Supplementary Material

Tamara R. Lenhard[1,2,3] Andreas Weinmann[2] Hichem Snoussi[3,4] Tobias Koch[1]

[1]Institute for the Protection of Terrestrial Infrastructures, German Aerospace Center (DLR), Sankt Augustin, Germany
[2]ACIDA Lab, Technical University of Applied Sciences Würzburg-Schweinfurt, Schweinfurt, Germany
[3]Data Science Institute, European University of Technology, European Union
[4]LIST3N, Université de Technologie de Troyes, Troyes, France

{tamara.lenhard, tobias.koch}@dlr.de, andreas.weinmann@thws.de, hichem.snoussi@utt.fr

A. Sequence-Level Tracking Performance

While the main paper primarily reports aggregated results on DUT Anti-UAV [13] and the custom datasets R1 and R2 (cf. Sec. 3.3, main paper), including comparisons with other tracking algorithms (cf. Tab. I), Tabs. II and III provide per-sequence evaluations. These results expose sequence-specific variations, offering a fine-grained characterization of SAMURAI and its detector-augmented extension.

Performance on DUT Anti-UAV. SAMURAI achieves stable tracking performance across most sequences, irrespective of whether initialization is based on ground-truth (GT) annotations or detector predictions (cf. GT vs. D, 2nd column, Tab. II). The detector-augmented extension further reinforces this stability, consistently matching or surpassing GT-based initialization and exhibiting robustness to initialization noise. In challenging sequences (*e.g.*, video05, video12, video16), detector-only initialization leads to noticeable performance degradation, whereas the detector-augmented variant mitigates these effects and recovers performance close to GT-level.

A representative example is video05, where the initial detection erroneously marks the mirror of a car as the drone (cf. Fig. II, bottom), while the actual drone is located at the top-center of the frame. In this situation, SAMURAI cannot recover from the erroneous initialization, whereas its detector-augmented extension leverages continuous detector feedback to correct the error and restore accurate tracking. Another example is video04, where the unfolding of an attached parachute induces significant appearance changes, resulting in erroneous bounding boxes from inaccurate segmentation masks (cf. Fig. II, top). With first-frame-only initialization (both GT- and detector-based), these errors accumulate and remain uncorrected. In contrast, detector augmentation enables SAMURAI to counteract these effects and sustain accurate tracking (cf. Fig. I).

Table I. Comparison of SAMURAI (with GT initialization) against state-of-the-art trackers on the DUT Anti-UAV dataset [13]. Baseline results for all trackers (except SAMURAI) are reported as presented in [13]. Best results are in **bold**.

Tracker	S ↑	P ↑	P_{norm} ↑
SiamFC [1]	0.381	0.623	0.526
ECO [5]	0.404	0.717	0.643
SPLT [11]	0.405	0.651	0.585
SiamRPN++ [10]	0.545	0.780	0.709
ATOM [6]	0.578	0.830	0.758
DiMP [2]	0.578	0.831	0.756
TransT [3]	0.586	0.832	0.765
LTMU [4]	0.608	0.783	0.858
SAMURAI [12]	**0.663**	**0.888**	**0.973**

On the other hand, sequences such as video06 and video10 (characterized by favorable tracking conditions, i.e., blue sky) achieve near-perfect performance across all metrics – even under first-frame-only initialization via detector predictions (cf. Tab. II).

Performance on Custom Data. On the custom datasets R1 and R2, SAMURAI exhibits pronounced sequence-level performance variations. In the POS3 sequences of both datasets, detector-only initialization without detector-based augmentation leads to substantial degradation, with low success rates, reduced mAP values, and elevated FNRs (cf. Tab. III and Fig. IV). However, when leveraging the detector-augmented version of SAMURAI, performance improves markedly: tracking scores often double, and detection quality rises to levels comparable to or even exceeding GT initialization. For instance, in POS3 (R1), the success rate increases from 0.289 to 0.560, while the FNR is reduced by more than half. Visual inspection (cf. Fig. III) further reveals that the observed improvements in mAP are driven not only by continuous prompting through detector-derived bounding boxes but also by the averaging mechanism embedded in the proposed Prediction Fusion Module (cf. Sec. 3.2).

Table II. Performance of SAMURAI and its detector-augmented extension (✓, 3rd column) on sequences from the DUT Anti-UAV dataset [13]. *GT* (2nd column) denotes initialization with ground truth, while *D* denotes detector-based initialization using the first YOLO-FEDER FusionNet prediction as the bounding-box prompt. Best results are highlighted in **bold**.

