Tracking Small Objects in Global Motion Conditions for an Ornithological Monitoring System

Natalia Obukhova, Alexandr Motyko,
Alexandr Pozdeev, Alexander Savelev,
Pavel Baranov, Konstantin Smirnov
Dmitry Sharivzyanov
St. Petersburg Electrotechnical University "LETI"
St. Petersburg, Russia
naobukhova@etu.ru, aamotyko@etu.ru,
puches4@gmail.com, algsavelev@gmail.com
psbaranov@etu.ru, konstantinandsmi@ya.ru,
sonderyx@ya.ru

Aleksei Samarin
ITMO University
St. Petersburg, Russia
avsamarin@itmo.ru

Egor Kotenko
St. Petersburg State University
St. Petersburg, Russia
kotenkoed@gmail.com

***Abstract*—In ornithological video monitoring tasks aimed at studying bird behavior, estimating population size, and tracking migration routes, a major challenge is the robust tracking of targets under global camera motion. Such motion, often caused by pan-tilt or mobile platforms, introduces significant distortions in the optical flow. At the same time, the tracked objects are typically small, low-contrast, and highly dynamic, which considerably reduces the robustness of conventional tracking methods. This study aims to develop and experimentally validate a tracking method that can operate in video sequences affected by global motion, while maintaining high accuracy and real-time performance. The proposed approach integrates a neural network-based tracker and trajectory prediction using a Kalman filter. The method was evaluated on a dataset simulating real ornithological monitoring scenarios, including highly detailed and dynamic backgrounds, moving cameras, variable lighting, and complex object trajectories. Experimental results showed that the tracking failure rate did not exceed 5×10^{-4}, while the average processing speed reached 21 frames per second. Compared to a conventional tracking method based on HOG+KCF and Kalman filtering, the proposed method achieved a 4-fold reduction in tracking failure rate and a 2.5-fold reduction in tracking failures under occlusion conditions. The developed method is designed for use in bird monitoring systems operating in natural and agricultural landscapes, where reliable object tracking is required in visually complex environments. The results demonstrate the potential of the proposed solution for both scientific and ornithological research, as well as applied environmental monitoring tasks.**

I. Introduction

The task of automatic bird tracking in video sequences is one of the important areas in computer vision, lying at the intersection of applied ecology, biomonitoring, and environmental observation. Modern video surveillance systems increasingly operate in real-world conditions that differ significantly from laboratory settings, characterized by high scene dynamics, changing backgrounds, camera motion, and numerous low-contrast objects of interest. In the context of ornithological applications, this creates a need for robust algorithms capable of reliably tracking birds in visually complex environments, including forested, coastal, and agricultural landscapes.

The issue of stable tracking of objects of interest in video sequences is becoming increasingly relevant across a range of applications. At the same time, the development of current-generation video monitoring systems faces several critical challenges, primarily related to the following factors:

- highly detailed and dynamic backgrounds;
- significant variation in object properties, in particular, rapid changes in the projection size of the object on the image plane;
- occurrence of occlusion events due to overlapping with the background or other objects;
- the need for high-speed video processing (close to real time) under limited computational resources.

Several studies [1]–[3] have proposed integrated systems that combine infrared cameras and radar for continuous monitoring, as well as modified YOLOv8-based architectures adapted for bird detection near protected areas. The system described in this study implements an ornithological monitoring scenario that is close in purpose to the approaches discussed in [1]–[3]. Its fundamental difference, however, lies in its ability to perform tracking under camera motion — a factor not addressed in the aforementioned publications.

This paper presents a method for automatic tracking of target objects in an optoelectronic ornithological monitoring system under global motion, heterogeneous and textured backgrounds, occlusions, and low-contrast targets. A key feature of the problem setting is the presence of multiple objects of interest that are small in size and exhibit significantly different trajectories and speeds.

The monitoring system consists of a pan-tilt camera and a computing unit equipped with an NVIDIA RTX 3090 GPU, enabling real-time neural network execution. The optical layout of the system is designed so that the projection of a medium-sized bird (e.g., a pigeon or crow), approximately 35 centimeters in length, occupies about 20 pixels on the

image when observed from a distance of 1500 meters. Such distances correspond to typical conditions for ornithological monitoring in field and security contexts. According to the specifications of modern radar and optical systems, reliable detection and tracking of medium-sized birds is achievable at distances ranging from 1000 to 2000 meters [4], [5]. Compact radars and combined sensor modules integrated with optical cameras are capable of reliably detecting and tracking birds even under poor visibility conditions [6], which confirms the applicability of this distance range to practical ornithological monitoring tasks. Furthermore, both larger birds at greater distances and smaller birds at closer ranges may be of interest for monitoring.

