Supplementary Materials: Planar Object Tracking via Weighted Optical Flow

Jonáš Šerých, Jiří Matas
CMP Visual Recognition Group, Department of Cybernetics,
Faculty of Electrical Engineering, Czech Technical University in Prague
{serycjon,matas}@fel.cvut.cz

1. Video Results

We show examples of WOFT POT-210 [8] tracking in `WOFT_POT_box.mp4` and of POIC [3] tracking in `WOFT_POIC_box.mp4`. The tracked video is shown in grayscale to improve visibility of the WOFT output. The `WOFT_weights.mp4` video shows the estimated flow weight maps projected to the current frame by the tracker homography pose (yellow – $w_i = 1$, purple – $w_i = 0$). Finally, we used the WOFT tracker homographies for simple augmented reality application demo – replacing the target surface with a static image – in `WOFT_AR.mp4`.

2. POT-280 Results

The alignment error threshold plots on POT-280 [7] are shown in Fig. 4. Table 1 shows the corresponding P@5 and P@15 scores.

We will publish the raw results of WOFT on all the tested datasets (POT-210 [8], its extension POT-280 [7], and POIC [3]).

3. Ground Truth Re-Annotation

Figure 2 depicts additional examples of the original POT-210 [8] ground-truth alignment compared to our re-annotation. On some frames a precise homography alignment was not possible – either due to strong motion blur, or due to imperfect planarity of the targets. A target non-planarity, *e.g.* a slight bend in otherwise flat-looking target, manifests itself the most when the target is viewed from extreme angles. We annotate such cases as precisely as possible (selected examples shown in Fig. 1) and mark the frames as problematic (we will publish this information together with the re-annotated homographies).

method	year	P@5	P@15
SuperGlue [11, 7]	2020	37.7	58.2
SOSNet [12, 7]	2019	51.9	67.1
HDN [14]	2022	56.7	88.9
SIFT [9, 7]	2004	57.2	68.4
LISRD [10, 7]	2020	57.3	77.6
WOFT (ours)		**76.9**	**93.2**

Table 1. Results on POT-280 [7] dataset. The proposed WOFT tracker sets a new state-of-the-art performance in both accuracy (P@5) and robustness (P@15).

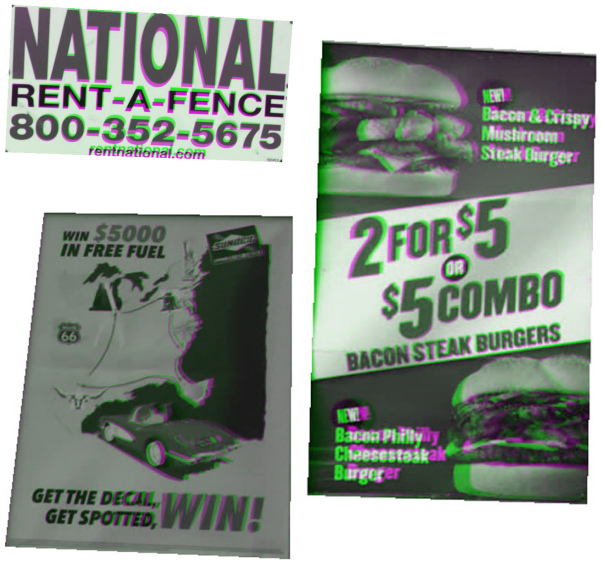

Figure 1. Selected not completely planar targets from POT-210 [8]. When viewed from extreme angle, slight target non-planarity becomes visible. It is then impossible to precisely align the image with the template view on the whole target surface. Top-left image: precise alignment on sides, imprecise alignment in the center. Bottom-left image: imprecise alignment in the center and bottom-right part. Right image: precise alignment in the center, imprecise elsewhere.

4. Speed-Accuracy Trade-Off

A faster WOFT tracker variant $\text{WOFT}_{\downarrow s}$ down-scales the input images to $H/s \times W/s$ and re-scales the output homographies to the original resolution. We show the speed-

Figure 2. Additional POT-210 [8] re-annotation examples. Left: original GT annotation, right: our precise re-annotation. The grayscale template in green channel, the GT-warped current frame in red and blue channels. Imprecise annotation causes green and magenta shadows, while precisely aligned images produce a grayscale result.

