Benchmarking EfficientTAM on FMO datasets

Senem Aktas, Charles Markham, John McDonald & Rozenn Dahyot

Department of Computer Science Maynooth University Ireland

Abstract

Fast and tiny object tracking remains a challenge in computer vision and in this paper we first introduce a JSON metadata file associated with four open source datasets of Fast Moving Objects (FMOs) image sequences. In addition, we extend the description of the FMOs datasets with additional ground truth information in JSON format (called FMOX) with object size information. Finally we use our FMOX file to test a recently proposed foundational model for tracking (called EfficientTAM) showing that its performance compares well with the pipelines originally taylored for these FMO datasets. Our comparison of these state-of-the-art techniques on FMOX is provided with Trajectory Intersection of Union (TIoU) scores. The code and JSON is shared open source allowing FMOX to be accessible and usable for other machine learning pipelines aiming to process FMO datasets.

Keywords: Fast Moving Object, FMO Dataset, Labeling, EfficientTAM

1 Introduction

The tracking and detection of small and/or fast-moving objects (FMOs) remains relatively underexplored, particularly in comparison to research on general objects [Zhang et al., 2022]. Specific challenges occur in such scenarios including motion blur, which can distort the appearance of objects and complicate the performance of general object detectors and trackers. Moreover, data annotation becomes more difficult due to the size and motion blur of FMOs, leading to a scarcity of available annotated datasets [Zhang et al., 2022, Yu et al., 2020, Rozumnyi et al., 2017]. Despite these issues, several public FMO datasets, including Falling Object [Kotera et al., 2020], FMOv2 [Rozumnyi et al., 2017], TbD [Kotera et al., 2019], and TbD-3D [Rozumnyi et al., 2020] are now available (see Section 2). Unlike larger objects, small objects often suffer from reduced visibility and lower image cover rate, leading to fewer appearance cues and increased background interference. Considering both issues, fast and small object tracking, are therefore still challenging for modern tracking techniques [Zhang et al., 2022, Haalck et al., 2024, Yu et al., 2020].

In this work, we define and characterize the terms *small* and *fast* in the context of object tracking in videos. Second, we provide a combined and extended dataset, which we term FMO eXtended (FMOX), combining the four datasets mentioned above with a more informative metadata encoding. This encoding is provided through a simplified JSON format, allowing the development of Machine Learning dataloaders and pipelines that operate over the entire unified dataset (Section 3). Finally, we assess EfficientTAM (Efficient Track Anything Model) [Xiong et al., 2024], a recent state-of-the-art technique for tracking, on these datasets, benchmarking its performance against leading FMO specific techniques from the literature. Code and FMOX JSON are made available at `https://github.com/CVMLmu/FMOX/`.

2 Fast-Moving Object Datasets

The following datasets[1] are considered here to create FMOX.

[1] available at `https://cmp.felk.cvut.cz/fmo/`

Falling Objects [Kotera et al., 2020]. The dataset comprises static image frames from video sequences recorded at 25/30 Frame Per Second (FPS). It comprises a collection of 6 different objects, including box, cell, key, and rubber, which are dropped from a table, as shown in Figure 1. It includes two sets of images as high-speed and low-speed sequences, with ground truth trajectories provided in text format.

FMOv2 [Rozumnyi et al., 2017]. This dataset includes 19 sports sequences featuring a variety of small objects, primarily balls, such as volleyballs and tennis balls, but also objects such as darts and frisbees, with PNG images of varying lengths. The ground truth trajectories are provided in Matlab for each sequence, with object mask PNG format and its run length encoding compressed text file. In the dataset, FMO displacement is evenly spread between 0-150 pixels while bounding boxes between consecutive frames have nearly zero intersection each time. Figure 2 illustrates samples from the dataset, with the FMOv2 extension sequence samples highlighted in red.

Figure 1: Falling Objects [Kotera et al., 2020] samples.

Figure 2: FMOv2 [Rozumnyi et al., 2017] samples.