In practical scenarios, the projection size of an object may vary from approximately 15×15 pixels (at the capturing stage) to around 200×200 pixels (at close range), given a frame resolution of 1920×1080. Therefore, the tracking method must account not only for background variability and camera motion, but also for substantial changes in the visual properties of the tracked objects. An additional challenge is achieving high-speed video processing under constrained resources, which imposes strict requirements on the efficiency and robustness of the tracking algorithm. The main feature of the described tracking task is the very complex trajectory of the object, because it is of natural, not artificial origin. This leads to the fact that classical algorithms and approaches used for such tasks encounter significant difficulties, and their accuracy characteristics are reduced.

II. Related Work

The task of object tracking in video streams is complicated by a variety of factors, including challenging environmental and motion-related conditions. Furthermore, tracking must be synchronized with the algorithms controlling the camera's position. These constraints place specific requirements on the choice and architecture of the tracking method.

Today, deep learning methods and neural networks are widely used in computer vision for a broad range of tasks. With the advent of high-performance computing resources, it has become feasible to employ deep neural networks for near-real-time object tracking. A promising direction in object tracking involves segmenting the object in each video frame, even when only a bounding box overlay is required. Generating a segmentation mask of the object of interest enables more precise estimation of its boundaries and position, significantly improving tracking robustness under complex backgrounds and partial occlusions [7].

Recent works on tracking small objects have proposed methods that incorporate both spatial and temporal features, such as ST-Motion TinyDet [8]. Among the single-frame algorithms, DN-FPN+Trans-R-CNN, DCFL-TinyDet [9], TAD (Tiny Airborne Detection) [10] and hybrid architectures combining transformers with YOLO (STF-YOLO) [11] are notable.

Despite the progress of transformer-based, diffusion-based, spatiotemporal, and hybrid trackers, their practical application in real-time video analytics under limited computing resources remains a challenge. Transformers and hybrid models with heavy transformer components demand substantial computational power during both training and inference. Diffusion models offer flexibility but still fall short in processing speed for real-time streaming data.

When computing resources are limited, classical (non-neural) tracking methods are often applied, trading accuracy for speed. Among them, discriminative correlation filter (DCF)-based trackers such as KCF [12], CSRT [13], and DLT [14] are particularly popular due to their high speed and low hardware requirements. Several improved KCF variants demonstrating competitive performance on benchmark datasets while maintaining speed are discussed in [15]–[17]. The authors of [18] propose a lightweight neural modification of DCF (DCFNet), also suitable for real-time applications. A comprehensive review of classical trackers and their deployment under resource constraints is presented in [19].

To ensure stable and robust object tracking in various scenarios, it is often necessary to combine multiple approaches. Each works best under specific conditions, and the system must include logic to switch between them. For example, [20] describes a combination of a HOG-based detector [21] and a correlation tracker along with a Kalman filter. The drawback of such methods lies in the need to fine-tune thresholds and frequent reinitializations, which reduce tracking efficiency under complex conditions.

In general, the choice of tracking method depends on both available hardware resources and task-specific requirements. For monitoring high-contrast objects (e.g., self-illuminated or with simple trajectories), classical low-cost approaches may suffice. In contrast, neural network-based methods may be preferable in more challenging cases.

This study proposes a tracking method for an ornithological monitoring system based on neural segmentation tracking and Kalman filtering. As a baseline for comparison, a classical algorithm using HOG descriptors, a KCF tracker, and a Kalman filter is used, as described in [20]. An important feature of the study is the specificity of the task - tracking natural objects (birds) with complex, poorly predictable trajectories. This contrasts with the traditional formulation of the problem - tracking artificial objects, which is the focus of most classical approaches.

III. Proposed Solution

The proposed tracking method employs the Segment Anything Model 2 (SAM2) [22] as the core segmentation module, combines it with a third-order Kalman filter in the image plane for motion estimation and trajectory prediction, leverages a neural network-based detector to locate target objects, and integrates a control module for a pan-tilt camera that uses the predicted motion to maintain continuous alignment with the target.

