output $\mathbf{H}_{\text{SAM}}$ acts as the pre-warping homography. If this fails too, WOFTSAM falls back ③ to $\mathbf{H}_{\text{SAM}}$.

SAM-H homography $\mathbf{H}_{\text{SAM}_t}$ directly. This three-level cascade reflects the observation that correspondence-based estimation should be preferred when it succeeds (higher precision), with SAM-H providing increasingly direct contributions as correspondence quality degrades.

3.1 Implementation details

For the segmentation tracker we use the `SAM2.1_hiera_tiny` variant in default configuration, except for two improvements from DAM4SAM [24] — we set the memory temporal stride to 5, and disable memory updates when the target is not present. We found this to perform better than both plain SAM 2 [21] and the full DAM4SAM [24]. See Sec. 4.2 for details. For optical flow homography estimation we use the pre-trained version of RAFT [23] from [22].

For the symmetry disambiguation, we use a template view either from the initial frame I_0, or from a later one where the target appears biggest-so-far, but before any of the four corners becomes lost, *i.e.*, before any re-detections. This helps in cases when the target has low resolution (far away or zoomed out) on the first frame. We scale the cropped target views to a standard 224×224 resolution and use the `dinov2_vits14_reg` model [19] features (16×16 featuremap resolution, 384 channels). Each featuremap vector is L2-normalized and the similarity is computed as a dot-product between the warped template view and the current view features.

4 Experiments

We evaluate the proposed methods on two standard planar tracking benchmarks, POT-210 [14] and PlanarTrack$_{\text{TST}}$ [15].

POT-210 tests precision under controlled, isolated challenges (one attribute per video). PlanarTrack tests robustness under unconstrained conditions with more

simultaneous challenges per video, including unconventional targets (transparent, reflective, dynamic appearance). The two benchmarks demand fundamentally different capabilities, making strong performance on both non-trivial.

In both benchmarks, the evaluation is based on the *alignment error*,

$$e_{\mathrm{AL}}(\mathbf{H};\mathbf{H}^*,X)=\sqrt{\frac{1}{4}\sum_{i=1}^{4}\|\mathcal{W}(\mathbf{H}^*,\mathbf{x}_i)-\mathcal{W}(\mathbf{H},\mathbf{x}_i)\|^2}, \tag{1}$$

comparing the positions of four template control points, $X=\{x_1,x_2,x_3,x_4\}$ warped to the current frame by the ground truth homography $\mathbf{H}^*$ and by the estimate $\mathbf{H}$.

The alignment error is thresholded to produce the *precision* metric. Following [22] we use two thresholds. The p@5 metric, *i.e.*, the fraction of frames where $e_{\mathrm{AL}}<5$ px, and the p@15 metric with the e_{AL} threshold set to 15 px.

Additionally, we evaluate on the MPOT-3K [28] multiple planar target tracking dataset, where the official e_{AL} threshold is set to 50 px ($Pr_{\epsilon=50}$ metric, here p@50).

	POT-210		$\mathrm{PT}_{\mathrm{TST}}$		$\mathrm{PT}_{\mathrm{TST}}$ re-annot	
method	p@5	p@15	p@5	p@15	p@5	p@15
SIFT [14, 17]	65.8	69.6	17.6	26.7	–	–
LISRD [14, 20]	68.3	79.2	22.0	37.2	–	–
HDN [26]	70.9	92.4	26.6	56.5	–	–
CGN [13]	78.5	93.3	–	–	–	–
HVC-Net [27]	<u>91.4</u>	<u>95.8</u>	42.0	63.1	48.8 (+6.8)	64.3 (+1.2)
WOFT [22]	90.0	95.4	43.6	64.8	48.9 (+5.3)	65.9 (+1.1)
WOFTSAM	**91.9**	**97.5**	<u>51.5</u>	<u>77.2</u>	<u>57.4</u> (+5.9)	<u>77.6</u> (+0.4)
SAM-H	64.4	89.0	**62.0**	**80.0**	**62.0** (+0.0)	**80.1** (+0.1)

Table 1: Precision on the POT-210 [14] benchmark with improved ground-truth from [22] and on $\mathrm{PlanarTrack}_{\mathrm{TST}}$ [15] ($\mathrm{PT}_{\mathrm{TST}}$). Higher is better. The proposed SAM-H tracker sets a new state-of-the-art on PlanarTrack because of its ability to track robustly, including unconventional targets present in parts of the benchmark. Thanks to precise correspondence-based homography estimation and the re-detection ability provided by SAM-H, the WOFTSAM sets a new state-of-the-art on POT-210 and out-performs all prior methods on PlanarTrack. The last two columns show the effect of precise re-annotation of the initial $\mathrm{PT}_{\mathrm{TST}}$ poses as described in Section 4.4.