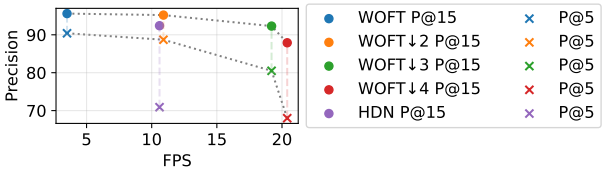

Figure 3. Speed-accuracy trade-off of $\text{WOFT}_{\downarrow s}$ variants as measured on the re-annotated POT-210 dataset. Down-scaling the input images with $s = 2$ or $s = 3$ significantly speeds up ($3\times$, respectively $6\times$) the WOFT tracker while retaining state-of-the-art accuracy. The second-best performing method on POT-210 – HDN [14] in purple.

accuracy trade-off for $s \in \{1, 2, 3, 4\}$ in figure 3. The $\text{WOFT}_{\downarrow 3}$ variant is better than the second best POT-210 method HDN [14], while running $1.8\times$ faster on the same GPU type.

5. Replacing RAFT with LiteFlowNet2

We have evaluated the proposed weighted flow homography (WFH) idea with LiteFlowNet2 [6] optical flow network. We have kept the same 3-layer CNN architecture for weight estimation as with RAFT (Sec. 3.1 in the paper). For inputs, we have used the cost-volume on the last LiteFlowNet2 NetE pyramid level (level 3). The cost-volume contains a 7×7 correlation response map for each position in the template feature map. We feed each of these 7×7 maps through the weight estimation CNN to get the corresponding flow weights w_i. The weight estimator training was kept the same, except we have only trained for 5 epochs.

