Leveraging Unreal Engine for UAV Object Tracking: The AirTrackSynth Synthetic Dataset

Mingyang Zhang[a], Kristof Van Beeck[b] and Toon Goedemé[c]
PSI-EAVISE Research Group, Department of Electrical Engineering, KU Leuven, Belgium
{*mingyang.zhang, kristof.vanbeeck, toon.goedeme*}*@kuleuven.be*

Keywords: Synthetic Data, UAV, Object Tracking, Multimodality, Deep Learning, Siamese Network.

Abstract: Nowadays, synthetic datasets are often used to advance the state-of-the-art in many application domains of computer vision. For these tasks, deep learning approaches are used which require vasts amounts of data. Acquiring these large annotated datasets is far from trivial, since it is very time-consuming, expensive and prone to errors during the labelling process. These synthetic datasets aim to offer solutions to the aforementioned problems. In this paper, we introduce our AirTrackSynth dataset, developed to train and evaluate deep learning models for UAV object tracking. This dataset, created using the Unreal Engine and AirSim, comprises 300GB of data in 200 well-structured video sequences. AirTrackSynth is notable for its extensive variety of objects and complex environments, setting a new standard in the field. This dataset is characterized by its multi-modal sensor data, accurate ground truth labels and a variety of environmental conditions, including distinct weather patterns, lighting conditions, and challenging viewpoints, thereby offering a rich platform to train robust object tracking models. Through the evaluation of the SiamFC object tracking algorithm on AirTrackSynth, we demonstrate the dataset's ability to present substantial challenges to existing methodologies and notably highlight the importance of synthetic data, especially when the availability of real data is limited. This enhancement in algorithmic performance under diverse and complex conditions underscores the critical role of synthetic data in developing advanced tracking technologies.

1 INTRODUCTION

Object tracking in computer vision, crucial for applications such as traffic monitoring, medical imaging, and autonomous vehicle tracking, is particularly significant in the context of unmanned aerial vehicles (UAVs). This task involves identifying and predicting the movements and characteristics of objects within video sequences, presenting unique challenges in real-world scenarios (Du et al., 2018). The pursuit of real-world data is complicated by privacy concerns, copyright infringement issues, and limitations of relying solely on RGB data, which can be affected by environmental factors like lighting and object transformations (Bhatt et al., 2021). Additionally, datasets need labeling, a process that is time-consuming and prone to errors. To overcome these obstacles and enhance deep-learning-based object-tracking models, we propose the construction of a synthetic dataset using a novel combination of the Unreal Engine 4 (UE4), Unreal Engine 5 (UE5), and the AirSim plugin (Shah et al., 2018).

The trajectory of object tracking algorithms has evolved significantly from traditional methods to sophisticated deep-learning approaches. Early techniques like Mean-shift (Zhou et al., 2009; Hu et al., 2008) and Kalman Filter (Weng et al., 2006; Patel and Thakore, 2013) methods laid the groundwork but encountered limitations with complex dynamics and occlusions (Yilmaz et al., 2006). The integration of deep learning transformed object tracking, with Convolutional Neural Networks (CNNs) substantially increasing accuracy and resilience (Nam and Han, 2016). Siamese networks have particularly excelled, with SiamFC (Bertinetto et al., 2016) pioneering a robust offline trained similarity metric. Subsequent innovations like SiamRPN (Li et al., 2018), SiamRPN++ (Li et al., 2019), and SiamMask (Wang et al., 2019b) enhanced adaptability to scale changes and segmentation capabilities. Recent transformer-based models like STARK (Yan et al., 2021) and TransT (Chen et al., 2021) represent cutting-edge tracking technology, leveraging advanced feature extraction and con-

[a] https://orcid.org/0000-0002-8530-6257
[b] https://orcid.org/0000-0002-3667-7406
[c] https://orcid.org/0000-0002-7477-8961

Zhang, M., Van Beeck, K. and Goedemé, T.
Leveraging Unreal Engine for UAV Object Tracking: The AirTrackSynth Synthetic Dataset.
DOI: 10.5220/0013319400003912

In *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2025) - Volume 2: VISAPP*, pages 743-750
ISBN: 978-989-758-728-3; ISSN: 2184-4321

text comprehension.