The third-order Kalman model is particularly suitable for natural targets (e.g., birds) exhibiting irregular trajectories and abrupt speed changes, where a simpler second-order model

would be insufficiently responsive. In this formulation, the state vector at time step k

$$s_k = [x_k, v_{kx}, a_{kx}, y_k, v_{ky}, a_{ky}]^T$$

encodes the target center coordinates (x, y) along with the projections of its velocity v and acceleration a on the respective axes. Observations are derived from the center of the detectors or the tracker's bounding box, with the discretization step equal to the inter-frame interval of the video stream. Initialization is performed upon the first reliable detection, setting the position to the observation and both velocity and acceleration to zero. The use of acceleration in the kinematic model enables rapid convergence after initial corrections, precise short-term prediction, robustness to brief dropouts, and jitter suppression under global camera motion. SAM2 utilizes the concept of promptable segmentation, enabling the extraction of a mask for any object based on a specific prompt (point or bounding box).

The model demonstrates strong generalization ability, requires no fine-tuning for a specific object class, and performs robustly under various conditions. The operation pipeline is illustrated in Algorithm 1.

The system operates in the following stages:

1) **Initialization.** Primary object detection is performed on the full frame. Our implementation is based on the SSD-ADSAR detector. A target is selected (in case of several objects were detected), captured, and both the Kalman filter and the region of interest (ROI) are initialized.
2) **Prediction.** For each incoming frame, the Kalman filter predicts the new target position. This prediction is used both to constrain the search area and to generate a control command for the pan-tilt camera, compensating for system latency and minimizing the risk of losing the target.
3) **Local Processing.** Within the predicted ROI:
   - SAM2 segmentation is executed, producing a mask and bounding box ($Rect_2$).
   - Local detection is performed using SSD-ADSAR, yielding another bounding box ($Rect_1$).
4) **Confirmation.** Confirmation is implemented based on the IoU (Intersection over Union) metric calculated for the above-formed bounding boxes. If the confirmation is successful. The Kalman filter is updated, and all counters are reset. If only the SAM2 mask is available, a partial update is performed. If neither mask nor detection is present, the Kalman prediction is used, and failure counters are incremented.
5) **Tracking interruption.** If the number of consecutive frames without a mask exceeds N_{loss}, or the number of frames without detector confirmation exceeds N_{conf}, the object is considered lost.

The algorithm's parameters are selected based on the hardware platform used in deployment.

Algorithm 1 Target tracking algorithm

Input: T_{IoU} (IoU threshold), N_{loss} (no-mask limit), N_{conf} (no-confirmation limit)
State: KF (Kalman filter), ROI (region of interest), $loss\_cnt \leftarrow 0$, $conf\_wait\_cnt \leftarrow 0$
Initialization: Run global SSD-ADSAR on full frame to obtain $Rect_0$ (target acquisition).
Initialize KF using $Rect_0$; set $ROI \leftarrow Rect_0$.
while new frame available **do**
  Predict target position using KF; update ROI around prediction.
  Run SAM2 on ROI to get $Mask$ (primary); run SSD-ADSAR on ROI to get $Rect_1$ (may be absent).
  if $Mask$ exists **then**
    Derive $Rect_2$ from $Mask$.
    if $Rect_1$ exists **and** $IoU(Rect_1, Rect_2) > T_{IoU}$ **then**
      Update KF using $Rect_2$; $loss\_cnt \leftarrow 0$; $conf\_wait\_cnt \leftarrow 0$.
    else
      Update KF using $Rect_2$ (possibly reduced-trust update).
      $loss\_cnt \leftarrow 0$; $conf\_wait\_cnt \leftarrow conf\_wait\_cnt + 1$.
      if $conf\_wait\_cnt \geq N_{conf}$ **then**
        break // terminate tracking: no detector confirmation for too long
      end if
    end if
  else
    $loss\_cnt \leftarrow loss\_cnt + 1$.
    if $loss\_cnt \geq N_{loss}$ **then**
      break // terminate tracking: mask missing for too long
    end if
    KF remains in predicted state.
  end if
  Generate and send PTZ control command based on current prediction of KF.
end while

The proposed method integrates the segmentation precision of SAM2 and the robustness of Kalman filtering with the detection capabilities of SSD-ADSAR. This combination enables reliable object tracking even under challenging conditions such as complex backgrounds, complex trajectories of movement, significant changes in speed, camera motion, and significant variations in target size and shape. The system operates near real-time, making it well-suited for practical deployment in automated monitoring and surveillance scenarios involving dynamic and unpredictable environments.