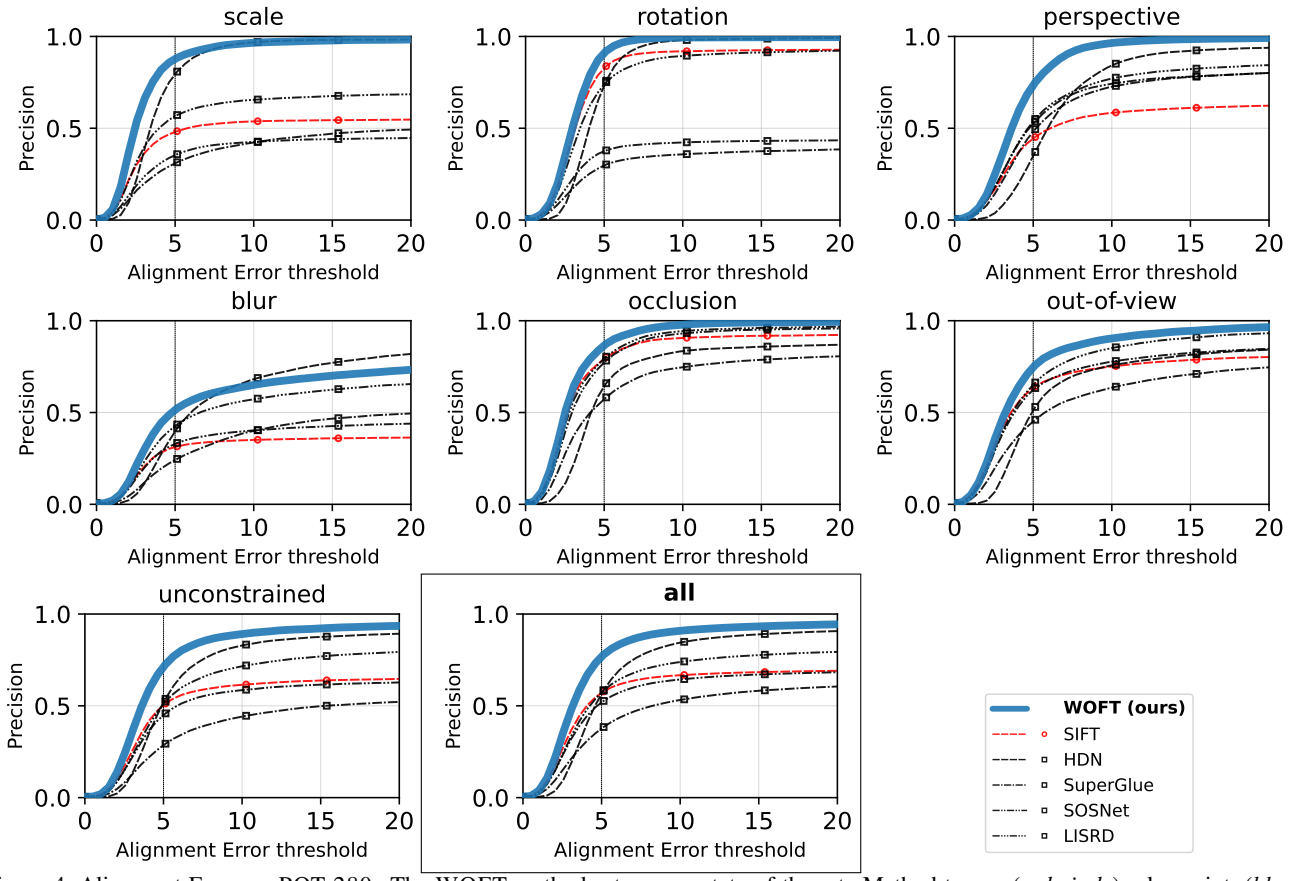

Figure 4. Alignment Error on POT-280. The WOFT method sets a new state-of-the-art. Method types: (*red circle*) – keypoint, (*black square*) – deep.

method	P@5	P@15
SIFT [9, 3]	43.8	54.5
SOL [5]	55.3	74.8
HDN [14]	74.4	94.5
Bit-Planes [1]	75.1	76.0
Gracker [13]	75.2	89.9
GOP-ESM [3]	90.8	93.1
SiamESM [2]	**96.1**	97.7
WOFT	**96.1**	**98.0**

Table 2. Results on POIC [3] dataset. The proposed WOFT tracker achieves state-of-the-art performance in both accuracy (P@5) and robustness (P@15).

We did not fine-tune the LiteFlowNet2 and used the `liteflownet2_ft_4x1_600k_sintel_kitti_320x768` configuration and checkpoint from MMFlow [4].

6. Additional POIC results

We provide the POIC dataset results in table 2. The proposed WOFT tracker achieves state-of-the-art performance. Additionally, we show all the per-sequence results in `POIC_results_per_sequence.pdf`.

7. POT-210 results grouped by method types

For faster comparison with other methods, we provide additional views of the POT-210 results figure 6 in the paper. We group the results by method type – *keypoint* trackers in Fig. 5, *direct* trackers in Fig. 6, and *deep learning* trackers in Fig. 7.

8. POT-210 Re-Annotated plots

The alignment error plots on just the re-annotated POT-210 frames are shown in Fig. 8.

9. Post-processing industry state-of-the-art

We have additionally evaluated a film post-processing industry standard planar tracking solution Mocha Pro 2022. The software is primarily made for interactive use, but also provides python API enabling fair benchmark evaluation. We have tested three variants of the Mocha Pro tracker hyper-parameters on POT-210. We used perspective (homography) model in all the experiments. We have tried I. the default parameters, II. increasing the *Min % Pixels Used* parameter to 100%, and III. increasing the target initialization by 10%. We have chosen the variants II. and III. according to the recommendations in Mocha Pro user guide. Variant III. performs best, but still significantly worse (P@5 32.8, P@15 52.0) than POT-210 state-of-the-art.

