Small Object Detection for Ornithological Monitoring Tasks

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*—Automatic bird detection represents one of the most critical technical challenges in ornithological monitoring systems, which are relevant for scientific wildlife observation, biodiversity assessment, and practical applications in agriculture and environmental management. Modern monitoring systems require high accuracy under real-world imaging conditions; however, automatic detection of birds is complicated by the presence of small and low-contrast objects embedded in complex and highly detailed natural scenes. An additional challenge is the high intra-class variability, which arises from the diversity of bird species, varying viewpoints, and differences in object size, both due to species-specific morphology and varying distances to the camera.**

This study is dedicated to the development of an effective method for detecting small and low-contrast objects in individual frames of a video stream. The proposed solution is based on a modified SSD-ADSAR architecture enhanced with a dual-stream attention mechanism. On the test dataset, the model achieved mAP@0.5 = 0.876 and mAP@0.5:0.95 = 0.645. The use of synthetically augmented data helped to mitigate the background-type imbalance and improved the model's robustness under complex visual conditions. The practical significance of this work lies in its applicability to real-time ornithological video monitoring systems, as well as to nature conservation, agricultural automation, and scientific ornithological research. The developed method is tailored to typical conditions of ornithological monitoring (such as small, fast-moving objects and cluttered natural backgrounds), and it outperforms existing solutions designed primarily for detecting artificial airborne objects in terms of detection accuracy.

I. Introduction

In recent years, the task of moving object automatic detection in video streams has received increasing attention. This has been facilitated by the development of high-quality imaging equipment and the emergence of computing systems capable of processing visual data in real time. These technological advancements have enabled the practical application of more computationally intensive computer vision methods. Nevertheless, the problem of detecting small objects in complex scenes with high background variability and significant intra-class dispersion remains unresolved and highly relevant. Within-class variability is primarily driven by differences in object scale, appearance, and viewing angle, all of which significantly complicate the training of robust detectors.

The development of video analytics systems for ornithological monitoring faces several challenges, including heterogeneous and cluttered natural backgrounds, dynamically changing lighting conditions, partial occlusions, and the presence of naturally moving elements in the scene. In recent years, interest in automated bird monitoring in natural habitats has grown substantially. To support algorithm development and evaluation, several specialized datasets have been proposed, such as AirBirds [1] and FBD-SV-2024 [2], which include video sequences captured in diverse environmental conditions. In addition, several studies [3]–[5] have presented integrated surveillance systems that combine infrared cameras and radar modules for continuous monitoring, along with adaptations of YOLOv8-based models tailored to bird detection in open natural environments.

In this work, we address the problem of automatic detection of objects of interest under conditions typical for ornithological monitoring in open environments—such as highly variable backgrounds, dynamic illumination, partial occlusion, and pronounced target class variability due to species diversity, object size, and perspective changes.

In real-world bird monitoring scenarios, observation distances typically range from several hundred meters to several kilometers. For instance, technical guidance [6] reports that optical and infrared systems are commonly used for reliable bird detection at distances of 300–600 meters under both daylight and nighttime conditions. More advanced setups employing long-range optical zoom and multi-sensor configurations have demonstrated the ability to detect medium and large birds at significantly greater distances. In particular, a study [7] reports successful detections at up to 2000 meters, and another work [8] using stereo imaging and radar tracking describes detections of large raptors at distances of up to 2800 meters. Similar performance levels are reported in the context of radar-based migration monitoring, where effective detection

ranges reach approximately 1000 meters [9].

The method proposed in this study is specifically tailored for small object detection, where the term "small" denotes targets with a projection size of approximately 15×15 to 20×20 pixels in a Full HD image frame. Under typical viewing angles, this corresponds to the projected size of an average-sized bird at distances of about 800 to 1500 meters—a range typical for ornithological monitoring systems. Notably, this size regime encompasses several of the most operationally significant species, including pigeons and crows, which are both relatively large and widely distributed, thus representing prevalent targets in such applications. Objects of this scale present a significant challenge for most standard detection algorithms, necessitating architectural and training adaptations to maintain high accuracy. Further complicating the problem are computational resource constraints, especially in applications requiring near real-time performance and deployment on mobile or power-limited monitoring platforms.