To provide a comparative perspective, the workflow of the competing tracking approach described in [20] is summarized below. Both trackers implement tracking with confirmation from a detector, are designed for similar operating conditions, including global motion and complex backgrounds, and use

Kalman-based motion prediction. The proposed method employs segmentation-based localization using a neural network model (SAM2), whereas the competing approach is based on feature-based tracking with HOG descriptors and KCF, supplemented by correlation analysis. The key distinction is that our method is specifically adapted for tracking natural, living targets with complex and irregular trajectories (e.g., birds), while the competing method has demonstrated strong performance primarily for man-made objects with more regular trajectories and predictable motion patterns.

In the competing approach, after the initial detection using a neural network model, the object in the current frame is identified via an HOG+KCF tracker, supplemented by a correlation-based detector and motion prediction generated by Kalman filtering. This combination improves tracking robustness under scale variation and global scene motion. For each frame, three sources of information (HOG+KCF, the correlation detector, and the Kalman filter) are analyzed, and their outputs are compared using the IoU metric. Depending on the degree of consistency between sources, different modules are re-initialized: when all sources agree, the reference image is updated; when there is disagreement with the Kalman filter, its parameters are re-estimated; and when a significant scale change occurs, only the correlation output is temporarily used. This fallback mode is activated based on a threshold on the correlation response area to balance stability and accuracy. Tracking is terminated if no agreement between sources is observed for N consecutive frames. A more detailed description of the algorithm is provided in [20].

IV. Experiments and Results

The objective of the experimental study is to quantitatively assess the tracking performance and compare it with alternative approaches. Two tracking methods were evaluated: a "classical" one based on the combination of HOG descriptors, KCF tracker, and Kalman filter, as described in [20], hereinafter referred to as HOG+KCF, and the proposed approach based on SAM2 segmentation and Kalman filtering.

To evaluate the tracking quality, a set of Full HD (1920×1080) videos was collected and used. The main metric was the number of tracking failures – instances where the tracker lost the target and required reacquisition via the detector.

Additionally, the processing speed (FPS, frames per second) was measured on the hardware platform used in the monitoring system (equipped with an Nvidia GTX 3090 GPU). The results are summarized in Table I.

As described in the dataset documentation, the tracking conditions are challenging. The bird often undergoes significant scale changes, hovers motionless, or blends into the background (e.g., a white bird against clouds or a black one against buildings and vegetation), crosses regions with varying backgrounds, or sharply changes its motion trajectory, sometimes accompanied by global scene shifts. Such scenarios reflect real-world conditions and significantly complicate tracking. Examples of substantial scale changes and strong background fusion are shown in Fig. 1 and Fig. 2.

Fig. 1. The example of an object with scale variation

From Table I, the failure rate (tracking dropout frequency) can be computed as follows [23]:

$$F = \frac{F_t}{N},$$

where F_t is the number of tracking failures, and N is the total number of frames. The calculation includes videos No.1–7, which are representative of the monitoring system. Video No.8, a synthetic composition of ten 3-second clips containing occlusions (ranging from 0.4 to 1.0 seconds), was evaluated separately.

According to the experiments, the proposed method achieved a failure rate of $5 \cdot 10^{-4}$, whereas the HOG+KCF tracker yielded $2 \cdot 10^{-3}$.

The performance of the HOG+KCF tracker depends on the object's size; hence, the FPS values reported in Table I vary. In general, HOG+KCF demonstrates higher speed compared to the proposed method; however, the achieved frame rate of 21 FPS is sufficient for practical deployment.