The development of object tracking algorithms has been supported by robust datasets. Notable examples include OTB2015 (Wu et al., 2015), VOT2018 (Kristan et al., 2018), GOT-10k (Huang et al., 2021), DTB70 (Li and Yeung, 2017), NfS (Danelljan et al., 2017), and LaSOT (Fan et al., 2019). Gaming engines have emerged as powerful tools for synthetic data generation, addressing real-world data acquisition challenges. The AM3D-Sim dataset (Hu et al., 2023) introduced dual-view detection for aerial monocular object detection, while UnrealCV (Qiu et al., 2017) and UnrealRox (Martinez-Gonzalez et al., 2020) developed frameworks for computer vision research and robotics simulations. Synthetic data from games like GTA5 (Wang et al., 2019a; Wang et al., 2021; Fabbri et al., 2021) has proven valuable in replacing real-world datasets while avoiding privacy issues.

While previous datasets cater to broader applications, our AirTrackSynth dataset focuses specifically on UAV object tracking challenges. Our proposed dataset, with 300GB of data across 200 video sequences, is notable for its extensive variety of objects and complex environments. It features multi-modal sensor data, accurate ground truth labels, and diverse environmental conditions, including distinct weather patterns, lighting conditions, and challenging viewpoints. Through evaluation using the SiamFC algorithm, we demonstrate the dataset's ability to present substantial challenges to existing methodologies and highlight the importance of synthetic data when real data availability is limited.

In this work, we address the necessity of such a dataset for UAV object tracking, where maintaining object view under challenging conditions is crucial. This paper details the methods for generating the AirTrackSynth dataset, compares its characteristics with existing datasets, and presents results of training and evaluating object trackers, demonstrating the efficacy of synthetic datasets in advancing real-world object tracking algorithms.

2 SYNTHETIC DATA GENERATION AND METHODOLOGY

Our research utilizes an integrated toolchain centered on Unreal Engine and AirSim for generating synthetic data, coupled with a sophisticated methodology for ensuring data diversity and quality. This section details our technical approach to data generation and the strategies employed for creating realistic tracking scenarios.

2.1 Core Tools and Implementation

The foundation of our data generation pipeline combines Unreal Engine's advanced rendering capabilities with AirSim's drone simulation features. We utilize both UE4 and UE5, leveraging UE5's Lumen global illumination and Nanite geometry system for enhanced realism. Our environments incorporate assets from the Epic Marketplace, including CityPark, CitySample, and DowntownWest, while Mixamo provides character animations for dynamic scenes.

To ensure compatibility between UE5 and AirSim, we modified AirSim's source code to address conflicts with UE5's dynamic characters and lighting effects. This integration enables us to capture complex aerial scenarios while maintaining visual fidelity and physical accuracy.

2.2 Flight Control Strategy

Our implementation focuses on precise UAV control through velocity and acceleration adjustments, employing three key components:

- **Discrete LQR Control:** Implements precise flight adjustments using:

$$u(t) = -K(x(t) - x_{ref}) \tag{1}$$

  where $u(t)$ represents control input, $x(t)$ current state, x_{ref} reference state, and K the gain matrix.

- **Acceleration Control:** Manages horizontal acceleration through attitude angles:

$$\Theta_h = -g^{-1} A_{\psi}^{-1} a \tag{2}$$

  with Θ_h as desired attitude angles and a as desired acceleration.