4.1 Main Results

The results of the proposed SAM-H and WOFTSAM planar trackers are summarized in Tab. 1. Both SAM-H and WOFTSAM substantially outperform all prior methods on PlanarTrack (p@5 +18.4 and +7.9 pp, p@15 +15.2 and +12.4 pp over the next-best WOFT [22]).

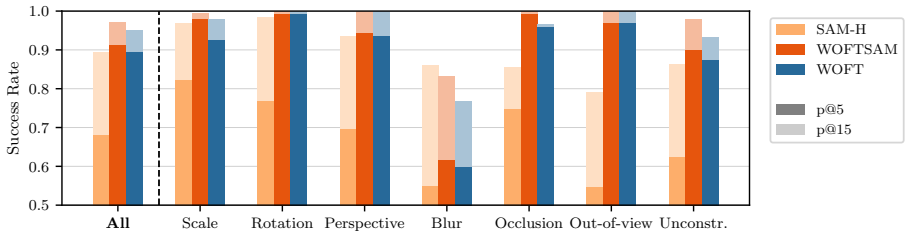

Fig. 4: POT-210 [14] overall (*All*) and per-attribute results. Precision measured on 5 px and 15 px alignment error thresholds on the re-annotated GT [22]. Although SAM-H is the least precise of the shown methods, using it as a robust re-detection mechanism in the proposed WOFTSAM outperforms the WOFT [22] baseline, almost halving its failure rate on the 15 px threshold. The improvement is particularly high on the sequences where re-detection is needed — *blur*, *occlusion*, and *unconstrained* — without decreasing performance on the rest of the benchmark.

POT-210 results On POT-210, where the prior methods already achieve >95% p@15 and the competition is in the high-precision regime, WOFTSAM sets a new state-of-the-art performance, almost halving the 15 px alignment errors compared to the baseline WOFT. The improvement comes from the proposed re-detection capability as confirmed by the challenging-attribute based analysis shown in Figure 4. Indeed, the most improved attributes are motion *blur* (target impossible to track due to heavy motion blur, followed by re-detection), *occlusion*, and *unconstrained* which contains both blur and occlusions. SAM-H mask-based homography estimation by itself is insufficient on POT-210 (64.4 p@5 vs 90.0 for WOFT). The SAM 2 segmentation masks do not provide boundaries precise enough to find the corners to 5 px accuracy. In contrast, the WOFT and WOFTSAM use dense optical-flow-based correspondences from the whole surface of the target, resulting in high precision. The drop in the p@15 score is caused mainly by the *occlusion* and *out-of-view* attributes, where often only one or two lines of the target quadrilateral are visible. WOFT and WOFTSAM can infer a precise pose from the texture on the visible part of the target. Note that once the whole boundary is visible again SAM-H tends to recover. See Section A in supplementary for success plots and more tracking examples.

PlanarTrack results. On PlanarTrack SAM-H sets a new state of the art performance by a large margin. This is primarily due to SAM2's ability to robustly re-detect the target after tracking loss and its ability to track unconventional highly reflective or transparent targets, and objects with changing appearance (*e.g.* TV screens), which was not possible with prior planar tracking methods.

WOFTSAM also outperforms all prior methods (+12.4 pp on p@15), but does not fully match SAM-H (-2.8pp on p@15). Figure 5 shows that SAM-H re-detection leads to consistent improvements over the baseline WOFT, with the benefit growing over time as the baseline accumulates irrecoverable failures.

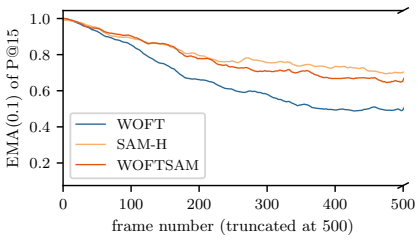

Fig. 5: Average p@15 performance on PlanarTrack$_{TST}$ as a function of the frame number, smoothed with exponential moving average (EMA). After about 3.5 seconds of video ($\approx$ frame 100), the SAM-H-based re-detection starts to boost the WOFTSAM performance compared to the baseline WOFT.