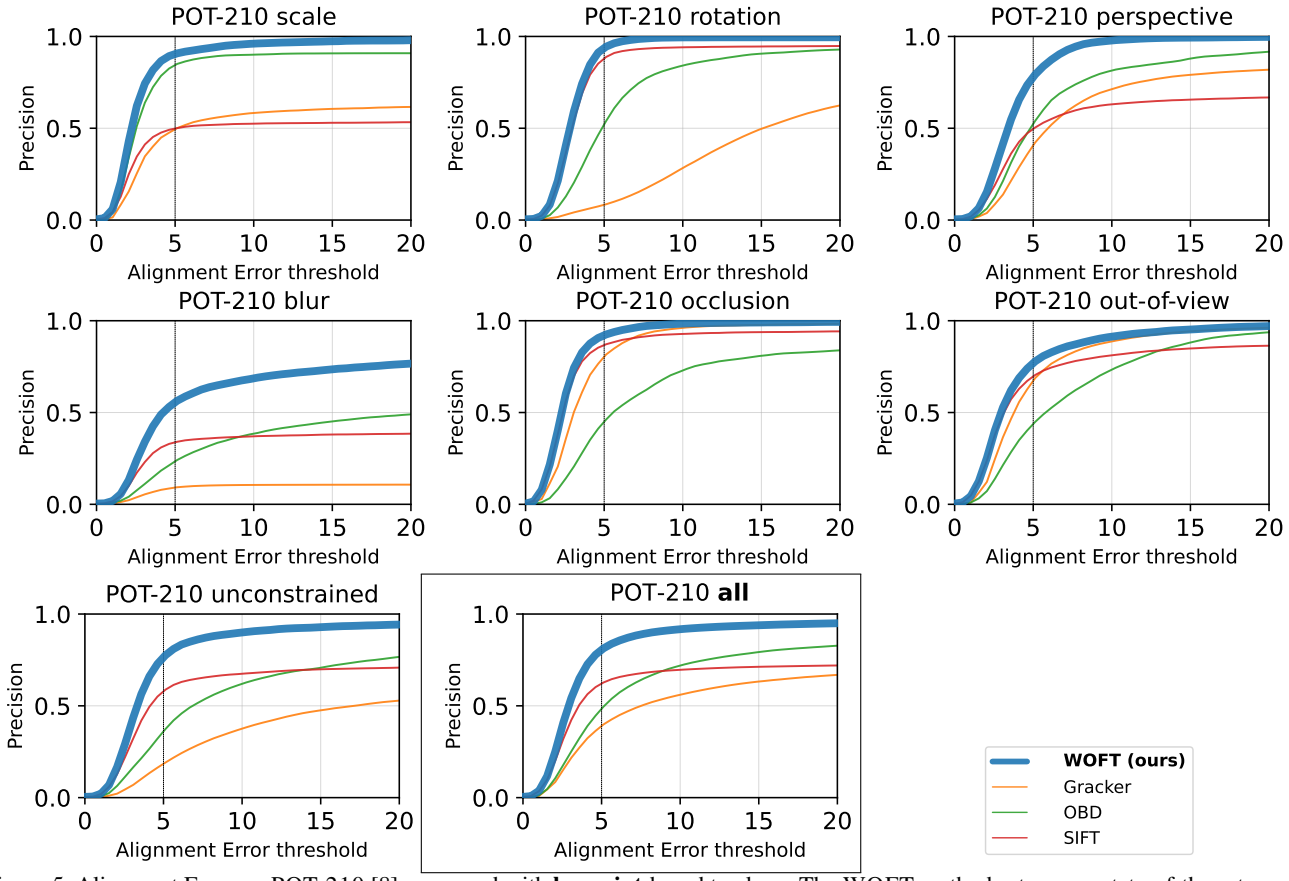

Figure 5. Alignment Error on POT-210 [8] compared with **keypoint**-based trackers. The WOFT method sets a new state-of-the-art.

References

[1] Hatem Alismail, Brett Browning, and Simon Lucey. Robust tracking in low light and sudden illumination changes. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 389–398. IEEE, 2016.

[2] Lin Chen, Yaowu Chen, Haibin Ling, Xiang Tian, and Yuesong Tian. Learning robust features for planar object tracking. *IEEE Access*, 7:90398–90411, 2019.

[3] Lin Chen, Haibin Ling, Yu Shen, Fan Zhou, Ping Wang, Xiang Tian, and Yaowu Chen. Robust visual tracking for planar objects using gradient orientation pyramid. *Journal of Electronic Imaging*, 28(1):1–16, 2019.

[4] MMFlow Contributors. MMFlow: Openmmlab optical flow toolbox and benchmark. `https://github.com/open-mmlab/mmflow`, 2021.

[5] Sam Hare, Amir Saffari, and Philip HS Torr. Efficient online structured output learning for keypoint-based object tracking. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1894–1901. IEEE, 2012.

[6] Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. A lightweight optical flow cnn—revisiting data fidelity and regularization. *IEEE transactions on pattern analysis and machine intelligence*, 43(8):2555–2569, 2020.

[7] 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.