- **Dynamic Camera Adjustment:** Maintains object centering using:

$$\theta_{cam} = \tan^{-1}\left(\frac{y_{obj} - y_{cam}}{x_{obj} - x_{cam}}\right) \tag{3}$$

2.3 Data Diversity Enhancement

We employ two primary strategies to ensure dataset diversity:

UAV Manipulation:

- Multiple camera positions (top, bottom, left, right, front, rear)
- Varied flight altitudes and distances

- Complex motion patterns including circling, ascending, and linear movements

Environment Manipulation:

- Dynamic lighting conditions through time-of-day adjustments
- Six distinct weather conditions (clear, rain, snow, sandstorm, autumn, fog)
- Diverse object types including humans, vehicles, and animals

This integrated approach enables the generation of rich, varied datasets that closely mirror real-world scenarios while maintaining control over environmental conditions and tracking parameters.

3 DATASET CHARACTERISTICS

Our AirTrackSynth dataset comprises 300GB of data across 200 video sequences, enriched with ground truth labels. Featuring a broad spectrum of data modalities, environments, objects, and scenarios, this dataset aims to create a new benchmark for object tracking research.

3.1 Data Multimodality

Beyond traditional RGB footage, AirTrackSynth extends into depth maps, infrared maps, segmentation maps, IMU values, and UAV statuses. This multimodal data approach is critical for developing sophisticated algorithms capable of navigating the complexities of real-life environments. Figure 1 exemplifies the multimodal data presented in our dataset.

3.2 Challenges in Object Tracking

Our AirTrackSynth dataset simulates a variety of complex scenarios to challenge the state-of-the-art in UAV-based object tracking. Drawing upon the diverse UAV manipulations and virtual environment adjustments outlined, it provides a rich testing ground that closely mirrors the unpredictability and dynamism inherent to real-world tracking tasks.

Firstly, the dataset introduces intricately designed challenges, testing algorithms against complex UAV flight patterns that emulate real operational conditions. These include varying altitudes, angles and motion dynamics that necessitate advanced adaptability and precision in tracking algorithms. Such UAV manipulation strategies ensure that algorithms can maintain robust performance despite the unpredictable movements of both UAVs and their targets, thereby pushing the envelope of current tracking capabilities.

Secondly, our AirTrackSynth offers an immersive simulation environment for tracking algorithms, presenting a wide array of real-world challenges accurately represented within virtual contexts. These challenges include different weather conditions, drastic appearance changes of the tracked object, partial and complete occlusion, presence of distractors and illumination changes.

Illustrated in Figure 2, the dataset showcases a variety of weather conditions—ranging from dusty and foggy atmospheres to autumn scenes with falling leaves, rainy environments with puddles, snowy landscapes, and bright sunny days. These weather scenarios are designed to test the resilience of tracking algorithms under diverse atmospheric conditions, each affecting visibility and object appearance in unique ways.

Further complicating the tracking task, Figure 3a and Figure 3b depict a scenario where the appearance of the object changes dramatically between successive frames. Figure 3c, Figure 3d, Figure 3e and Figure 3f highlight the case of partial or complete occlusion, where a man is obscured by elements within the environment, challenging algorithms to maintain track of the subject despite significant visual obstructions. Additionally, in Figure 3g, Figure 3h and Figure 3i, the presence of distractors alongside drastic illumination changes across frames introduces a scenario where false localizations of the tracked object are highly probable, underscoring the importance of developing algorithms capable of distinguishing the target from misleading cues in the environment.

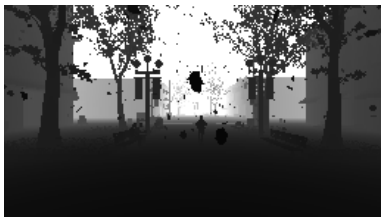
By presenting these multifaceted challenges, the AirTrackSynth dataset serves as a crucial tool for advancing object tracking research. It not only benchmarks the resilience of existing technologies but also inspires the development of innovative solutions capable of overcoming the complexities of tracking in dynamic, real-world environments. The inclusion of detailed environmental manipulations and UAV flight dynamics ensures that AirTrackSynth reflects a wide range of scenarios that algorithms must be prepared to handle.