The failure rate for HOG+KCF observed in this study differs from that reported in [20]. This discrepancy is attributed to the different target types: the referenced study dealt with synthetic, predictable targets, while our work focused on real birds whose

TABLE I. EXPERIMENTAL TRACKERS COMPARISON RESULTS

Video No.	Description	Frame count	Failures SAM2 + Kalman	FPS SAM2 + Kalman	Failures HOG + KCF	FPS HOG + KCF
1	White bird, complex background (ground, vegetation)	1625	0	21	3	24
2	Black bird, complex background (sky, high noise), global motion	675	0		3	37
3	Black bird, complex background (mountains, vegetation), occlusion	1495	0		2	30
4	White bird, simple background (sky with clouds), size variation	314	0		1	24
5	Black bird (sky with clouds, vegetation, houses), rapid variation	1428	3		7	23
6	Black bird, simple background (twilight, grey sky)	1660	0		1	30
7	White bird, simple background (sky with clouds), size variation	320	1		0	33
8	White and black birds, montage with occlusions, urban background	900	2		5	24

Fig. 2. The examples of a scene with a complex background

motion is less regular. For such dynamic scenarios, Kalman filtering alone proves less effective. Although HOG+KCF shows satisfactory performance, the proposed method demonstrates clear advantages. This includes a 2.5x reduction in tracking failures under occlusion, which can be largely attributed to the SAM2 model's internal image memory (embedding), improving robustness under partial visibility, deformation, and complex motion patterns.

An interesting point of discussion is the ability of the SAM2-based tracker to handle extremely small objects, located generally outside the target characteristics of the system and corresponding, for example, to objects at very long distances (relative to those typical for monitoring systems). Figure 3 illustrates frame sequences and masks produced by SAM2 during tracking of an object with a minimal projected size of approximately 5×5 pixels.

As seen in Figure 3, when the projection is that small, both the image and the segmentation mask are blurry and poorly aligned in shape. In such conditions, especially with a highly detailed background, reliable tracking is unrealistic. Nonetheless, in some cases common to ornithological monitoring (e.g., high-contrast targets on low-textured backgrounds), successful tracking can still be achieved. It should be noted that the object size here is far below the target specification for the system, and such tracking success is an additional advantage.

Experiments have shown that the proposed method performs robustly under both simple and complex conditions. It delivers sufficient processing speed and demonstrates high accuracy in detection and tracking, even under difficult scenarios.

V. DISCUSSION AND CONCLUSION

This paper presents a method for automatic tracking of objects of interest adapted to the specific challenges of ornithological monitoring, such as global motion in video data, high object dynamics, occlusion, cluttered backgrounds, complex trajectories, and the speed mode of a natural object. The proposed approach integrates a segmentation-based tracker built upon the SAM2 model, a hybrid detector powered by neural networks used for initial object detection and trajectory confirmation, and Kalman filter-driven motion estimation.

Utilizing motion estimation via the Kalman filter enabled increased tracking stability by smoothing abrupt changes in the visual scene caused by rapid camera movement. Additionally, it allowed for reduced latency between object position estima-

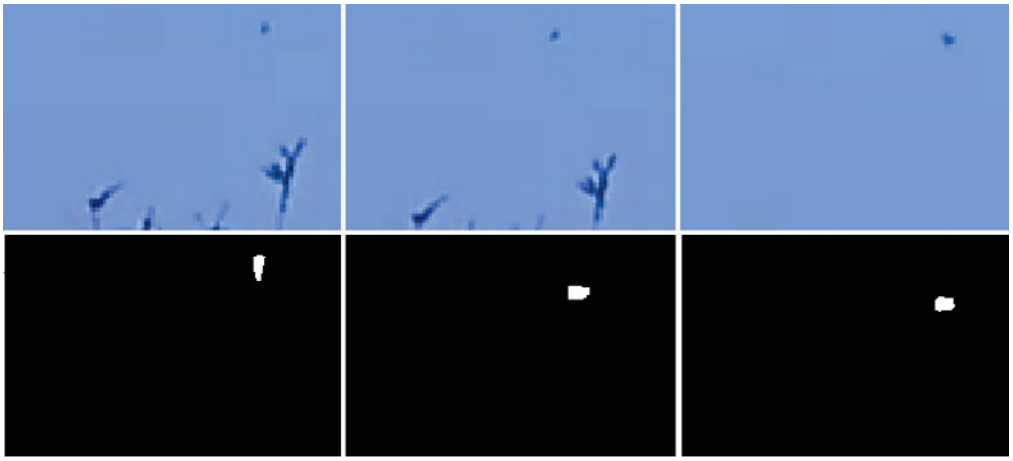

Fig. 3. Zoomed-in video frames of tracking a small object of interest and corresponding masks generated by the SAM2 tracker

tion and pan-tilt actuator response, ensuring reliable operation in mobile camera systems.