			Tracking Metrics			Detection Metrics				
Seq.	Init. Method	Detector Augmentation	S ↑	P ↑	P_{norm} ↑	mAP ↑			FNR ↓	FDR ↓
						@0.25	@0.5	@0.5-0.95		
video01	GT	–	0.808	0.878	0.959	0.989	**0.983**	0.697	0.021	**0.000**
	D	–	0.807	**0.878**	0.959	0.989	0.981	**0.697**	0.022	0.001
	D	✓	**0.809**	0.872	**0.965**	**0.990**	0.977	0.690	**0.017**	0.003
video02	GT	–	0.730	0.926	0.980	0.995	0.995	0.466	0.000	0.000
	D	–	0.724	0.924	0.980	0.995	0.995	0.454	0.000	0.000
	D	✓	**0.769**	**0.941**	**0.980**	**0.995**	**0.995**	**0.557**	**0.000**	**0.000**
video03	GT	–	0.802	0.952	0.980	0.995	0.995	0.694	0.000	0.000
	D	–	0.802	0.951	0.980	0.995	0.995	0.694	0.000	0.000
	D	✓	**0.842**	**0.958**	**0.980**	**0.995**	**0.995**	**0.754**	**0.000**	**0.000**
video04	GT	–	0.412	0.351	0.938	0.716	0.401	0.190	0.226	0.226
	D	–	0.412	0.352	0.938	0.716	0.401	0.190	0.226	0.226
	D	✓	**0.872**	**0.929**	**0.980**	**0.995**	**0.990**	**0.765**	**0.000**	**0.000**
video05	GT	–	**0.807**	**0.910**	**0.969**	**0.939**	**0.933**	**0.679**	**0.044**	**0.040**
	D	–	0.010	0.000	0.014	0.000	0.000	0.000	1.000	1.000
	D	✓	0.699	0.847	0.914	0.904	0.882	0.478	0.082	0.057
video06	GT	–	0.893	0.935	0.980	0.995	0.995	0.817	0.000	0.000
	D	–	0.893	0.935	0.980	0.995	0.995	0.817	0.000	0.000
	D	✓	**0.904**	**0.947**	**0.980**	**0.995**	**0.995**	**0.841**	**0.000**	**0.000**
video07	GT	–	0.868	0.905	0.980	0.943	0.943	0.728	0.000	0.073
	D	–	**0.868**	0.905	**0.980**	**0.943**	**0.943**	0.728	**0.000**	**0.073**
	D	✓	0.866	**0.914**	0.979	0.942	0.941	**0.733**	0.001	0.074
video08	GT	–	0.834	0.933	0.980	0.969	0.967	0.683	0.000	0.028
	D	–	0.834	0.932	**0.980**	0.969	0.967	0.683	0.001	0.028
	D	✓	**0.846**	**0.936**	0.979	**0.969**	**0.967**	**0.708**	**0.001**	**0.028**
video09	GT	–	0.879	0.938	0.980	0.898	0.898	0.713	0.000	0.109
	D	–	0.879	0.938	0.980	0.898	0.898	0.713	0.000	0.109
	D	✓	**0.912**	**0.946**	**0.980**	**0.898**	**0.898**	**0.776**	**0.000**	**0.109**
video10	GT	–	0.820	0.917	0.980	0.995	0.993	0.682	0.000	0.002
	D	–	0.818	0.916	0.980	0.995	0.992	0.677	0.000	0.002
	D	✓	**0.857**	**0.929**	**0.980**	**0.995**	**0.995**	**0.753**	**0.000**	**0.002**
video11	GT	–	0.729	0.917	0.980	0.994	0.846	0.464	0.000	0.002
	D	–	0.727	0.918	0.980	0.994	0.839	0.463	0.000	0.002
	D	✓	**0.797**	**0.938**	**0.980**	**0.994**	**0.990**	**0.611**	**0.000**	**0.002**
video12	GT	–	0.438	0.850	0.932	0.905	0.203	0.065	0.093	0.065
	D	–	0.426	0.851	0.934	0.903	0.169	0.058	0.095	0.066
	D	✓	**0.523**	**0.895**	**0.958**	**0.940**	**0.524**	**0.139**	**0.060**	**0.047**
video13	GT	–	0.479	0.922	0.980	0.987	0.374	0.076	0.001	0.001
	D	–	0.450	0.920	**0.980**	0.974	0.306	0.058	0.002	0.003
	D	✓	**0.517**	**0.936**	0.979	**0.993**	**0.510**	**0.105**	**0.001**	**0.001**
video14	GT	–	0.568	0.906	0.980	0.994	0.731	0.218	0.000	0.000
	D	–	0.560	0.904	0.980	0.995	0.705	0.196	0.000	0.000
	D	✓	**0.682**	**0.922**	**0.980**	**0.995**	**0.975**	**0.396**	**0.000**	**0.000**
video15	GT	–	0.516	0.933	0.980	0.988	0.519	0.095	0.000	0.014
	D	–	0.494	0.930	0.980	0.988	0.381	0.066	0.000	**0.014**
	D	✓	**0.642**	**0.939**	**0.980**	**0.988**	**0.925**	**0.310**	**0.000**	0.015
video16	GT	–	0.491	0.903	0.974	0.942	0.374	0.083	**0.032**	0.026
	D	–	0.449	**0.903**	**0.974**	0.922	0.235	0.046	0.041	0.034
	D	✓	**0.542**	0.889	0.937	**0.975**	**0.628**	**0.162**	0.045	**0.001**
video17	GT	–	0.488	0.908	0.979	0.941	0.440	0.095	0.040	0.039
	D	–	0.478	0.907	**0.979**	0.934	0.405	0.085	0.050	0.045
	D	✓	**0.591**	**0.923**	0.977	**0.983**	**0.791**	**0.229**	**0.008**	**0.004**
video18	GT	–	0.590	0.938	0.971	0.995	0.632	0.203	0.009	0.000
	D	–	0.580	0.937	0.971	0.994	0.580	0.188	0.009	0.000
	D	✓	**0.628**	**0.950**	**0.971**	**0.995**	**0.806**	**0.268**	**0.009**	**0.000**
video19	GT	–	0.508	0.934	0.980	**0.993**	0.485	0.099	**0.002**	0.002
	D	–	0.487	0.934	**0.980**	0.990	0.380	0.073	0.003	0.003
	D	✓	**0.512**	**0.935**	0.971	0.984	**0.583**	**0.129**	0.011	**0.002**
video20	GT	–	0.601	0.910	0.980	0.995	0.692	0.218	0.000	0.000
	D	–	0.590	0.907	**0.980**	**0.995**	0.660	0.199	**0.000**	**0.000**
	D	✓	**0.691**	**0.941**	0.979	0.994	**0.935**	**0.389**	0.002	0.001