4 EVALUATION OF THE DATASET

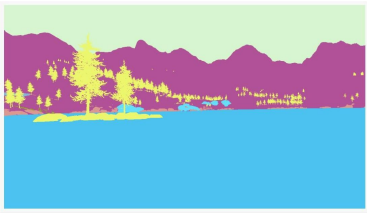
In this section we present a detailed evaluation of our AirTrackSynth dataset, using the SiamFC algorithm (Bertinetto et al., 2016). Our analysis spans across multiple benchmarks, each presenting unique

(a) RGB

(b) Depth



(c) Normalized Disparity



(d) Segmentation

```
382,370,614,529
726,176,715,723
633,151,670,748
653,130,656,769
701,119,647,780
718,116,625,783
728,121,597,778
```

(e) Object Bounding Box

VehicleName	TimeStamp	POS_X	POS_Y	POS_Z	Q_W	Q_X	Q_Y	Q_Z	vx	vy	vz
SimpleFlight	1672622292501	0	0	0.0989935	0.984808	-0	0	0.173648	0	0	0.238048
SimpleFlight	1672622292933	0	0	0.198748	0.984808	-0	0	0.173648	0	0	0.224812
SimpleFlight	1672622293584	0	0	0.341049	0.984808	-0	0	0.173648	0	0	0.213671
SimpleFlight	1672622294226	0	0	0.476294	0.984808	-0	0	0.173648	0	0	0.20827
SimpleFlight	1672622294949	0	0	0.625695	0.984808	-0	0	0.173648	0	0	0.20538
SimpleFlight	1672622295687	0	0	0.776714	0.984808	-0	0	0.173648	0	0	0.204055
SimpleFlight	1672622296254	0	0	0.754747	0.984808	0	0	0.173648	0	0	0 -0
SimpleFlight	1672622296992	0	0	0.754747	0.984808	0	0	0.173648	0	0	0 0.(

(f) Sensor Data

Figure 1: Example of Multimodal Data from Our Dataset.

challenges to object tracking algorithms, to ascertain the dataset's efficacy in enhancing tracking performance, especially under constraints of limited real-world data availability.

4.1 Object Tracker Used for Evaluation

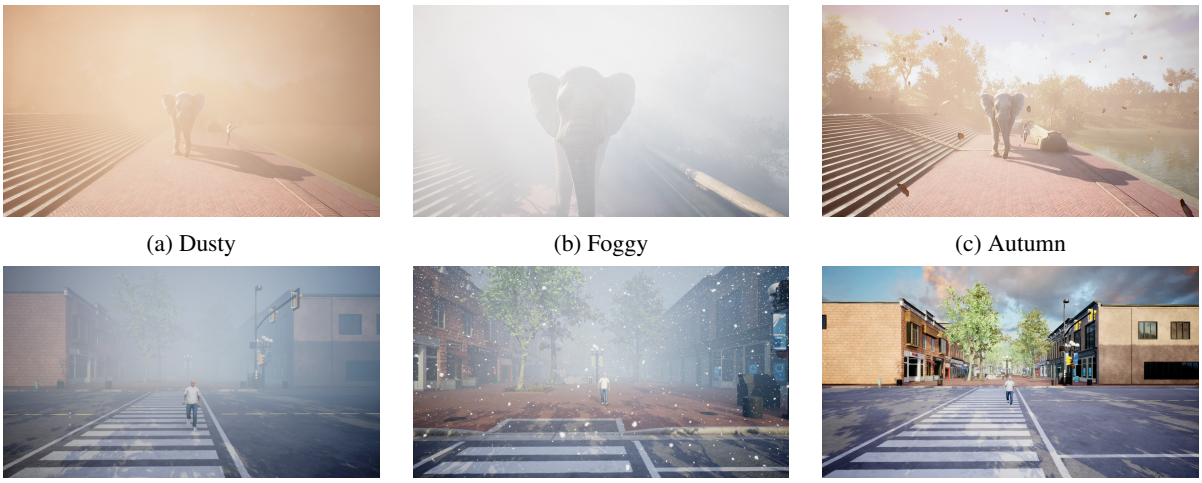
To evaluate the effectiveness of the AirTrackSynth dataset comprehensively, we selected the SiamFC algorithm, a pioneering object tracking model known for its innovative use of Siamese networks. This choice was motivated by SiamFC's status as a seminal work in the application of Siamese networks to visual object tracking, making it an excellent representative for Siamese-based tracking methods. Despite being a relatively early model, SiamFC demonstrates robust performance in visual tracking tasks compared to many conventional methods, establishing it as a relevant benchmark in the object tracking domain.