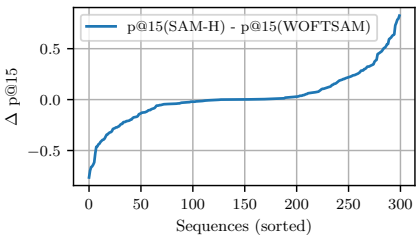

Fig. 6: WOFTSAM and SAM-H complementarity on PlanarTrack$_{TST}$ [15]. Per-sequence differences in p@15 scores between SAM-H and WOFTSAM show a balance between the sequences where WOFTSAM excels (*bottom left*) and the sequences where SAM-H excels (*top right*).

Per-sequence analysis (Fig. 6) shows that roughly half the sequences are tracked better by SAM-H and the other half by WOFTSAM; a per-sequence oracle selecting between WOFTSAM and SAM-H reaches 86.9 p@15 (+6.9pp w.r.t SAM-H), confirming complementarity. WOFTSAM excels on well textured objects and during partial occlusions, where dense correspondences provide higher precision than mask boundaries. SAM-H is stronger on sequences where correspondence-based estimation is unreliable, but this is undetected by the inlier-ratio heuristic, *e.g.*, on a mirror target (see Figure 7), where optical flow confidently tracks reflected content that is geometrically consistent, but does not correspond to the intended physical target surface. Conversely, SAM-H fails when SAM 2 masks do not track the planar object well, *e.g.*, when the planar target is a part of a non-planar object not fully visible in the initial frame and SAM 2 starts tracking the entire object rather than its selected front face.

Developing a more reliable integration of the segmentation-based and the correspondence-based approaches is an open problem. Nevertheless, even the simple track, re-detect, fallback cascade used in WOFTSAM substantially outperforms all prior methods on both benchmarks. The large gains from SAM-H demonstrate that segmentation-based homography estimation is a powerful new tool for planar tracking.

MPOT-3K results. The benchmark [28] focuses mainly on multi object tracking (MOT) scenario and consequently uses MOT-style metrics. Instead, we track each object (by object_id) individually, computing the standard planar tracking metrics. The results shown in Tab. 2 demonstrate good performance, especially on the official 50px threshold, on which WOFTSAM outperforms (p@50 95.1 compared to p@50 92.85) the baseline PRTrack [28], which was designed for this particular dataset. Moreover the PRTrack was evaluated by bi-partite Hungarian matching of the tracker outputs and the GT poses, resulting in over-optimistic

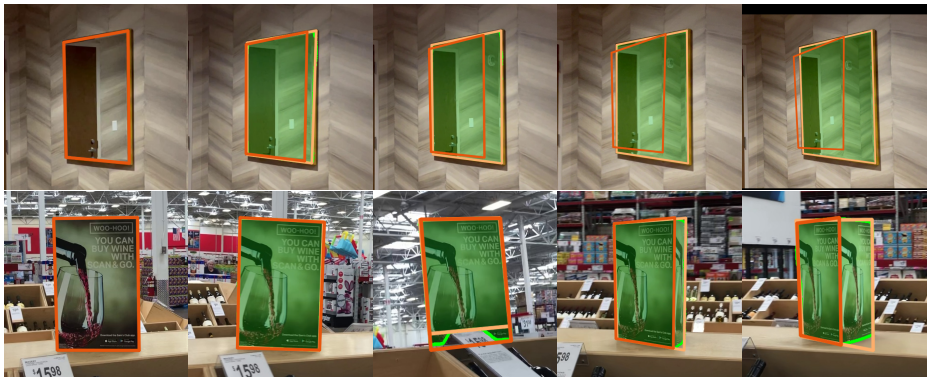

Fig. 7: PlanarTrack [15] examples, where the correspondence-based WOFTSAM (*dark orange*) significantly differ from the segmentation-based SAM-H (*light orange*). SAM 2 mask shown as a *green* overlay, frames are cropped around the object. *Top:* While SAM and consequently SAM-H tracks the mirror surface as annotated in the dataset, WOFTSAM tracks a different planar surface (the doors and part of the adjacent wall) reflected in the mirror. *Bottom:* Two failure modes of SAM-H. (i) SAM 2 segments correctly, but the homography estimation is incorrect due to an occlusion (*middle image*). (ii) SAM 2 "spills" beyond the prompted surface (*last two images*).

results compared to our 1-on-1 prediction-GT evaluation. *E.g.*, target identity switches do not affect the PRTrack p@50 score. We have observed that our methods are often visually more precise than the semi-automatically annotated MPOT-3K GT.

method	p@5	p@15	p@50
WOFTSAM	67.0	89.9	95.1
SAM-H	54.2	81.2	89.9

Table 2: MPOT-3K [28] evaluation. The p@5 and p@15 results are provided for completeness, the p@50 is more representative due to rough semi-automatic GT annotation of the dataset.