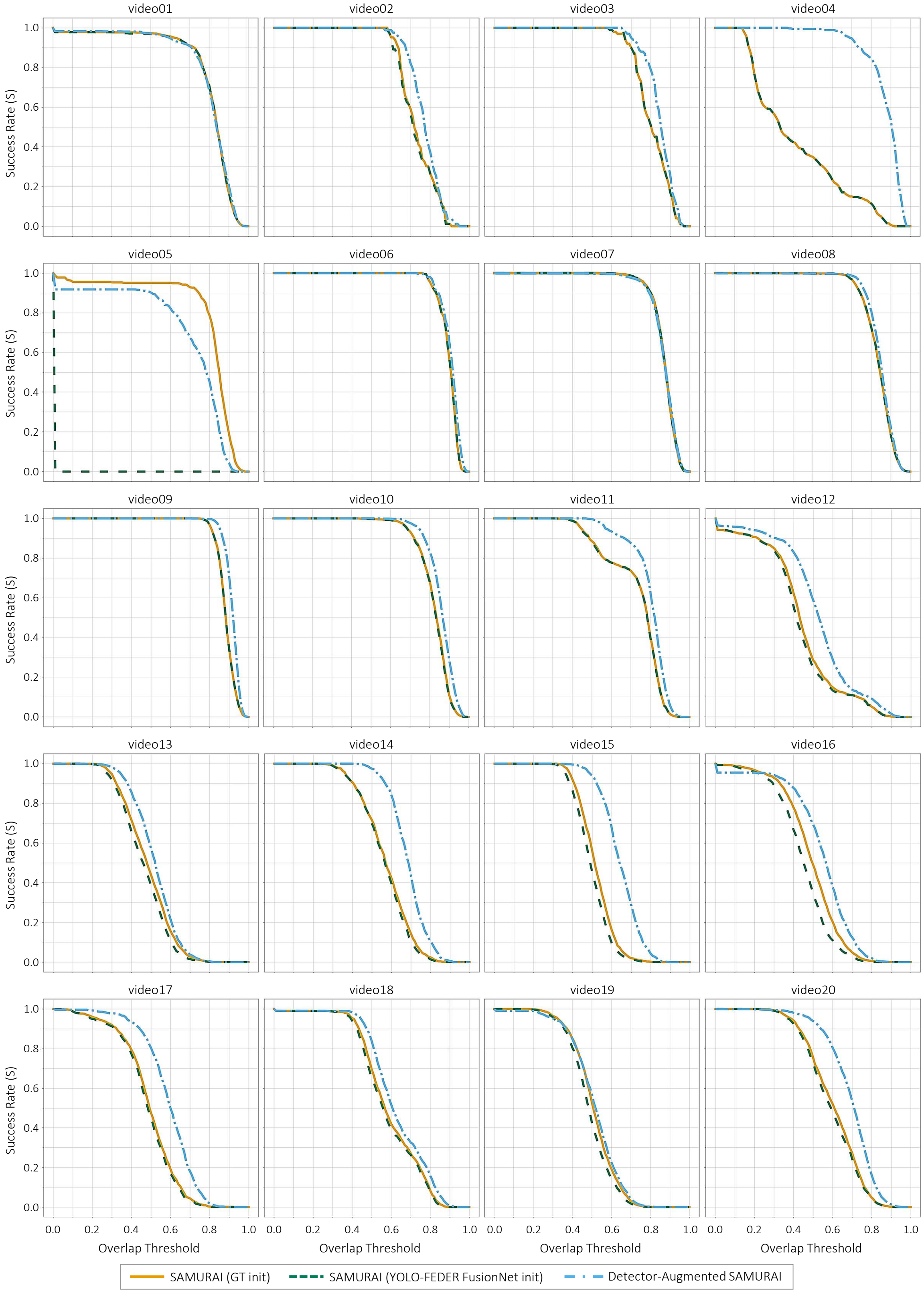

Figure I. Success plots on sequences from DUT Anti-UAV [13], comparing SAMURAI with ground-truth initialization, SAMURAI with first-frame YOLO-FEDER FusionNet initialization, and the detector-augmented SAMURAI.