A key advantage of SiamFC lies in its real-time capability, which is crucial for applications such as autonomous UAV drones where quick processing is essential. The model's architecture is designed for efficiency, allowing it to operate in real-time scenarios. This efficiency is further complemented by the simplicity of SiamFC's design, making it easier to train and implement, especially on embedded hardware. Such simplicity contrasts with more complex transformer-based models, which, while potentially more accurate, may be less suitable for resource-constrained environments.

4.2 Benchmark Datasets and Metrics

We evaluated our dataset using several established benchmarks: DTB70 (Li and Yeung, 2017) (70 sequences focusing on small object tracking), OTB2015 (Wu et al., 2015) (100 sequences with varied scenarios), VOT2018 (Kristan et al., 2018) (challenging tracking sequences), NfS (240 FPS) (Danelljan et al., 2017) (high-speed tracking), LaSOT (Fan and Ling, 2019) (1400 videos across 70 categories), and GOT-10k (Huang et al., 2019) (over 10,000 video clips).

For evaluation metrics, we used Success Score and Precision Score for OTB2015, DTB70, LaSOT, and NfS datasets, measuring overlap rate and center point accuracy respectively. VOT2018 was evaluated using Accuracy (spatial precision) and Robustness (failure rate) metrics. For GOT-10k, we employed Average Overlap (AO) and Success Rates at thresholds 0.5 and 0.75 (SR0.5, SR0.75). The metrics used for each dataset are commonly used in the literature.



(a) Dusty (b) Foggy (c) Autumn

(d) Rainy (e) Snowy (f) Sunny

Figure 2: Examples of frames with different weather conditions introduced in the AirTrackSynth dataset.

4.3 Experiment Setup

We evaluated the impact of combining synthetic and real data using the GOT-10k dataset as baseline. Our experiments used training sets with varying ratios of real-to-synthetic data, progressively increasing real video sequences from 0 to 500 while maintaining consistent synthetic data. A control group using only real data was maintained for comparison.

The experiments were conducted using an NVIDIA RTX 3060 Ti GPU with Intel i7-12700K CPU and 64GB RAM. The SiamFC model was trained with an initial learning rate of 0.01, batch size of 32, and 50 epochs, following the original implementation's optimization parameters.

4.4 Results and Analysis

This section presents and analyzes the test results of the SiamFC model trained with various combinations of synthetic and real data across multiple datasets. We evaluate the impact of combining synthetic data with real data on tracking performance, comparing it to training with real data alone.

Table 1 presents a comprehensive overview of the SiamFC model's performance across six diverse datasets, comparing various combinations of real and synthetic training data. The results demonstrate the impact of synthetic data on the model's tracking capabilities across different scenarios and metrics.

For GOT-10k, the integration of synthetic data significantly improves the model's performance, especially when real data is limited. The results show large performance gains when combining synthetic data with small amounts of real data, demonstrating the substantial benefit of synthetic data in training robust models.