II. Related Work

Historically, the task of automatic object detection in images and videos began with methods based on predefined feature extraction filters. Early widespread solutions included cascade classifiers such as Viola–Jones [10] and handcrafted features like Haar and LBP [11], primarily used for face detection. Later, more advanced descriptors were introduced, such as SIFT [12], SURF [13], and HOG (Histograms of Oriented Gradients), which proved effective when combined with support vector machines (SVM) in the classical "HOG + SVM" pipeline [14]. However, these approaches suffer from limited generalization capabilities in challenging visual conditions, are sensitive to scale and geometric transformations, and typically demonstrate poor robustness to occlusions and background variations.

Modern video analytics largely relies on deep learning techniques for object detection, segmentation, tracking, and classification in video streams. Some of the first deep learning-based object detectors were two-stage architectures, such as R-CNN, Fast R-CNN, and Faster R-CNN, which achieved high accuracy but exhibited relatively low inference speed.

A major leap in real-time object detection was made possible by one-stage detectors such as SSD (Single Shot Detector) [15]. Introduced in 2015, SSD generates predictions at multiple feature levels in a single forward pass. While it is computationally efficient and effectively utilizes multi-scale features, its performance on complex datasets (e.g., COCO [16]) remains relatively modest.

EfficientDet [17], introduced in 2019, represented a further evolution of one-stage CNN-based detectors, aiming to balance detection accuracy and computational efficiency. It employs the EfficientNet backbone for feature extraction and introduces a Bidirectional Feature Pyramid Network (BiFPN) to better fuse multi-scale features. The smaller variants (D0–D2) enable near real-time inference with moderate mAP, while the larger ones (D6–D7) achieve higher accuracy (around 50–52% mAP on COCO) at the cost of reduced speed, processing only a few frames per second.

Among single-stage convolutional detectors, the YOLO (You Only Look Once) family continues to stand out. Modern versions (YOLOv8–YOLOv11) achieve high detection accuracy with real-time inference speeds. For instance, YOLOv7, released in 2022, achieved 56.8% mAP on the COCO dataset with a throughput exceeding 30 FPS, surpassing earlier detectors in terms of the accuracy–speed trade-off [18]. However, the structural limitations of YOLO-based architectures have led to a plateau in performance gains in subsequent versions.

In 2020, transformer-based object detectors emerged, beginning with DETR (DEtection TRansformer) [19]. DETR combines a CNN-based feature extractor (e.g., ResNet) with a transformer encoder–decoder architecture, which models global object relationships. While DETR achieves high accuracy, it requires long training times, performs suboptimally on datasets with high within-class variability, and suffers from limited spatial resolution in feature representations.

To address these limitations, Deformable DETR [20] was introduced. It replaces standard attention with deformable attention, focusing computation on a sparse set of key sampling points around reference locations, and employs multi-scale feature maps (via FPN). This approach significantly improves performance on small objects and achieves comparable accuracy to DETR (43–45% AP) while reducing training time (e.g., 50 epochs).

In 2022, the DINO detector [21] was proposed, currently considered one of the most accurate and efficient DETR-based models. It incorporates several key innovations, such as robustness to slight misalignments in annotations, improved anchor box initialization, and a two-stage bounding box prediction module. DINO achieves 49.4% AP on COCO with a ResNet-50 backbone after 12 training epochs and up to 51.3% AP after 24 epochs.

DiffusionDet [22] extends object detection into the domain of diffusion models. Instead of predicting bounding boxes directly, it models the detection task as a denoising process, similar to how generative diffusion models work. The model learns to progressively refine noisy representations into accurate object locations and categories through iterative inference steps. On COCO with a ResNet-50 backbone, DiffusionDet achieves approximately 45.8% mAP.

Despite significant advances in transformer-based, diffusion-based, spatiotemporal, and hybrid detection models, their practical use in resource-constrained real-time video analytics systems remains limited. Transformer-based and hybrid architectures typically require substantial computational resources during both training and inference. While diffusion models offer greater flexibility in modeling, they still fail to meet the real-time throughput requirements for streaming video input.

In scenarios with limited computing power and the need for near real-time performance, lightweight single-stage convolutional detectors—specifically adapted to the target deployment conditions—remain the most practical and balanced solution.