TbD [Kotera et al., 2019]. This dataset comprises striking golf, tennis, volleyball and badminton balls as well as a falling cube and coin. 14 image sequences are provided with ground-truth trajectories including 12 sequences from sports videos with mostly spherical objects and almost no appearance changes over time. These 30 FPS low-speed videos are obtained from 240 FPS high-speed videos using temporal averaging. Semi-manual annotation are applied for creating ground truth by labeling the first frame and applying a tracker, and then correcting the annotations.

TbD-3D [Rozumnyi et al., 2020]. Similar to TbD, the TbD-3D dataset includes 10 PNG image sequences, focusing on objects moving in 3D that undergo significant changes in appearance within a single low-speed videos. 9 of these sequences depict a spherical object, specifically a ball, while the last sequence features a circular object. The 6D poses (3D position and rotation) of FMOs are provided as ground truth with manually annotated 3D object location (2D position and radius) and estimated 3D object rotation. Videos were captured in raw format with a high-speed camera operating at 240 FPS. Both high-speed and low-speed versions of the image sequences are provided. Ground truths are available in both MATLAB format and text format as trajectories for the low-speed image sequences. Figure 4 shows a sample frame from each sequence.

3 FMOX

In this section, we define FMOX, a JSON structured annotation format that simplifies the handling of the four public FMO datasets described in Section 2). It is designed to be easily interpretable and compatible, enabling researchers and practitioners to quickly understand the structure and content of the data, allowing them to use the datasets with minimal effort. We also provide object size categories as additional labeling for the datasets. By considering size as a key factor in the annotations, we aim to enhance the relevance of the dataset for various research applications, such as small object detection and tracking.

Figure 3: TbD [Kotera et al., 2019] samples.

Figure 4: TbD-3D [Rozumnyi et al., 2020] samples.

Object Size. In the computer vision community, the term "small objects" characterizes as covering an area of 32×32 pixels or less [Tong et al., 2020, Lin et al., 2014, Zhang et al., 2022]. For instance, the Microsoft COCO benchmark [Lin et al., 2014] categorizes objects into three size categories for evaluation as `small`, `medium`, and `large` based on the dimensions of the bounding boxes that encapsulate the objects. Object size categorization facilitates the analysis of detector and model performance across various object sizes [Zhu et al., 2016]. Table 1 presents object size categories as defined in the literature. Table 2 presents our object size categories used in FMOX.

Table 3 summarizes the FMOX statistics, primarily focusing on object size information. Additionally, the listing 1 displays the structure of FMOX for the four analyzed FMO datasets.

Study	Resolution	Categorization
Traffic sign [Zhu et al., 2016]	2048×2048	`small` (0,32], `medium` (32,96], `large` (96,400]
[Tong et al., 2020]		`small` (0,32], `medium` (32,96], `large` (96,∞]
TinyPerson [Yu et al., 2020]	1920×1080	`tiny` [2,20] : `tiny1` [2,8], `tiny2` [8,12], `tiny3` [12,20], `small` [20,32], `all` [2,∞]
[Ying et al., 2025]		`extremely tiny` [1,8), `tiny` [8,16), `small` [16,32)

Table 1: Examples of object size category definitions using side length of square bounding box.

`Extremely Tiny`	`Tiny`	`Small`	`Medium`	`Large`
$[1 \times 1, 8 \times 8)$	$[8 \times 8, 16 \times 16)$	$[16 \times 16, 32 \times 32)$	$[32 \times 32, 96 \times 96)$	$[96 \times 96, \infty)$

Table 2: FMOX object size categories.

Fast Movement. Fast-Moving Objects (FMOs) exhibit significant displacement that exceeds their size within the exposure time across consecutive frames [Rozumnyi et al., 2017]. Movement of FMOs result in pronounced 6D pose changes in sequential frames; specifically, the 3D rotation of the object alters due to high angular velocity, causing it to appear partially visible and manifest as shadowy streaks [Rozumnyi et al., 2020, Rozumnyi et al., 2017]. According to [Rozumnyi et al., 2017], FMOs are perceived as translucent lines that are larger than their actual size, which is typically less than 100 pixels. Their motion in the x, y, z planes leads to substantial changes in their 3D location, affecting their 2D position and radius, as well as the distance between the object and the camera [Rozumnyi et al., 2020].