LaSOT, known for its challenging large-scale and long-term tracking conditions, shows the effectiveness of synthetic data augmentation. The mixed training approach (real + synthetic) outperforms the real-data-only setups across almost all metrics, indicating improved performance in complex tracking scenarios.

The NfS dataset, characterized by high-speed object tracking, reveals significant improvements with the inclusion of synthetic data. This is particularly evident in setups where the amount of real data is limited, underscoring synthetic data's utility in preparing the model for challenging high-speed tracking scenarios.

On the DTB70 dataset, we observe a clear improvement when combining synthetic data with real data. All Real + Synthetic data configurations outperform their real-data-only counterparts, highlighting the synthetic data's role in enhancing the model's generalization capabilities.

For OTB2015, training with synthetic data yields significant improvements in both precision and success rates. The enhancements are more pronounced when using only 1 or 8 real videos, highlighting synthetic data's critical role in boosting tracking accuracy and success, especially when real data is scarce.

In the context of VOT2018, a dataset renowned for its demanding tracking tasks, the addition of synthetic data alongside real data significantly enhances the robustness of the SiamFC tracker. This improvement in robustness is crucial for applications like autonomous driving and surveillance, where the ability to handle unpredictable elements is essential.

Overall, this comprehensive analysis across multiple datasets reveals a consistent trend: incorporating

(a) Appearance change ($T = t$) (b) Appearance change ($T = t+1$) (c) Occlusion ($T = t$)

(d) Occlusion ($T = t+1$) (e) Occlusion ($T = t+2$) (f) Occlusion ($T = t+3$)

(g) Distractor and illumination change ($T = t$) (h) Distractor and illumination change ($T = t+1$) (i) Distractor and illumination change ($T = t+2$)

Figure 3: Challenging scenarios in object tracking within the AirTrackSynth dataset.

synthetic data with real data significantly enhances the performance of the SiamFC model, particularly in settings where real data is limited. The improvement is evident across various tracking challenges, from high-speed scenarios to long-term, large-scale tracking. The enhancement in performance metrics such as accuracy, precision, success rate, and robustness underscores the value of synthetic data in providing a diverse range of scenarios that real data alone might not capture.

The consistent improvement in robustness observed in the VOT2018 dataset is particularly notable, demonstrating the synthetic data's role in preparing the model for complex tracking environments. This finding is crucial for applications where robustness is critical, such as autonomous driving and surveillance, where unpredictable elements are present.

5 CONCLUSIONS

In this work, we introduced a novel synthetic dataset created using the Unreal Engine and the AirSim simulator, designed to address the complex needs of the object tracking task in computer vision. Our dataset stands out by offering multi-modal data, encompassing a variety of weather and lighting conditions, and specifically addressing challenging scenarios that are critical for advancing object tracking algorithms.

The synthetic dataset's diversity and richness in scenarios—from varying weather conditions to intricate lighting dynamics—provide a comprehensive testing and training ground for developing robust object tracking algorithms. Moreover, the inclusion of hard scenarios, such as rapid object motion, occlusions, and illumination changes, ensures that models trained on this dataset are well-equipped to handle real-world complexities.

Through experimental validation, we have demonstrated the significant value of integrating synthetic data with real-world data, particularly in contexts where real data is scarce or limited in diversity. Our results, obtained across several benchmarks, including DTB70, GOT-10k, LaSOT, NfS, OTB2015 and VOT2018, clearly show that models trained on a combination of real and synthetic data exhibit superior performance in terms of accuracy, precision, success rates and robustness compared to models trained exclusively on real data.

The findings from our study underscore the synthetic data's crucial role in enhancing the generalization capabilities of object tracking models. This work

Table 1: Comprehensive Results of SiamFC Model Across Multiple Datasets.