[8] Pengpeng Liang, Yifan Wu, Hu Lu, Liming Wang, Chunyuan Liao, and Haibin Ling. Planar object tracking in the wild: A benchmark. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 651–658. IEEE, 2018.

[9] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[10] Rémi Pautrat, Viktor Larsson, Martin R Oswald, and Marc Pollefeys. Online invariance selection for local feature descriptors. In *European Conference on Computer Vision*, pages 707–724. Springer, 2020.

[11] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4938–4947, 2020.

[12] Yurun Tian, Xin Yu, Bin Fan, Fuchao Wu, Huub Heijnen, and Vassileios Balntas. SOSNet: Second order similarity regularization for local descriptor learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11016–11025, 2019.

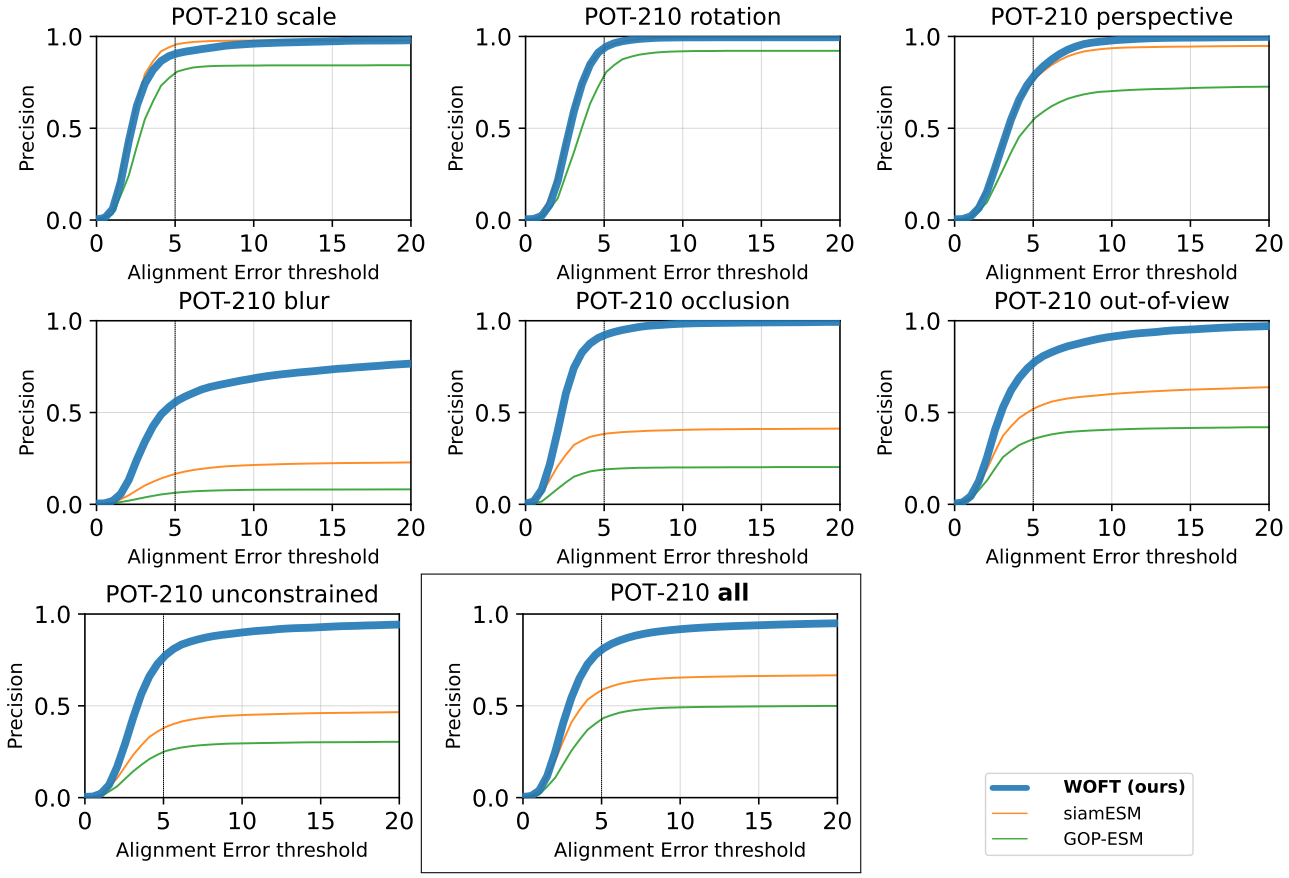

Figure 6. Alignment Error on POT-210 [8] compared with **direct**-type trackers. The WOFT method sets a new state-of-the-art.