4.2 Ablations

The Tab. 3 supports our choice of the segmentation tracker variant SAM2$^+$, *i.e.* the SAM 2 [21] changed to use memory temporal stride 5 and disabling memory updates when the object is not visible. These are two simple, but highly effective modifications introduced by DAM4SAM, outperforming both the original SAM 2 and the full DAM4SAM [24]. The other changes introduced in DAM4SAM are slightly hurting the performance in our mostly distractor-free

method	PT_{TST} p@5	p@15
SAM 2 [21]	59.6	77.2
DAM4SAM [24]	58.6	77.7
SAM2$^+$	**62.0**	**80.0**

Table 3: Ablation — the choice of the segmentation module for the SAM-H homography estimation. SAM2$^+$is the original SAM 2 [21], changed to use memory temporal stride 5 and disabling memory updates when the object is not visible.

Θ_r	PT_{TST} p@5	p@15
0	58.9	76.6
2	61.5	79.6
5	62.0	80.0
8	62.0	80.1

Table 4: SAM-H sensitivity to the symmetry disambiguation consistency parameter Θ_r. Disabling the multi-frame consistency requirement ($\Theta_r = 0$) results in a significant performance drop. Results do not improve beyond the default $\Theta_r = 5$.

scenario. WOFTSAM performance drops only -1.1 pp p@5 and -1.5 pp p@15 on $\text{PlanarTrack}_{\text{TST}}$ when using plain SAM 2 (as opposed to the SAM2$^+$), because the cascade often selects the correspondence-based estimate.

Table 4 shows tracking results for different values of the Θ_r, which sets the minimal number of consecutive frames with stable appearance-based symmetry disambiguation, after which the appearance-selected cyclical corner shift replaces the one based solely on the motion model. Not requiring consistence on consecutive frames, *i.e.*, setting $\Theta_r = 0$ results in a -3.4pp p@15 performance drop. The results do not significantly improve beyond the default $\Theta_r = 5$.

Θ_i	POT-210 p@5	p@15	PT_{TST} p@5	p@15
10%	91.9	97.3	50.2	75.0
20%	91.9	97.5	51.5	77.2
30%	91.6	97.7	52.0	77.6

Table 5: WOFTSAM sensitivity to the failure detection inlier threshold Θ_i. The results are stable over a wide range of values. In PT_{TST}, the results slightly improve when increasing the threshold, consistently with the WOFTSAM vs SAM-H results. Even with the worst-performing 10% threshold, WOFTSAM out-performs the second-best WOFT by a large margin (+10.2pp).

For failure detection in WOFTSAM, we use the $\Theta_i = 20\%$ inlier threshold from WOFT. The Table 5 contains results with different inlier thresholds, showing that WOFTSAM is stable over wide range of Θ_i values.

Cascade analysis, component usage and failure rates. The frequency of each of the three cascade stages being selected is: ① tracking 85.3%, ② re-detection 3.7%, ③ fallback 11.0%. Re-detection rescue rate is 69.8%, *i.e.*, the frequency of alignment error of ② being at least 15 px better than that of ①, and at the same

time ② is selected. The frequency of failure detection miss-classification, *i.e.*, selected ① tracking, but SAM-H would outperform the result by > 15 px, is only 8.6%. In conclusion, most time is spent in the high-accuracy correspondence-based mode and the re-detection mechanisms are useful.

See the supplementary for more ablations (Sec. D for Hough transform, Sec. E for SAM-H comparison against simple baselines).

4.3 Computational efficiency

We evaluated the WOFTSAM efficiency on the PlanarTrack$_{\text{TST}}$ dataset. Overall the proposed WOFTSAM runs at 2.4 FPS using a NVIDIA RTX A5000 GPU, requiring 4.5GB of GPU RAM and Intel(R) Xeon(R) Gold 6326 CPU. While not real-time, it is enough for applications like tracking in movie post-production. Like in WOFT (which runs at a similar speed of 2.7 FPS), the most computationally costly part is the optical flow. The segmentation only takes 42ms per frame, more details in supplementary Section C.