Dataset	GOT-10k	LaSOT	NfS	DTB70	OTB2015	VOT2018
	AO / SR0.50 / SR0.75	Success / Precision	Success / Precision	Success / Accuracy	Success / Precision	Accuracy / Robustness
Real 1	0.167 / 0.127 / 0.023	0.1051 / 0.0646	0.107 / 0.120	0.143 / 0.213	0.131 / 0.140	0.3249 / 311.4349
Real 8	0.445 / 0.472 / 0.240	0.2484 / 0.2270	0.356 / 0.417	0.326 / 0.480	0.437 / 0.559	0.4560 / 85.3599
Real 50	0.455 / 0.483 / 0.227	0.2449 / 0.2164	0.371 / 0.455	0.343 / 0.441	0.447 / 0.589	0.4546 / 70.0071
Real 500	0.458 / 0.510 / 0.239	0.2518 / 0.2337	0.413 / 0.489	0.383 / 0.585	0.468 / 0.626	0.4499 / 71.0992
Real 1 + Full Synthetic	0.426 / 0.449 / 0.199	0.2090 / 0.1930	0.288 / 0.359	0.347 / 0.536	0.399 / 0.535	0.4593 / 85.1166
Real 8 + Full Synthetic	0.455 / 0.492 / 0.230	0.2649 / 0.2444	0.395 / 0.473	0.393 / 0.594	0.452 / 0.600	0.4707 / 75.2022
Real 50 + Full Synthetic	0.428 / 0.453 / 0.195	0.2395 / 0.2172	0.345 / 0.425	0.366 / 0.548	0.451 / 0.603	0.4557 / 72.0395
Real 500 + Full Synthetic	0.489 / 0.544 / 0.239	0.2555 / 0.2406	0.399 / 0.479	0.403 / 0.637	0.469 / 0.648	0.4479 / 65.0460

not only validates the effectiveness of our synthetic dataset but also highlights the potential of synthetic data to complement and augment real data, pushing the boundaries of what is achievable in object tracking research.

Building on the solid foundation laid by this research, future work will pivot towards harnessing the full potential of multimodal data present in our synthetic dataset.

The primary focus will be on developing and fine-tuning models capable of effectively fusing multimodal data to achieve a more comprehensive understanding of the tracking environments. Furthermore, to thoroughly validate the versatility and robustness of our synthetic dataset, it is required to test its efficacy across a broader spectrum of tracking algorithms. By expanding the array of tested tracking models, including those leveraging advanced neural architectures, we aim to establish our synthetic dataset as a benchmark for future developments in object tracking.

ACKNOWLEDGEMENTS

This project has been partially funded my the VLAIO Tetra Project *AI To The Source* and the Flanders AI Research Program.

REFERENCES

Bertinetto, L., Valmadre, J., Henriques, J. F., Vedaldi, A., and Torr, P. H. (2016). Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*, pages 850–865. Springer.

Bhatt, D., Patel, C., Talsania, H. N., Patel, J., Vaghela, R., Pandya, S., Modi, K. J., and Ghayvat, H. (2021). Cnn variants for computer vision: History, architecture, application, challenges and future scope. *Electronics*, 10(20).

Chen, X., Yan, B., Zhu, J., Wang, D., Yang, X., and Lu, H. (2021). Transformer tracking. In *Proceedings of the*

IEEE/CVF conference on computer vision and pattern recognition, pages 8126–8135.

Danelljan, M., Bhat, G., Shahbaz Khan, F., and Felsberg, M. (2017). The need for speed: A benchmark for visual object tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1125–1134.

Du, D., Qi, Y., Yu, H., Yang, Y.-F., Duan, K., Li, G., Zhang, W., Huang, Q., and Tian, Q. (2018). The unmanned aerial vehicle benchmark: Object detection and tracking. In *European Conference on Computer Vision (ECCV)*.

Fabbri, M., Brasó, G., Maugeri, G., Cetintas, O., Gasparini, R., Ošep, A., Calderara, S., Leal-Taixé, L., and Cucchiara, R. (2021). Motsynth: How can synthetic data help pedestrian detection and tracking? In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10829–10839.