[13] Tao Wang and Haibin Ling. Gracker: A graph-based planar object tracker. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1494–1501, 2017.

[14] Xinrui Zhan, Yueran Liu, Jianke Zhu, and Yang Li. Homography decomposition networks for planar object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3234–3242, 2022.

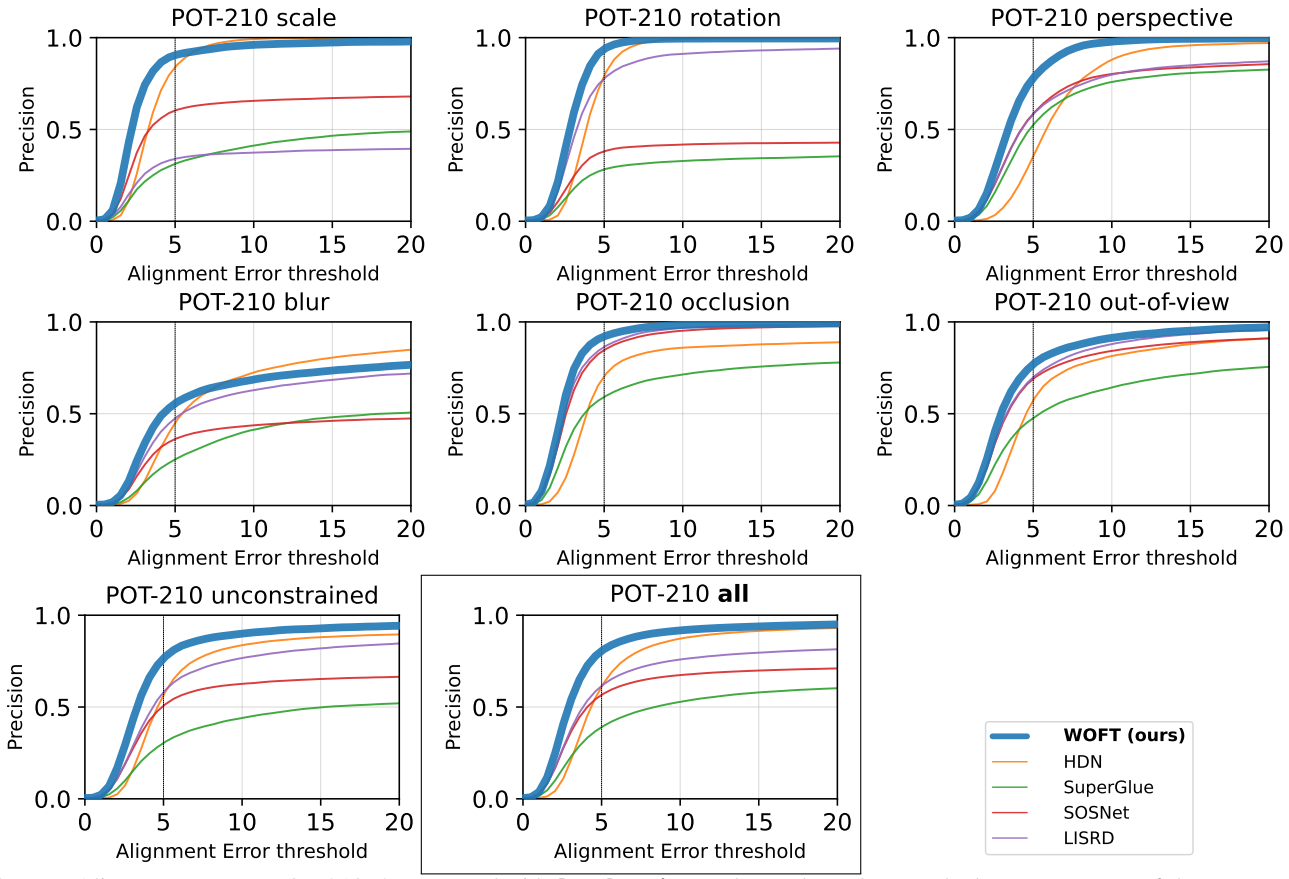

Figure 7. Alignment Error on POT-210 [8] compared with **deep learning** trackers. The WOFT method sets a new state-of-the-art.