4.4 Improving PlanarTrack GT Quality

The annotation quality analysis in the recently published journal version of PlanarTrack [12] estimates the mean alignment error of the original ground truth to be 5.71 px, with more than 10% of annotations exceeding error of 15 px. We have found that inaccurate ground truth accounts for more than half of the p@5 score drop between SAM-H and WOFTSAM as described next.

Most planar trackers — including those based on optical flow estimation, like WOFT and the proposed WOFTSAM, keypoint-based trackers, and direct alignment methods — attempt to align the texture of the target in the current frame and in the initial frame. The resulting homography is then applied to the initial control points to get the current frame control point positions. In this scenario, the resulting alignment error of the tracker is affected by the ground truth annotation errors on both the initial and the current frame.

On PlanarTrack, the target is often small on the initial frame and increases in size during the sequence (the target is larger than the init in 59% of frames). This increases the influence of initial frame GT annotations, *i.e.* a N px error in the initial frame results in approximately $s_t \cdot N$ px error in the current frame I_t when the target is $s_t\times$ larger than on I_0 as shown in Figure 8.

We have carefully re-annotated all the PlanarTrack$_{\text{TST}}$ initial frames with singe pixel or even sub-pixel precision, trying to perfectly align the initial control points with the corners of the target object (see Figure 9, more details in Section B in supplementary). The mean alignment error of the original ground truth on initial frames when compared to the precise re-annotation is only 1.9 px. However, this has significant influence on the evaluation of correspondence based trackers due to the aforementioned scaling as shown in Table 1. In particular, more than half of the PlanarTrack p@5 score difference between WOFTSAM and SAM-H is caused by the inaccurate original initial frame annotations.

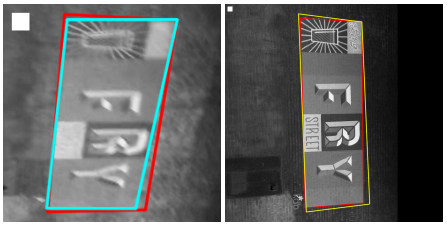

Fig. 8: GT annotations (*red*) on frames I_0 (*left*) and I_{382} (*right*) of PlanarTrack `Seq_00027`. Small pixel error on the low-resolution initial target, compared to precise ground truth (*cyan*) results in large error when warped (*yellow*) to a frame with a close-up target view. In particular, the corner pixel errors grew from 6.7, 0.5, 11.1, and 4.1 to 20.1, 1.5, 25.2, and 16.8. The *top-left* corner of each image shows a 20 px × 20 px white square serving as a scale indicator. Best viewed zoomed in.

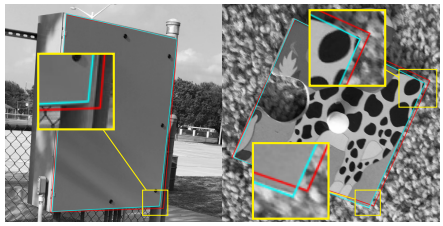

Fig. 9: Examples of PlanarTrack$_{\text{TST}}$ initial frame precise re-annotation. The original GT (*red*) goes outside the tracked object, while the new improved GT (*cyan*) exactly aligns with the target. Best viewed zoomed in, magnified cutouts in *yellow*. The benchmark scores are affected by both the initial and the current frame GT annotation quality for most planar trackers. Our re-annotation corrects half of the sources of the benchmarking errors.

Optical flow based methods WOFT [22] and the proposed WOFTSAM greatly benefit from precise initialization. This is because they propagate any errors in initialization to the whole sequence. The effect on the SAM-H results are negligible, because SAM-H does not primarily estimate the homography aligning the frames I_0 and I_t. Instead, it directly searches for the corners of the tracked object in I_t and the initial corner points only serve as a rough initialization of the SAM 2 tracker. Note that this also aligns well with what the PlanarTrack annotators did: clicking on the corners of the tracked objects. In essence, SAM-H is able to "guess" the GT even when initialized inaccurately.

Note that we only re-annotated the initial frame of each sequence as improving the annotations on all the 133320 frames would take $\approx 444\times$ more time and the first frame already accounts for approximately half of the benchmarking error.