Fan, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Bai, H., Xu, Y., Liao, C., and Ling, H. (2019). Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Fan, H. and Ling, H. (2019). Lasot: A high-quality benchmark for large-scale single object tracking. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5374–5383.

Hu, J.-S., Juan, C.-W., and Wang, J.-J. (2008). A spatial-color mean-shift object tracking algorithm with scale and orientation estimation. *Pattern Recognition Letters*, 29(16):2165–2173.

Hu, Y., Fang, S., Xie, W., and Chen, S. (2023). Aerial monocular 3d object detection. *IEEE Robotics and Automation Letters*, 8(4):1959–1966.

Huang, L., Zhao, X., and Huang, K. (2019). Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *arXiv preprint arXiv:1810.11981*.

Huang, L., Zhao, X., and Huang, K. (2021). GOT-10k: a large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1562–1577.

Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pflugfelder, R., Cehovin Zajc, L., Vojir, T., Hager, G., Lukezic, A., Eldesokey, A., et al. (2018). The sixth visual object tracking vot2018 challenge results. In *European Conference on Computer Vision*, pages 0–0.

Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., and Yan, J. (2019). SiamRPN++: Evolution of Siamese visual tracking with very deep networks. In *the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4282–4291.

Li, B., Yan, J., Wu, W., Zhu, Z., and Hu, X. (2018). High performance visual tracking with Siamese region proposal network. In *the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8971–8980.

Li, S. and Yeung, D.-Y. (2017). Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Martinez-Gonzalez, P., Oprea, S., Garcia-Garcia, A., Jover-Alvarez, A., Orts-Escolano, S., and Garcia-Rodriguez, J. (2020). Unrealrox: an extremely photorealistic virtual reality environment for robotics simulations and synthetic data generation. *Virtual Reality*, 24:271–288.

Nam, H. and Han, B. (2016). Learning multi-domain convolutional neural networks for visual tracking. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4293–4302.

Patel, H. A. and Thakore, D. G. (2013). Moving object tracking using kalman filter. *International Journal of Computer Science and Mobile Computing*, 2(4):326–332.

Qiu, W., Zhong, F., Zhang, Y., Qiao, S., Xiao, Z., Kim, T. S., and Wang, Y. (2017). Unrealcv: Virtual worlds for computer vision. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, page 1221–1224, New York, NY, USA. Association for Computing Machinery.

Shah, S., Dey, D., Lovett, C., and Kapoor, A. (2018). Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In Hutter, M. and Siegwart, R., editors, *Field and Service Robotics*, pages 621–635, Cham. Springer International Publishing.

Wang, Q., Gao, J., Lin, W., and Yuan, Y. (2019a). Learning from synthetic data for crowd counting in the wild. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8190–8199.

Wang, Q., Gao, J., Lin, W., and Yuan, Y. (2021). Pixel-wise crowd understanding via synthetic data. *International Journal of Computer Vision*, 129(1):225–245.

Wang, Q., Zhang, L., Bertinetto, L., Hu, W., and Torr, P. H. (2019b). Fast online object tracking and segmentation: a unifying approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1328–1338.

Weng, S.-K., Kuo, C.-M., and Tu, S.-K. (2006). Video object tracking using adaptive kalman filter. *Journal of Visual Communication and Image Representation*, 17(6):1190–1208.

Wu, Y., Lim, J., and Yang, M.-H. (2015). Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1834–1848.

Yan, B., Peng, H., Fu, J., Wang, D., and Lu, H. (2021). Learning spatio-temporal Transformer for visual tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10448–10457.

Yilmaz, A., Javed, O., and Shah, M. (2006). Object tracking: A survey. *ACM computing surveys (CSUR)*, 38(4):13.

Zhou, H., Yuan, Y., and Shi, C. (2009). Object tracking using sift features and mean shift. *Computer vision and image understanding*, 113(3):345–352.