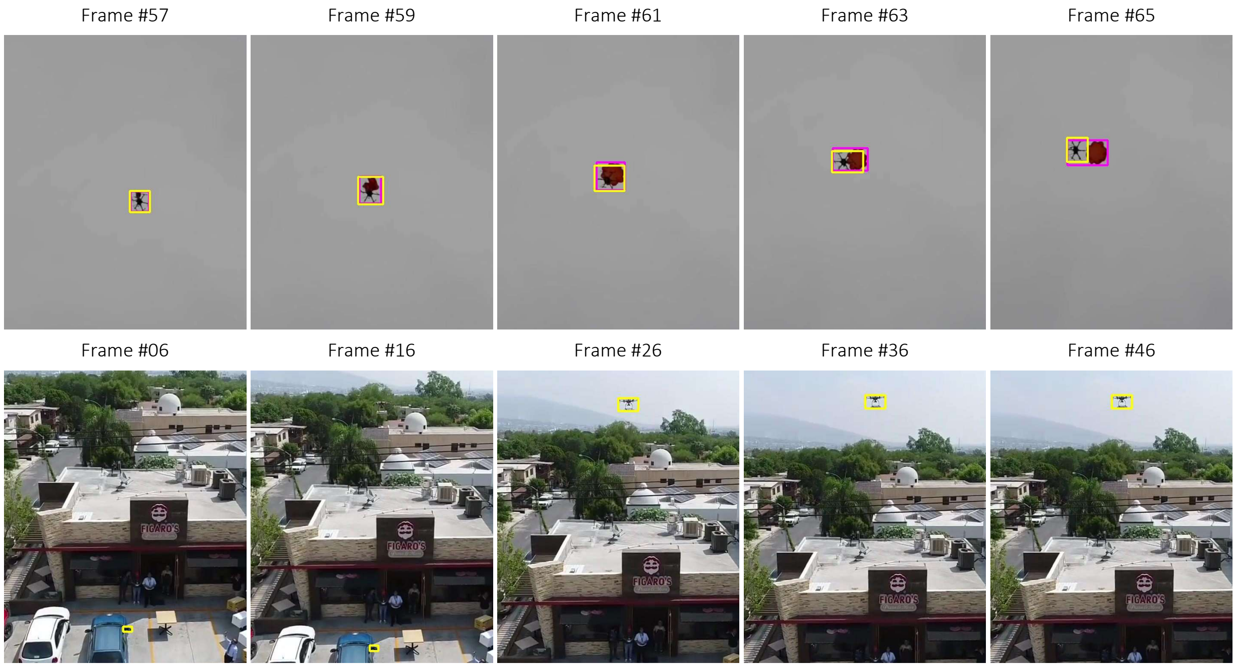

Figure II. Exemplary comparison between SAMURAI with first-frame YOLO-FEDER FusionNet initialization (magenta) and its detector-augmented extension (yellow), illustrating the benefits of continuous decoder-based prompting. Top row: First-frame initialization propagates erroneous masks under appearance variations (magenta), while continuous prompting corrects drift by re-aligning with detector outputs (yellow). Bottom row: First-frame initialization fails to recover from erroneous starting box, whereas detector-augmented SAMURAI leverages ongoing detections to reestablish accurate tracking (cf. frame 26, yellow).

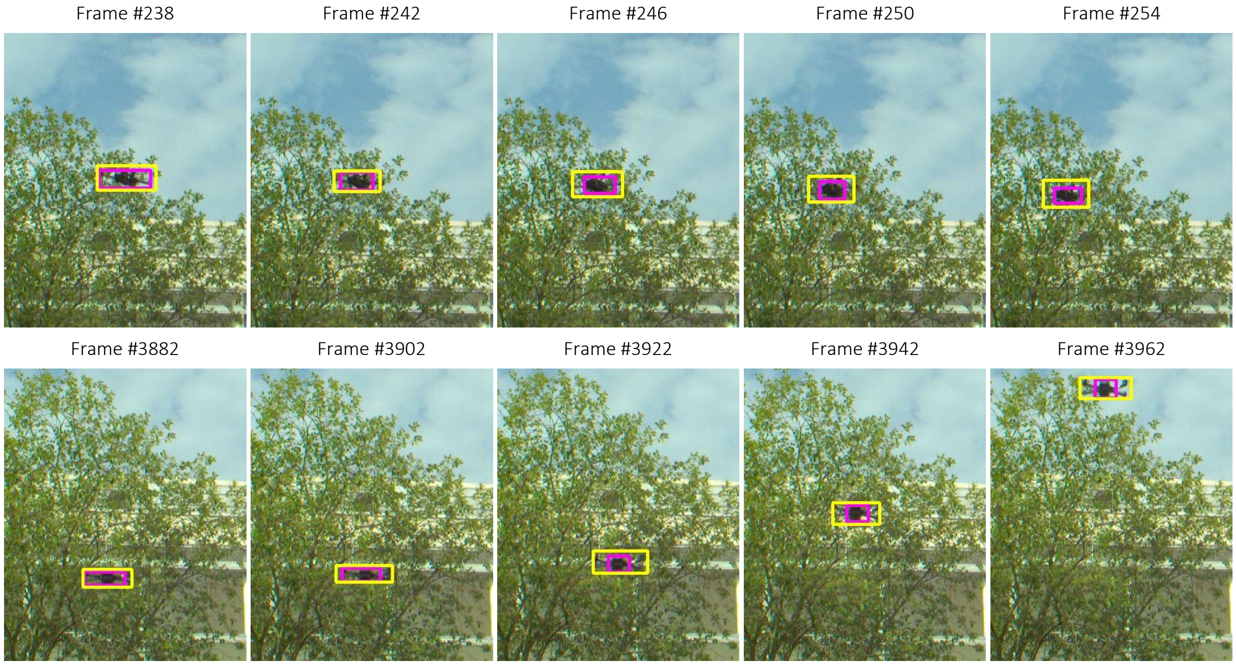

Figure III. Comparison between SAMURAI with first-frame-only initialization (magenta) and detector-augmented SAMURAI (yellow) with bounding-box averaging. The detector-augmented variant remains accurate in textured regions (*e.g.*, tree crowns), whereas first-frame initialization tends to degrade to partial drone-body predictions.