5 Limitations and Future Directions

The proposed SAM-H and WOFTSAM methods set a new state-of-the-art on the currently most challenging planar tracking benchmark PlanarTrack. However, they have some limitations that should be taken into account when deciding what tracker to use for a particular problem.

First, the proposed SAM-H procedure assumes an approximately quadrilateral target shape (which is often the case both in planar tracking benchmarks and in real world) and fails if there are not four stable lines on the target segmentation boundary. A general mask-to-homography method that would be robust to inaccurate and partially occluded masks and that would work under the strong

perspective distortions, rotations and scale changes common in planar tracking is a challenging open problem.

Second, there is a room for improvement of the integration of the two paradigms, correspondence- and mask-based, as shown by the per-sequence oracle experiment. This requires reliably estimating the quality of each homography candidate, which is non-trivial. Appearance-based criteria fail on texture-less, reflective, or dynamically changing targets, while mask-based criteria have the symmetry ambiguity problem and are unreliable during partial occlusions. Especially when occluded by objects with linear boundaries that preserve the quadrilateral shape of the segmentation mask (see Fig. 10).

Fig. 10: An example where SAM-H (*light orange*) fails during an occlusion. It fits the homography to the visible part of the object (SAM 2 mask shown as a *green* overlay). The correspondence-based WOFTSAM (*dark orange*) continues tracking the full object. In the POT-210 [14] dataset, most occlusions are of this type.

Third, the SAM-H procedure is not robust to some types of failures of the underlying segmentation tracker, like when it starts to track the whole 3D object and not just its specified planar face (see Fig. 7 and Fig. 11). Also, while SAM-H re-detects well *after* occlusions, it may fail during partial occlusions. However, the proposed approach is segmentation method agnostic and replacing SAM 2 with a future segmentation tracker (possibly even capable of amodal segmentation behind occlusions) may solve these issues for free.

6 Conclusions

We have proposed to leverage modern robust segmentation tracking methods for tracking planar objects with 8-DoF homography poses, a geometrical representation that is not provided directly by the segmentation masks. The SAM-H tracker sets, by a large margin, a new state-of-the-art on the currently most challenging PlanarTrack [15] planar tracking benchmark. The WOFTSAM tracker combines the strengths of correspondence-based homography estimation with re-detection ability provided by SAM-H, consistently improving over the baseline WOFT [22] and achieving a new state-of-the-art on the POT-210 [14] dataset. The segmentation- and correspondence-based homography estimation have complementary strengths, whose tighter integration and full exploitation is a promising direction for future work.

(a) Blur (b) Texture-less (c) Reflective (d) Spill (e) Box spill (f) Occlusion (g) Distractor

Fig. 11: Examples of SAM 2 [21] output on selected PlanarTrack [15] sequences. The images were cropped around the target from the original full-size frames. The initial frames shown on *top* with the ground truth pose in *red*. The *middle row* shows the selected frame with SAM 2 segmentation overlays in *green*. The *bottom row* shows the frame without overlays for better idea of the input data. The first three examples, (Figs. 11a to 11c), are difficult to track with correspondence-based methods, but are segmented well by SAM 2. In Fig. 11d, the segmentation "spills" outside the intended target, but gets corrected in the SAM-H Hough line detection step. Similar situation, but not recoverable by SAM-H is shown in Fig. 11e. Segmentation works well during partial occlusions, but recovering the full pose based only on the mask is not always possible (Fig. 11f). Distractors sometimes cause SAM 2 tracking failure (Fig. 11g).

We also provide precise re-annotations of PlanarTrack initial poses, showing that even small annotation errors significantly affect high-precision evaluation. Future planar tracking benchmarks should include non-quadrilateral targets and ensure high ground-truth annotation precision to enable meaningful evaluation at tight error thresholds. We publish the code for SAM-H and WOFTSAM, precise ground-truth re-annotations of the PlanarTrack initial frames, and the annotation tools at `https://github.com/serycjon/WOFTSAM`.

Acknowledgements. This work was supported by the National Recovery Plan project CEDMO 2.0 NPO (MPO 60273/24/21300/21000), the EC Digital Europe Programme project CEDMO 2.0 no. 101158609, and by the Research Center for Informatics project CZ.02.1.01/0.0/0.0/16_019/0000765 funded by OP VVV.