This streaking effect results in motion blur, complicating the ability to discern the object's features, such as its shape and edges, particularly in a single image [Rozumnyi et al., 2017]. Evaluation results from their proposed model indicate that significant motion against a distinct background yields the best tracking outcomes. Furthermore, it has been noted [Rozumnyi et al., 2017] that FMOs may go undetected when their motion is minimal or when the background color closely resembles that of the object. Additionally, local movements of large non-FMOs can sometimes be misidentified as FMOs.

Statistical analysis of FMO dataset [Rozumnyi et al., 2017] demonstrated that in two consecutive frames the overlap of the ground truth bounding boxes is zero. The distance between the center of the object is on average ten times higher than non-FMO datasets and the displacement is uniformly spread between 0-150 pixels. However, in two consecutive frames of non-FMO datasets, ground truth bounding boxes overlap is close to one or above 0.5 in 94% of cases, and the displacement is below 10 pixels in 91% of cases. In small and fast-moving object benchmark [Zhang et al., 2022], it is expected that target center moves by at least 50% of its size.

FMOX Structure. The four examined FMO datasets exhibit different formats, including variations in annotation styles such as text annotations, segmentation masks and Matlab versions. These variations in annotations could limit the reproducibility of experiments, which inspired us to develop an easy-to-use JSON format annotation called FMOX, with the goal of improving accessibility, readability, and usability for users. The JSON structure (shown in Listing 1) contains objects' bounding boxes as (x_1, y_1, x_2, y_2) format and object size category. To obtain the bounding boxes for FMOv2, the object mask images provided are processed with OpenCV's contour detection function `findContours` (with parameters `RETR_EXTERNAL`, `CHAIN_APPROX_SIMPLE` and threshold value of 70). For the TbD dataset, ground-truth trajectory text annotations are utilized to obtain bounding boxes which consist of object annotations for the entire sequence of frames (and not only for FMO frames). For the Falling Object and TbD-3D datasets, the data loading component of the repository `fmo-deblurring-benchmark`[2] is leveraged.

Listing 1: Structure of FMOX

```
{ "databases": [ {
      "dataset_name": "Falling_Object",
      "version": "1.0",
      "description": "Falling_Object annotations.",
      "sub_datasets": [
                {"subdb_name": "v_box_GTgamma",
                 "images": [
                    {
                        "img_index": 1,
                        "image_file_name": "00000027.png",
                        "annotations": [
                            {
                                "bbox_xyxy": [161, 259, 245, 333],
                                "object_wh": [84, 74],
                                "size_category": "medium"
                            }
                        ]},
                    {
                        "img_index": 2,
                        "image_file_name": "00000028.png",
                        "annotations": [ ---- ]
                    }   ] } -------- ]  }, ------- }
```

4 EfficientTAM performance on FMOX

To provide a baseline measure of performance, we test Efficient Track Anything Model (EfficientTAM) [Xiong et al., 2024] on FMOX. No additional training is conducted to the pretrained model `efficienttam_s` which we use with its default parameters[3]. The FMOX ground-truth annotation from the first image of each sequence is used to initialize the EfficientTAM with a target to track in that sequence. Both point (chosen as the center of the bounding box) and bounding box initializations were assessed for initializing the target. EfficientTAM

[2] `https://github.com/rozumden/fmo-deblurring-benchmark` [Rozumnyi et al., 2021c]

[3] `https://github.com/yformer/EfficientTAM`.