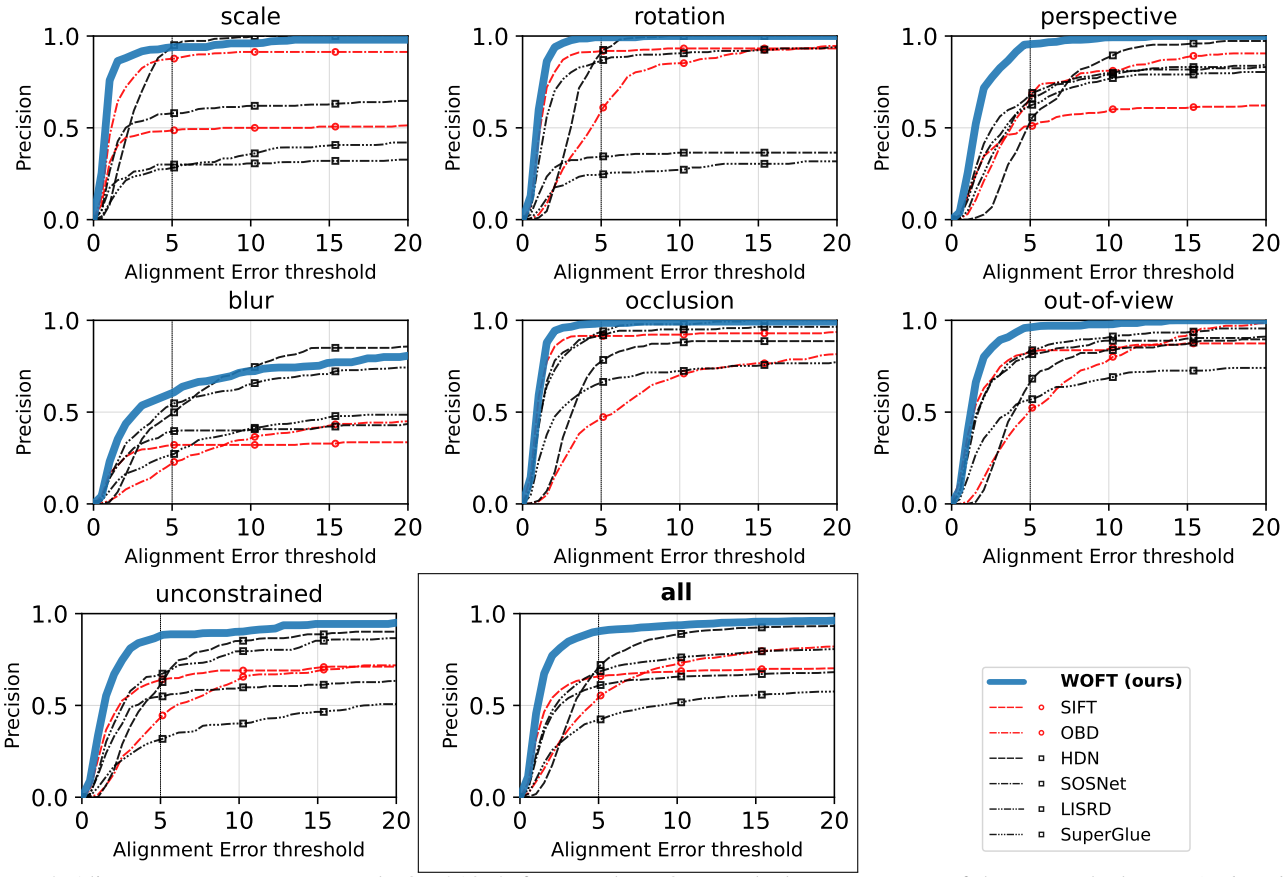

Figure 8. Alignment Error on re-annotated POT-210 [8] frames. The WOFT method sets a new state-of-the-art. Method types: (*red circle*) – keypoint, (*green triangle*) – direct, (*black square*) – deep.