References

1. Baker, S., Matthews, I.: Lucas-kanade 20 years on: A unifying framework. International journal of computer vision **56**(3), 221–255 (2004)
2. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (SURF). Computer vision and image understanding **110**(3), 346–359 (2008)

3. Benhimane, S., Malis, E.: 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, pp. 943–948. IEEE (2004)
4. Benhimane, S., Malis, E.: Homography-based 2d visual tracking and servoing. The International Journal of Robotics Research **26**(7), 661–676 (2007)
5. Chen, L., Chen, Y., Ling, H., Tian, X., Tian, Y.: Learning robust features for planar object tracking. IEEE Access **7**, 90398–90411 (2019)
6. Chen, L., Ling, H., Shen, Y., Zhou, F., Wang, P., Tian, X., Chen, Y.: Robust visual tracking for planar objects using gradient orientation pyramid. Journal of Electronic Imaging **28**(1), 1–16 (2019). https://doi.org/10.1117/1.JEI.28.1.013007
7. Chen, L., Zhou, F., Shen, Y., Tian, X., Ling, H., Chen, Y.: Illumination insensitive efficient second-order minimization for planar object tracking. In: 2017 IEEE International Conference on Robotics and Automation (ICRA). pp. 4429–4436. IEEE (2017)
8. Duda, R.O., Hart, P.E.: Use of the Hough transformation to detect lines and curves in pictures. Communications of the ACM **15**(1), 11–15 (1972)
9. Fischler, M.A., Bolles, R.C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24**(6), 381–395 (1981)
10. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press, ISBN: 0521540518, 2 edn. (2004)
11. Hough, P.V.: Method and means for recognizing complex patterns (1962)
12. Jiao, Y., Liu, X., Liu, X., Yuan, X., Fan, H., Zhang, L.: Planartrack: A high-quality and challenging benchmark for large-scale planar object tracking. Computer Vision and Image Understanding **262**, 104553 (2025). https://doi.org/https://doi.org/10.1016/j.cviu.2025.104553
13. Li, K., Liu, H., Wang, T.: Centroid-based graph matching networks for planar object tracking. Machine Vision and Applications **34**(2), 31 (2023)
14. Liang, P., Wu, Y., Lu, H., Wang, L., Liao, C., Ling, H.: Planar object tracking in the wild: A benchmark. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). pp. 651–658. IEEE (2018)
15. Liu, X., Liu, X., Yi, Z., Zhou, X., Le, T., Zhang, L., Huang, Y., Yang, Q., Fan, H.: PlanarTrack: A large-scale challenging benchmark for planar object tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 20449–20458 (Oct 2023)
16. Liu, Y., Shen, Z., Lin, Z., Peng, S., Bao, H., Zhou, X.: GIFT: Learning transformation-invariant dense visual descriptors via group cnns. Advances in Neural Information Processing Systems **32** (2019)
17. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision **60**(2), 91–110 (2004)
18. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: Proceedings of the 7th International Joint Conference on Artificial Intelligence-Volume 2. pp. 674–679 (1981)
19. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: DINOv2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
20. Pautrat, R., Larsson, V., Oswald, M.R., Pollefeys, M.: Online invariance selection for local feature descriptors. In: European Conference on Computer Vision. pp. 707–724. Springer (2020)

21. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollár, P., Feichtenhofer, C.: SAM 2: Segment anything in images and videos. arXiv preprint (2024)
22. Šerých, J., Matas, J.: Planar object tracking via weighted optical flow. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1593–1602 (2023)
23. Teed, Z., Deng, J.: RAFT: Recurrent all-pairs field transforms for optical flow. In: European Conference on Computer Vision. pp. 402–419. Springer (2020)
24. Videnovic, J., Lukezic, A., Kristan, M.: A distractor-aware memory for visual object tracking with sam2. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24255–24264 (2025)
25. Wang, T., Ling, H.: Gracker: A graph-based planar object tracker. IEEE transactions on pattern analysis and machine intelligence **40**(6), 1494–1501 (2017)
26. Zhan, X., Liu, Y., Zhu, J., Li, Y.: Homography decomposition networks for planar object tracking. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 3234–3242 (2022)
27. Zhang, H., Ling, Y.: HVC-Net: Unifying homography, visibility, and confidence learning for planar object tracking. In: Computer Vision and Pattern Regognition (CVPR) (2022)
28. Zhang, Z., Liu, S., Yang, J.: Multiple planar object tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23460–23470 (2023)