	Sequence Name	Analysis			
		Total Frame Number	FMO Exists Frame Number	**(Ours)** TIoU (↑)	Object Size Levels
Falling Object	v_box_GTgamma	62	22	**0.904**	{'medium': 22}
	v_cell_GTgamma	62	14	**0.730**	{'medium': 14}
	v_key_GTgamma	62	19	**0.651**	{'medium': 19}
	v_marker_GTgamma	62	11	**0.799**	{'medium': 11}
	v_pen_GTgamma	62	13	**0.558**	{'large': 8, 'medium': 5}
	v_rubber_GTgamma	62	15	**0.614**	{'medium': 15}
FMOv2	atp_serves^{+}	655	463	0.135 *	{'extremely_tiny': 1, 'small': 79, 'tiny': 383}
	blue	53	21	**0.775**	{'large': 1, 'medium': 20}
	darts1	75	51	**0.738**	{'large': 3, 'medium': 33}
	darts_window1	50	9	0.023 *	{'medium': 5}
	frisbee^{+}	100	68	0.490	{'large': 16, 'medium': 4}
	hockey^{+}	350	323	**0.527**	{'extremely_tiny': 48, 'small': 6, 'tiny': 7}
	more_balls $^{+}$	300	287	*Not applied*	{'medium': 129, 'small': 1112, 'tiny': 49}
	ping_pong_paint	120	111	0.036	{'extremely_tiny': 1, 'medium': 68, 'small': 6, 'tiny': 1}
	ping_pong_side^{+}	445	444	**0.629**	{'extremely_tiny': 2, 'medium': 172, 'small': 183, 'tiny': 79}
	ping_pong_top^{+}	350	350	0.396	{'extremely_tiny': 1, 'large': 2, 'medium': 242, 'small': 59, 'tiny': 148}
	softball	96	35	0.009 *	{'medium': 14, 'small': 13, 'tiny': 1}
	squash	250	242	*NAN*	{'extremely_tiny': 129, 'tiny': 5}
	tennis1	116	91	*NAN*	{'extremely_tiny': 63, 'tiny': 1}
	tennis2	278	274	0.005	{'extremely_tiny': 151, 'small': 6, 'tiny': 62}
	tennis_serve_back^{+}	156	78	0.312	{'extremely_tiny': 31, 'small': 10, 'tiny': 18}
	tennis_serve_side	68	35	**0.821**	{'medium': 1, 'small': 12, 'tiny': 5}
	volleyball1	50	33	**0.905**	{'large': 12, 'medium': 1}
	volleyball_passing	66	66	**0.895**	{'large': 4, 'medium': 62}
	william_tell	119	67	**0.783**	{'extremely_tiny': 1, 'large': 7, 'medium': 7, 'small': 5, 'tiny': 12}
TbD	VS_badminton_white_GX010058-8	125	40	0.010 *	{'tiny': 56, 'extremely_tiny': 36, 'small': 27, 'medium': 6}
	VS_badminton_yellow_GX010060-8	125	57	0.265	{'tiny': 63, 'extremely_tiny': 36, 'small': 19, 'medium': 7}
	fall_cube	28	20	**0.902**	{'medium': 4, 'small': 2, 'tiny': 5, 'extremely_tiny': 17}
	hit_tennis	57	30	**0.878**	{'extremely_tiny': 47, 'tiny': 9, 'small': 1}
	hit_tennis2	26	26	0.094 *	{'extremely_tiny': 4, 'tiny': 14, 'small': 5, 'medium': 3}
	VS_pingpong_GX010051-8	95	58	**0.756**	{'small': 36, 'tiny': 46, 'extremely_tiny': 13}
	VS_roll_golf-gc-12	16	16	**0.858**	{'small': 3, 'medium': 5, 'tiny': 3, 'extremely_tiny': 5}
	VS_tennis_GX010073-8	118	38	**0.807**	{'small': 66, 'tiny': 32, 'extremely_tiny': 20}
	throw_floor	73	40	0.003 *	{'medium': 15, 'large': 1, 'small': 6, 'tiny': 7, 'extremely_tiny': 44}
	throw_soft	75	60	0.008 *	{'small': 16, 'large': 1, 'medium': 13, 'tiny': 7, 'extremely_tiny': 38}
	throw_tennis	71	45	0.003 *	{'medium': 16, 'small': 19, 'tiny': 9, 'extremely_tiny': 27}
	VS_volleyball_GX010068-12	72	41	**0.872**	{'small': 11, 'medium': 5, 'tiny': 25, 'extremely_tiny': 31}
TbD-3D	HighFPS_GT_depth2	48	48	**0.860**	{'large': 2, 'medium': 46}
	HighFPS_GT_depthb2	81	81	**0.823**	{'medium': 81}
	HighFPS_GT_depthf1	46	46	**0.833**	{'medium': 46}
	HighFPS_GT_depthf2	50	50	**0.816**	{'medium': 50}
	HighFPS_GT_depthf3	37	37	**0.816**	{'medium': 37}
	HighFPS_GT_out1	57	57	**0.899**	{'large': 1, 'medium': 56}
	HighFPS_GT_out2	50	50	**0.909**	{'medium': 50}
	HighFPS_GT_outa1	47	47	**0.923**	{'large': 14, 'medium': 33}
	HighFPS_GT_outb1	41	41	**0.830**	{'medium': 41}
	HighFPS_GT_outf1	60	60	**0.895**	{'medium': 60}