Table III. Performance of SAMURAI and its detector-augmented extension (✓, 4th column) on sequences from R1 and R2. *GT* (3rd column) denotes initialization with ground truth, while *D* denotes detector-based initialization using the first YOLO-FEDER FusionNet prediction as the bounding-box prompt. Best results are highlighted in **bold**.

				Tracking Metrics			Detection Metrics				
Dataset	Seq.	Init. Method	Detector Augmentation	S ↑	P ↑	P_{norm} ↑	mAP ↑ @0.25	@0.5	@0.5-0.95	FNR ↓	FDR ↓
R1	POS3	GT	–	0.124	0.200	0.213	0.584	0.390	0.120	0.784	0.007
		D	–	0.289	0.508	0.548	0.740	0.395	0.115	0.448	0.013
		D	✓	**0.560**	**0.762**	**0.808**	**0.907**	**0.792**	**0.352**	**0.179**	**0.005**
	POS7	GT	–	0.360	0.551	0.591	0.793	0.597	0.187	0.398	0.002
		D	–	0.397	0.622	0.668	0.829	0.596	0.177	0.319	0.002
		D	✓	**0.710**	**0.874**	**0.930**	**0.970**	**0.914**	**0.499**	**0.052**	**0.002**
R2	POS3	GT	–	0.178	0.339	0.670	0.480	0.071	0.012	0.614	0.631
		D	–	0.168	0.344	0.658	0.481	0.044	0.007	0.621	0.636
		D	✓	**0.454**	**0.833**	**0.888**	**0.934**	**0.427**	**0.101**	**0.116**	**0.026**
	POS7	GT	–	0.578	0.846	0.972	0.994	0.726	0.236	0.009	0.001
		D	–	0.580	**0.846**	**0.972**	**0.994**	0.730	0.239	**0.009**	**0.001**
		D	✓	**0.684**	0.831	0.966	0.984	**0.861**	**0.427**	0.015	0.007

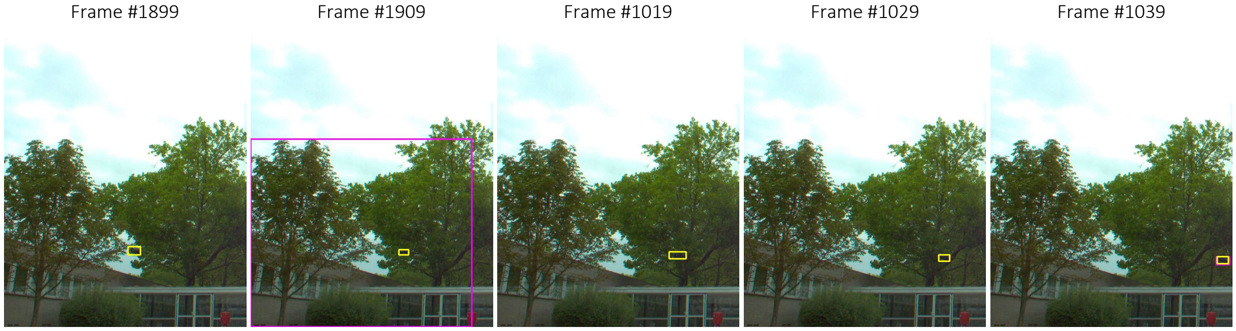
Frame #1899
Frame #1909
Frame #1019
Frame #1029
Frame #1039

Figure IV. SAMURAI with first-frame YOLO-FEDER FusionNet initialization (magenta) yields unstable (occasionally oversized) predictions when the drone leaves and reenters the FOV, whereas the detector-augmented SAMURAI (yellow) maintains robust tracking.

B. Qualitative Analysis of Detector-Augmented SAMURAI Limitations

Fig. V presents representative tracking failure cases of SAMURAI's detector-augmented extension on R1 (top) and R2 (bottom). These failures predominantly occur in scenarios where drone targets appear at very small scales or are partially occluded, resulting in limited visual evidence for reliable association and thus in tracking instabilities.

C. Detection Performance

Tabs. IV and V report the sequence-level detection performance of YOLO-FEDER FusionNet [7, 8] on the custom datasets R1 and R2, as well as on the publicly available DUT Anti-UAV dataset [13] (tracking subset). On R1 and R2, the detector achieves consistently high mAP with low FNRs and FDRs for POS7, whereas POS3 exhibits comparatively higher FNRs. Results on the DUT Anti-UAV dataset confirm this trend: most sequences achieve near-perfect detection at lower IoU thresholds and retain competitive performance under stricter evaluation. Except for video12 and video16, both FNR and FDR remain consistently low across sequences. Overall, YOLO-FEDER FusionNet provides promising detection performance across diverse conditions.

Table IV. YOLO-FEDER FusionNet performance on R1 and R2.