Experimental evaluation confirmed the robustness of the proposed method under challenging conditions: the tracking failure rate did not exceed 5×10^{-4}, while achieving a real-time processing rate of 21 FPS on FullHD video. In contrast, the failure rate of a traditional HOG+KCF tracker under similar conditions was found to be approximately four times higher.

The results support the following conclusions:

- the proposed method ensures stable tracking for birds under global motion, changes in object appearance, and partial occlusions;
- it demonstrates high performance, compatible with real-time constraints;
- the system is well-suited for deployment in field scenarios with limited computational resources;
- the observed improvement (a 4-fold reduction in tracking failure rate) over the baseline method is explained by the ability of the proposed approach to handle complex and non-linear object trajectories, which are typical in bird tracking scenarios and where classical methods tend to be less effective.

While these findings highlight the technical strengths of the system, several broader implications and practical considerations should also be noted. From an ecological perspective, reliable long-term bird tracking can provide critical data for biodiversity assessment, habitat monitoring, and conservation planning. Automated systems of this kind could reduce dependence on manual field observations, enabling continuous, large-scale monitoring of species populations and migration dynamics. In agricultural and airport contexts, such monitoring also holds applied value for mitigating bird strikes and managing human–wildlife interactions.

At the same time, practical constraints may limit direct deployment. While the RTX 3090 platform provides sufficient throughput, many real-world scenarios rely on resource-constrained edge devices. Although our method is optimized for efficiency, its performance on mobile GPUs or embedded platforms remains to be tested. Approximate segmentation, lightweight backbones, or model distillation techniques could reduce computational demands, though this may come at some cost to accuracy. Future experiments on Jetson-class devices or FPGA-based accelerators will therefore be essential to evaluate deployment feasibility in remote or autonomous stations.

Another important aspect concerns datasets and evaluation metrics. In this work, we relied on a dataset simulating real ornithological conditions; however, the diversity of species, environments, and weather conditions in the field is far greater. Expanding the dataset to cover more species, flight behaviors, and seasonal variations, as well as including multimodal inputs (thermal, radar-assisted), would strengthen generalization. Moreover, while failure rate and FPS were key evaluation metrics, additional measures such as ID switches, trajectory continuity, and long-term re-identification could provide a more nuanced picture of system performance.

Finally, although our method outperformed the chosen HOG+KCF baseline, further comparisons with state-of-the-art lightweight trackers would improve the robustness of conclusions. Explicit discussion of limitations (for example, reduced accuracy on extremely small or distant targets and reliance on clear optical conditions) helps delineate the boundaries of applicability. These limitations define clear directions for future research, including multimodal sensing, improved resilience under adverse weather, and integration into distributed ecological monitoring networks.

In summary, the proposed method achieves stable, accurate, and real-time bird tracking under global motion and complex backgrounds, offering tangible benefits for ecological monitoring and conservation practice. Future work should focus on testing the system on constrained hardware, expanding datasets and metrics, and pursuing multimodal and distributed approaches to ensure scalability and long-term impact.

FUNDING

The work was carried out with the financial support of the Ministry of Science and Higher Education of the Russian Federation within the framework of realization of the complex project on creation of high-tech production on the theme "Multimodal complex of airport airspace control" (Agreement on granting a subsidy from the federal budget for the development of cooperation between a state scientific institution and an organization belonging to real sector of the economy for the purpose of realization of the complex project on creation of high-tech manufacturing no. 075-11-2025-023 dated February 27, 2025) and within the framework of the Resolution of the Government of the Russian Federation no. 218 dated April 9, 2010 on the basis of the head executor: federal state autonomous educational institution of higher education the Saint Petersburg Electrotechnical University "LETI" (SPb ETU "LETI").