Table 3: FMOX dataset information with Trajectory-Intersection of Union (TIoU) results computed on each sequence with EfficientTAM [Xiong et al., 2024] (column labelled **(Ours)**; TIoU above 0.5 are in bold font). *Not applied*: multi object sequence (request multi initialization with no id for comparison); *: tracker could not initialized due to high motion blur. *NAN* is one of the outputs of the TIoU function used. Sequences noted $^{+}$ indicate that multiple objects occur but labels provided are not with object instance ids.

first performs segmentation on the first frame to then track the detected region in following frames. When the initialization is point-based, it is often observed that the detection of the target object fails whereas when using

the bounding box, the segmentation correctly captures the object of interest. Hence, tracking is tested with bounding box initialization on the first frame. Note however, that even with the bounding box, if the frame exhibits strong motion blur, EfficientTAM can still fail to segment the object to be tracked on the first frame. The performance of EfficientTAM, as measured by the Trajectory-Intersection of Union (TIoU) [Rozumnyi et al., 2021a], is reported in Tables 3 and 4, as TIoU is employed in studies of fast-moving objects to evaluate FMO datasets.

To evaluate the performance of EfficientTAM on four datasets, we utilized the Defmo[4] pipeline to perform TIoU [Rozumnyi et al., 2021a] calculations with FMOX. To feed the pipeline, FMOX object bounding boxes are transformed into trajectories by calculating the center coordinates of the bounding boxes and then interpolating according to the number of segment (nsplits) parameter. Coding details are available on GitHub `https://github.com/CVMLmu/FMOX/`.

Datasets \ Studies	Defmo [Rozumnyi et al., 2021c]	FmoDetect [Rozumnyi et al., 2021b]	TbD [Kotera et al., 2019]	TbD-3D [Rozumnyi et al., 2020]	**(Ours)** with EfficientTAM
Falling Object	0.684**	N/A	0.539	0.539	0.7093*
TbD	0.550**	(a) 0.519 (b) 0.715*	0.542	0.542	0.4546
TbD-3D	0.879*	N/A	0.598	0.598	0.8604**

(a) real-time with trajectories estimated by the network, (b) with the proposed deblurring, (N/A) : not defined

Table 4: Average TIoU (↑) performance comparison of our results with EfficientTAM [Xiong et al., 2024] on FMO datasets with FMOX (column **(Ours)**). For each dataset, * indicate best result and ** second best result. Other Average TIoUs shown are directly extracted from the cited papers.

Overall EfficientTAM performs very well in particular with the Falling Object and TbD-3D datasets (cf. Fig. 5 and Tab. 4). However, a general issue with the FMOv2 and TbD datasets is that some sequences, could not be initialized due to strong motion-blurred FMOs (see notation asterisk (*) in Table 3). We believe that initializing the EfficientTAM with less motion-blurred FMOs frames would yield higher scores. Moreover, for the FMOv2 dataset, some sequences contain multiple FMOs, such as `frisbee` and `more_balls`. However, without unique IDs for these objects, it is not possible to effectively compare the ground truth with the estimated results. In the `frisbee` sequence, two frisbee objects travel nearly the same distance and direction but at different times. We initialized the tracker for the first object, which covers almost half of the trajectory, resulting in a TIoU value of approximately 0.5. In the `more_balls` sequence, multiple balls appear and disappear repeatedly, which is why we have not included this sequence in our evaluation for the time being. Also, in FMOX, we have corrected several masks in the `ping_pong_paint` sequence, which contained only a small mask of ball from a different tennis game that was interfering with the tracking initialization. Similarly, the `william_tell` sequence was distorted by extra masks — specifically, traces left by pieces of the apple that was shot — misdirecting the trajectory of the main target (bullet). After making these mask corrections, the TIoU improved to 0.783.