Dataset	Seq.	mAP ↑ @0.25	@0.5	@0.5-0.95	FNR ↓	FDR ↓
R1	POS3	0.782	0.759	0.342	0.312	0.058
	POS7	0.944	0.925	0.503	0.088	0.046
R2	POS3	0.902	0.433	0.117	0.214	0.007
	POS7	0.983	0.906	0.449	0.078	0.013

When compared to SAMURAI with GT initialization, YOLO-FEDER FusionNet achieves superior performance, particularly on the custom datasets R1 and R2 (cf. Tabs. III and IV). However, in combination with SAMURAI – also referred to as detector-augmented SAMURAI – additional improvements are obtained beyond standalone YOLO-FEDER FusionNet. While gains in bounding-box localization are modest, with mAP values comparable to or slightly exceeding those of YOLO-FEDER FusionNet, the most sig-

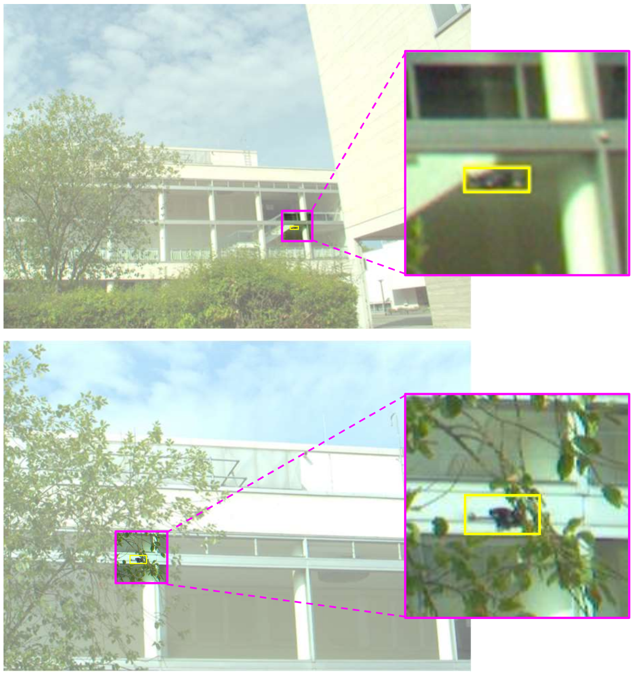

Figure V. Qualitative examples of representative tracking failure cases for detector-augmented SAMURAI. Zoomed-in regions are shown on the right-hand side to enhance the visibility of small-scale objects. Yellow bounding boxes denote GT drone localization. (top: R1; bottom: R2)

Table V. YOLO-FEDER FusionNet performance on DUT Anti-UAV [13].

Seq.	mAP ↑			FNR ↓	FDR ↓
	@0.25	@0.5	@0.5-0.95		
video01	0.972	0.964	0.661	0.059	0.012
video02	0.995	0.995	0.569	0.000	0.000
video03	0.995	0.995	0.743	0.000	0.000
video04	0.995	0.995	0.829	0.006	0.015
video05	0.967	0.942	0.500	0.056	0.025
video06	0.995	0.995	0.861	0.000	0.000
video07	0.994	0.994	0.783	0.008	0.006
video08	0.995	0.995	0.729	0.003	0.000
video09	0.995	0.995	0.844	0.000	0.010
video10	0.995	0.995	0.740	0.000	0.020
video11	0.993	0.993	0.684	0.000	0.000
video12	0.933	0.601	0.184	0.113	0.065
video13	0.982	0.549	0.109	0.060	0.012
video14	0.995	0.986	0.390	0.000	0.031
video15	0.995	0.964	0.330	0.002	0.001
video16	0.919	0.761	0.199	0.157	0.018
video17	0.983	0.827	0.228	0.060	0.034
video18	0.995	0.837	0.335	0.010	0.005
video19	0.995	0.605	0.134	0.000	0.003
video20	0.995	0.964	0.441	0.006	0.042

nificant benefit is reflected in FNR, with reductions of up to 41.99% on R1 and R2.

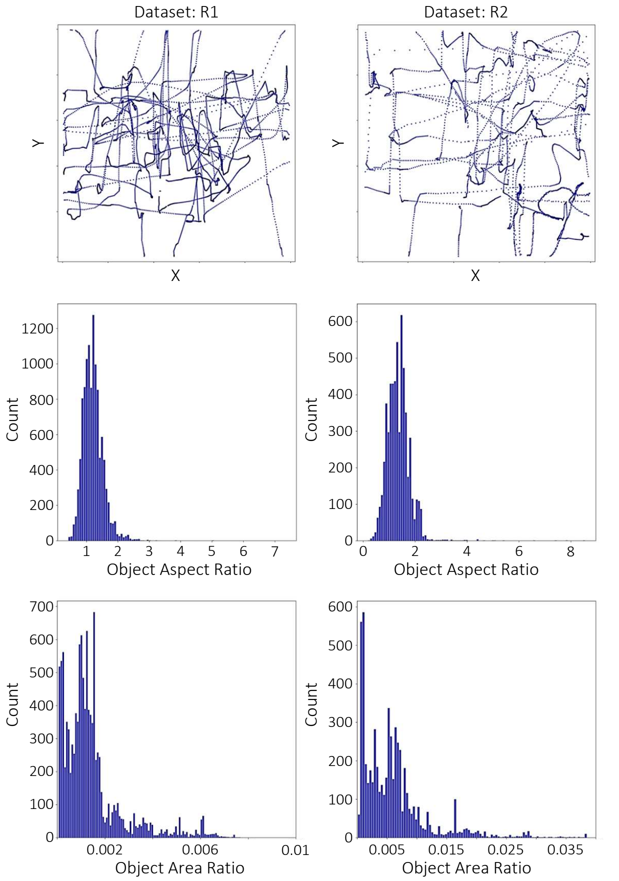

Figure VI. Distributions of drone positions (top), object aspect ratios (middle), and object area ratios (bottom) across all sequences in R1 and R2.