REFERENCES

[1] D. Dziak, D. Gradolewski, S. Witkowski, D. Kaniecki, A. Jaworski, M. Skakuj, and W. J. Kulesza, "Airport wildlife hazard management system," *Elektronika ir Elektrotechnika*, vol. 28, no. 3, pp. 45–53, Jun. 2022. [Online]. Available: https://eejournal.ktu.lt/index.php/elt/article/view/31418

[2] Y. Zhang and Y. Shi, "Bird detection method for airport perimeters based on an improved yolov8," in *Proceedings of the 5th International Conference on Artificial Intelligence and Computer Engineering*, ser. ICAICE '24. New York, NY, USA: Association for Computing Machinery, 2025, p. 389–393. [Online]. Available: https://doi.org/10.1145/3716895.3716964

[3] E. Sabziyan Varnousfaderani and S. A. Shihab, "Bird strikes in aviation: A systematic review for informing future directions," *Aerospace Science and Technology*, vol. 163, p. 110303, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1270963825003748

[4] S. B. Radar, "Birdscan mr1 radar system," 2023, accessed: 2025-08-08. [Online]. Available: https://swiss-birdradar.com/systems/radar-birdscan-mr1/

[5] "Detection of bird activity using radar," 2022, accessed: 2025-08-08. [Online]. Available: https://skybrary.aero/articles/detection-bird-activity-using-radar

[6] "Thermal imaging in ornithology," 2023, accessed: 2025-08-08. [Online]. Available: https://en.wikipedia.org/wiki/Thermal_imaging_in_ornithology

[7] T. Stanczyk, "Masks and boxes: Combining the best of both worlds for multi-object tracking," *arXiv preprint arXiv:2401.12345*, 2024. [Online]. Available: https://arxiv.org/abs/2401.12345

[8] X. Yang, G. Wang, W. Hu, J. Gao, S. Lin, L. Li, K. Gao, and Y. Wang, "Video tiny-object detection guided by the spatial-temporal motion information," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023, pp. 3054–3063.

[9] C. Xu, J. Ding, J. Wang, W. Yang, H. Yu, L. Yu, and G.-S. Xia, "Dynamic coarse-to-fine learning for oriented tiny object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 7318–7328.

[10] Y. Lyu, Z. Liu, H. Li, D. Guo, and Y. Fu, "A real-time and lightweight method for tiny airborne object detection," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Vancouver, BC, Canada, 2023, pp. 3016–3025.

[11] M. Shi, D. Zheng, T. Wu, W. Zhang, R. Fu, and K. Huang, "Small object detection algorithm incorporating swin transformer for tea buds," *PLOS ONE*, vol. 19, no. 3, p. e0299902, 2024.

[12] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 583–596, 2015.

[13] A. Lukezic, T. Vojir, L. Cehovin Zajc, J. Matas, and M. Kristan, "Discriminative correlation filter with channel and spatial reliability," *International Journal of Computer Vision*, vol. 126, no. 7, pp. 671–688, 2018.

[14] N. Wang and D.-Y. Yeung, "Learning a deep compact image representation for visual tracking," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 26. Curran Associates, Inc., 2013, pp. 809–817.

[15] A. Bibi and B. Ghanem, "Multi-template scale-adaptive kernelized correlation filters," 2015.

[16] S. Yadav and S. Payandeh, "Critical overview of visual tracking with kernel correlation filter," *Technologies*, vol. 9, no. 4, p. 93, 2021.

[17] B. Uzkent and Y. Seo, "Enkcf: Ensemble of kernelized correlation filters for high-speed object tracking," *arXiv preprint arXiv:1801.06729*, 2018.

[18] Q. Wang, J. Gao, J. Xing, M. Zhang, and W. Hu, "Dcfnet: Discriminant correlation filters network for visual tracking," *arXiv preprint arXiv:1704.04057*, 2017.

[19] S. Javed, M. Danelljan, F. S. Khan *et al.*, "Visual object tracking with discriminative filters and siamese networks: A survey and outlook," *arXiv preprint arXiv:2112.02838*, 2021.

[20] N. A. Obukhova, A. A. Motyko, A. A. Pozdeev, and K. A. Smirnov, "Automatic detection and tracking of objects in video data with global motion," in *2024 36th Conference of Open Innovations Association (FRUCT)*, 2024, pp. 549–556.

[21] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 886–893.

[22] A. Kirillov, E. Mintun, N. Ravi, H. Mao, L. Rolland, L. v. d. Gustafson, A. Shamsian, I. Alabdulmohsin, X. Chen, I. Misra, P. Dollár, and R. Girshick, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 3992–4003.

[23] A. E. Shchelkunov, V. V. Kovalev, K. I. Morev, and I. V. Sidko, "Metrics for evaluating automatic tracking algorithms," *Izvestiya SFedU. Engineering Sciences*, no. 1, pp. 233–245, 2020, (In Russian).