5 Conclusion

We propose an enhanced metadata description file, FMOX, associated with four video datasets featuring Fast Moving Objects (FMOs). Using FMOX, we evaluate the recently released foundation model, EfficientTAM, for tracking FMOs. Our experimental results demonstrate its competitive performance and limitations, including difficulties in initializing tracking for strongly motion-blurred objects, in challenging scenarios. Notably, EfficientTAM achieves superior performance without requiring specialized training or modifications to its default parameters, yielding average Trajectory-Intersection over Union (TIoU) scores of 0.7093 for Falling Objects

[4] `https://github.com/rozumden/DeFMO`.

and 0.8604 for TbD-3D datasets. These results showcase its effectiveness in tracking FMOs. Future work will investigate additional metrics to TIoU for performance assessment of trackers with FMOX, as well as evaluating impact of object size and motion. Some recommendations in using EfficientTAM can be made from this experiments such as best performance is obtained when the tracker is initialized on a non-blurry object in the image. Moreover in the cases where the tested sequence displays several instances of the same object (e.g. tennis balls), then the tracker can be distracted by a competing instance leading to a low TIoU for the sequence.

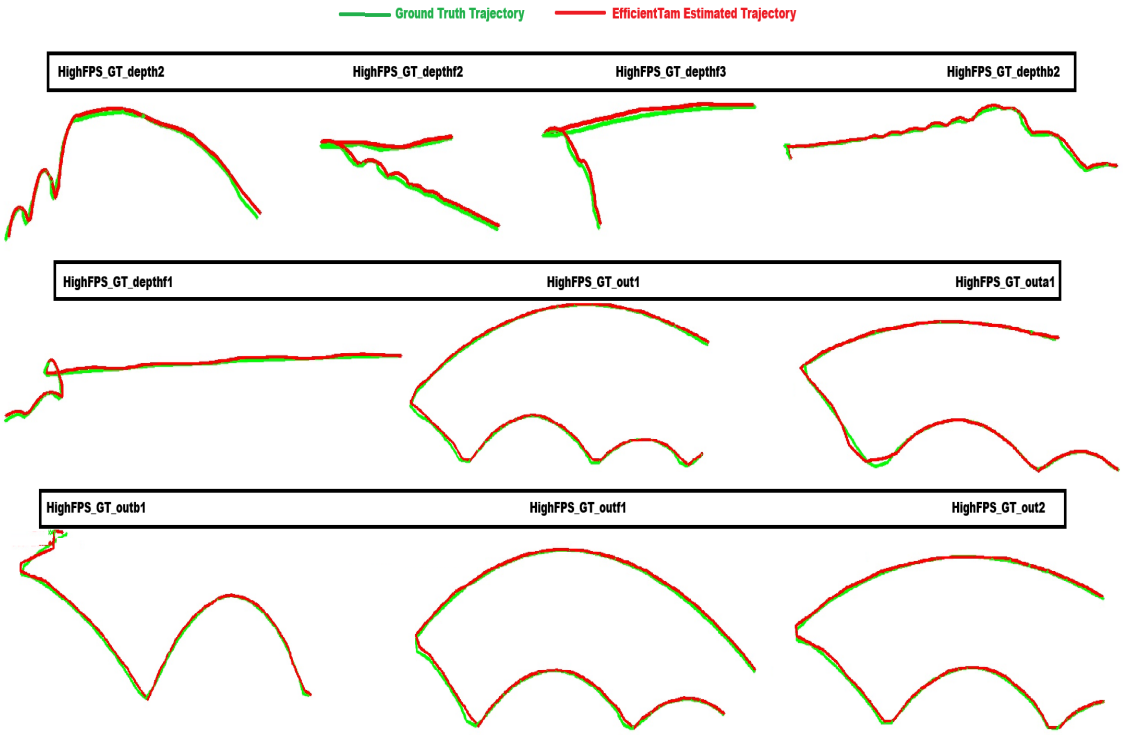

Figure 5: EfficientTAM estimated trajectories on TbD-3D dataset. Green color indicates ground truth trajectory while red color for EfficientTAM estimated trajectory. TIoU values are above 0.81 for all sequences. Objects (mostly balls) quite big, while having motion blur still object is pretty visible all along sequences.

Acknowledgments

This research was supported by funding through the Maynooth University Hume Doctoral Awards. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

[Haalck et al., 2024] Haalck, L., Thiele, S., and Risse, B. (2024). Tracking tiny insects in cluttered natural environments using refinable recurrent neural networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7126–7135.

[Kotera et al., 2020] Kotera, J., Matas, J., and Šroubek, F. (2020). Restoration of fast moving objects. *IEEE Transactions on Image Processing*, 29:8577–8589.

[Kotera et al., 2019] Kotera, J., Rozumnyi, D., Sroubek, F., and Matas, J. (2019). Intra-frame object tracking by deblatting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0.