D. Custom Dataset Details

The custom-recorded datasets R1 and R2 (cf. Sec. 3.3, main paper) encompass urban outdoor environments with varying structural compositions, combining architectural features, vegetation, and open-sky regions (see Fig. VII). R1 is characterized by more pronounced architectural structure, such as multi-story facades, accompanied by moderate vegetation. In contrast, R2 contains denser foliage and greater natural clutter, with buildings more frequently occluded, resulting in visually richer and more complex scenes that are predominantly vegetation-driven. This makes R2 especially valuable for drone-detection research, as identifying drones against highly textured vegetative backgrounds is inherently challenging due to reduced visual contrast and pronounced camouflage effects [7]. Beyond the environmental variations, the drones in R1 and R2 also exhibit distinct spatial and geometric properties (see Fig. VI): R1 features denser and more complex motion patterns, with drone instances tending toward more compact aspect ratios

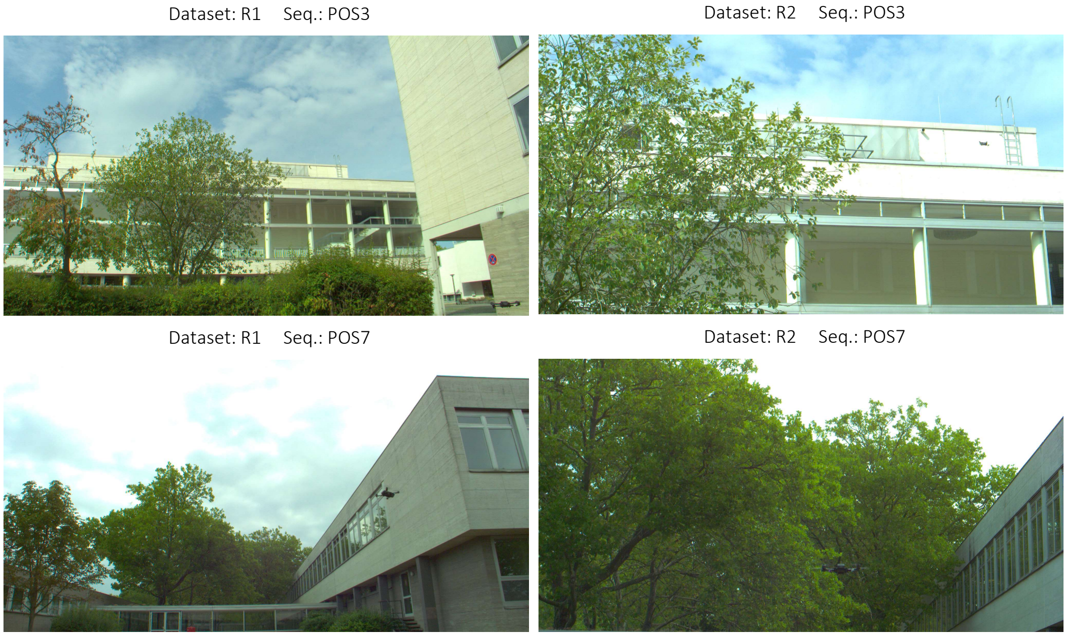

Figure VII. Representative frames from the four custom-recorded sequences in R1 and R2, highlighting the sequence-level FOV and the visual attributes of the surrounding environment.

and slightly larger relative areas. Conversely, R2 features a broader aspect-ratio distribution and smaller area ratios. Both datasets are released as part of this work and are publicly available at [9].

E. Visual Variability of DUT Anti-UAV

The publicly available DUT Anti-UAV dataset encompasses diverse outdoor environments, ranging from sky-dominant scenes (Fig. VIII, 1st row) to forested areas with dense vegetation (2nd row), as well as suburban and urban settings featuring varied architectural elements (rows 3-4). It captures multiple drone types from diverse viewpoints – varying in angle, distance, and altitude – leading to substantial changes in scale and appearance. Overall, DUT Anti-UAV exhibits significant diversity in drone appearance, background texture, illumination, and overall scene complexity.

References

[1] Luca Bertinetto, Jack Valmadre, João F. Henriques, et al. Fully-Convolutional Siamese Networks for Object Tracking. In *Eur. Conf. Comput. Vis. Workshops*, pages 850–865, 2016. 1

[2] Goutam Bhat, Martin Danelljan, Luc Van Gool, et al. Learning Discriminative Model Prediction for Tracking. In *IEEE/CVF Int. Conf. Comput. Vis.*, pages 6181–6190, 2019. 1

[3] Xin Chen, Bin Yan, Jiawen Zhu, et al. Transformer Tracking. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 8122–8131, 2021. 1

[4] Kenan Dai, Yunhua Zhang, Dong Wang, et al. High-Performance Long-Term Tracking With Meta-Updater. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 6297–6306, 2020. 1

[5] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, et al. ECO: Efficient Convolution Operators for Tracking. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 6931–6939, 2017. 1

[6] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, et al. ATOM: Accurate Tracking by Overlap Maximization. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 4655–4664, 2019. 1

[7] Tamara R. Lenhard, Andreas Weinmann, Stefan Jäger, et al. YOLO-FEDER FusionNet: A Novel Deep Learning Architecture for Drone Detection. In *IEEE Int. Conf. Image Process.*, pages 2299–2305, 2024. 5, 6

[8] Tamara R. Lenhard, Andreas Weinmann, and Tobias Koch. Performance Optimization of YOLO-FEDER FusionNet for Robust Drone Detection in Visually Complex Environments. *ArXiv*, abs/2509.14012, 2025. 5

[9] Tamara R. Lenhard, Andreas Weinmann, Hichem Snoussi, and Tobias Koch. Long-Duration Drone Tracking Dataset. *Zenodo*, 2026. https://doi.org/10.5281/zenodo.17182190. 7

[10] Bo Li, Wei Wu, Qiang Wang, et al. SiamRPN++: Evolution of Siamese Visual Tracking With Very Deep Networks. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 4277–4286, 2019. 1

[11] Bin Yan, Haojie Zhao, Dong Wang, et al. 'Skimming-Perusal' Tracking: A Framework for Real-Time and Robust Long-Term Tracking. In *IEEE/CVF Int. Conf. Comput. Vis.*, pages 2385–2393, 2019. 1

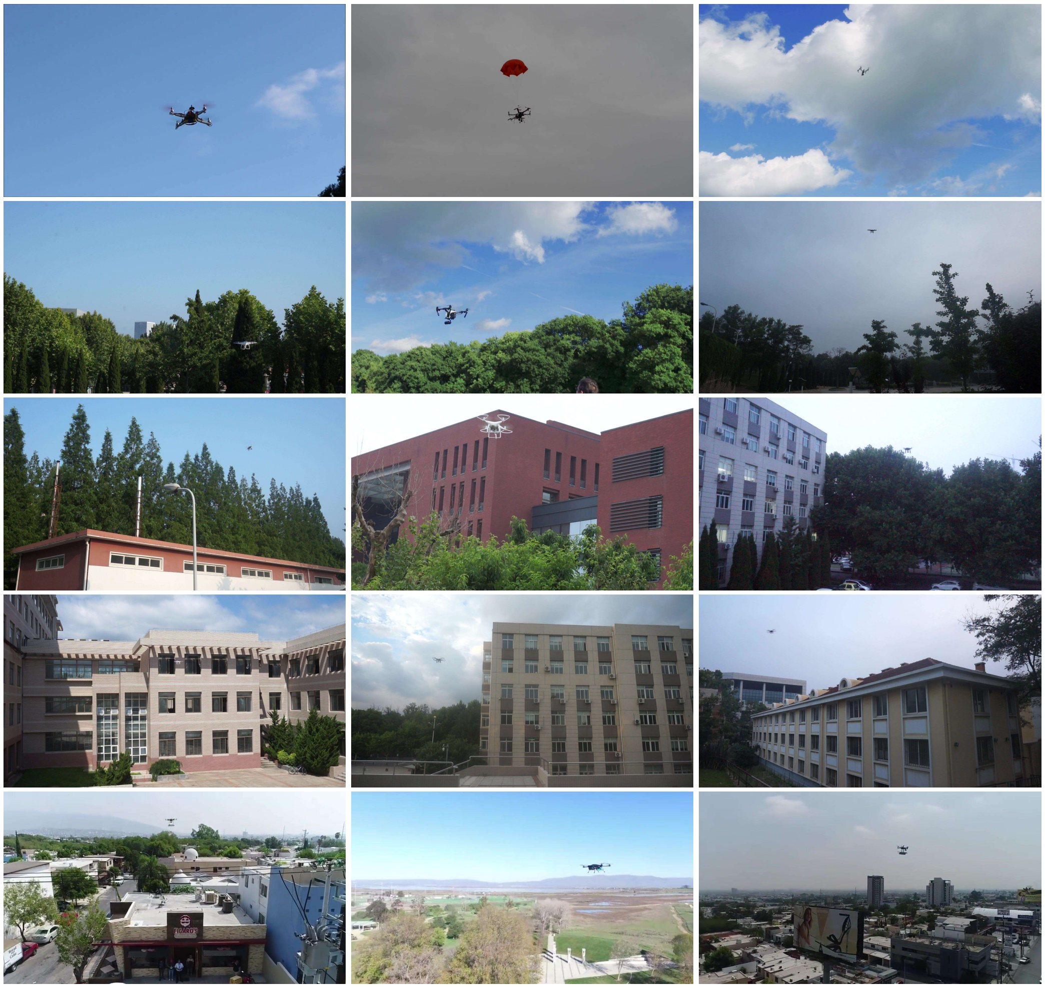

Figure VIII. Visual diversity of the DUT Anti-UAV dataset [13].

[12] Cheng-Yen Yang, Hsiang-Wei Huang, Wenhao Chai, et al. SAMURAI: Adapting Segment Anything Model for Zero-Shot Visual Tracking with Motion-Aware Memory. *ArXiv*, abs/2411.11922, 2024. 1

[13] Jie Zhao, Jingshu Zhang, Dongdong Li, et al. Vision-Based Anti-UAV Detection and Tracking. *IEEE Trans. Intell. Transp. Syst.*, 23(12):25323–25334, 2022. 1, 2, 3, 5, 6, 8