[Lin et al., 2014] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer.

[Rozumnyi et al., 2020] Rozumnyi, D., Kotera, J., Sroubek, F., and Matas, J. (2020). Sub-frame appearance and 6d pose estimation of fast moving objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6778–6786.

[Rozumnyi et al., 2017] Rozumnyi, D., Kotera, J., Sroubek, F., Novotny, L., and Matas, J. (2017). The world of fast moving objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5203–5211.

[Rozumnyi et al., 2021a] Rozumnyi, D., Kotera, J., roubek, F., and Matas, J. (2021a). Tracking by deblatting. *International Journal of Computer Vision*, 129(9):25832604.

[Rozumnyi et al., 2021b] Rozumnyi, D., Matas, J., Šroubek, F., Pollefeys, M., and Oswald, M. R. (2021b). Fmodetect: Robust detection of fast moving objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3541–3549. `https://github.com/rozumden/FMODetect`.

[Rozumnyi et al., 2021c] Rozumnyi, D., Oswald, M. R., Ferrari, V., Matas, J., and Pollefeys, M. (2021c). Defmo: Deblurring and shape recovery of fast moving objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3456–3465. `https://github.com/rozumden/DeFMO`.

[Tong et al., 2020] Tong, K., Wu, Y., and Zhou, F. (2020). Recent advances in small object detection based on deep learning: A review. *Image and Vision Computing*, 97:103910.

[Xiong et al., 2024] Xiong, Y., Zhou, C., Xiang, X., Wu, L., Zhu, C., Liu, Z., Suri, S., Varadarajan, B., Akula, R., Iandola, F., et al. (2024). Efficient track anything. *arXiv preprint arXiv:2411.18933*. `https://yformer.github.io/efficient-track-anything/`.

[Ying et al., 2025] Ying, X., Xiao, C., An, W., Li, R., He, X., Li, B., Cao, X., Li, Z., Wang, Y., Hu, M., et al. (2025). Visible-thermal tiny object detection: A benchmark dataset and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[Yu et al., 2020] Yu, X., Han, Z., Gong, Y., Jan, N., Zhao, J., Ye, Q., Chen, J., Feng, Y., Zhang, B., Wang, X., et al. (2020). The 1st tiny object detection challenge: Methods and results. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 315–323. Springer.

[Zhang et al., 2022] Zhang, Z., Wu, F., Qiu, Y., Liang, J., and Li, S. (2022). Tracking small and fast moving objects: A benchmark. In *Proceedings of the Asian Conference on Computer Vision*, pages 4514–4530.

[Zhu et al., 2016] Zhu, Z., Liang, D., Zhang, S., Huang, X., Li, B., and Hu, S. (2016). Traffic-sign detection and classification in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2110–2118.