III. Proposed Solution

For the task of small object detection, this work employs a modified version of the SSD-ADSAR architecture [23], specifically adapted to the problem's unique challenges. We extended the Multi-Head Dual Stream Attention block from three to five heads, each configured to focus on either global or local attention. Each head now processes tokens of different resolutions, enabling robust detection of both small and large objects, even against complex backgrounds. In addition, we increased the network depth while proportionally reducing the width of individual layers, keeping the overall parameter count unchanged. This modification resulted in deeper feature representations without increasing computational cost, and led to improved prediction accuracy.

Therefore, the model was designed to be robust to background changes caused by camera motion and capable of detecting low-contrast and small-scale objects. The overall structure of the proposed solution is illustrated in Fig 1. The key component of the model is the dual-stream self-attention module, ADSAR.

Let us consider the ADSAR module in more detail. Let $X \in \mathbb{R}^{H\times W\times D}$ be the input feature representation to the ADSAR block after convolutional operations in the feature extraction stage (where H, W, and D denote the spatial dimensions and depth of the input tensor). The features are projected into queries Q, keys K, and values V as follows:

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V,$$

where $W^Q, W^K, W^V \in \mathbb{R}^{D\times D'}$ are learnable projection matrices.

To separate global and local information, two attention masks M_l (local) and M_g (global) are introduced:

$$M_l(i,j) = \begin{cases} 0, & \text{if } j \in N(i), \\ -\infty, & \text{otherwise} \end{cases}$$

$$M_g(i,j) = \begin{cases} 0, & \text{if } j \notin N(i), \\ -\infty, & \text{otherwise} \end{cases}$$

where $N(i)$ denotes a fixed-size neighborhood around position i. Using these masks, separate attention weights are computed for the local and global streams:

$$A_l = \text{softmax}\left(\frac{QK^T}{\sqrt{D'}} + M_l\right),$$

$$A_g = \text{softmax}\left(\frac{QK^T}{\sqrt{D'}} + M_g\right).$$

The attention responses for the local and global streams are then calculated as:

$$R_l = A_l V, \quad R_g = A_g V,$$

and their difference is computed as:

$$\text{Res} = R_l - R_g,$$

which serves to enhance relevant features of small objects. Finally, the resulting representation is integrated back with the original features as follows:

$$X' = X + \text{Res} \cdot F,$$

where $F \in \mathbb{R}^{D'\times D}$ is a learnable projection matrix that maps the result back to the original dimensional space.

Thanks to this architecture, SSD-ADSAR achieves superior mAP metrics for small object detection and better localization accuracy compared to other solutions such as RetinaNet and YOLOv11. A detailed comparison of SSD-ADSAR performance with competing architectures, including YOLOv8, SSD, and DETR, is provided in [23]. In particular, experimental results show that SSD-ADSAR outperforms YOLOv8 by more than 5% in key metrics such as AP and IoU.

IV. Experimental Evaluation

A. Motivation and Model Selection

Inference speed is a critical factor when selecting a neural network architecture for real-time or near-real-time computer vision tasks. It is commonly believed that models based on the EfficientDet architecture offer performance comparable to SSD in terms of inference speed, while YOLOv8 and newer architectures tend to outperform SSD in this regard. However, in practical engineering applications, the choice of architecture depends on a variety of factors. A recent comprehensive survey [24] convincingly demonstrates that inference speed is largely determined by system-level parameters.

When comparing performance across models, it is essential to specify the exact configurations under comparison, including model variants (particularly their size), hardware specifications (e.g., CPU architecture, memory bandwidth, GPU availability and type), and runtime conditions (such as parallelization strategies, buffering techniques, and inference frameworks).

Thus, the inference speed of a given architecture must be evaluated in the context of the specific requirements and constraints of the target vision system. In our case, the SSD-ADSAR-based model, when executed with a parallelized inference mechanism (processing independently the image fragments into which the original frame is divided), achieves a throughput exceeding 40 frames per second for an input resolution of 1920×1080 on an NVIDIA RTX 3090 GPU. This performance is sufficient from a systems perspective, especially considering the standard input rate of 30 FPS from typical camera hardware.

B. Data Description and Preprocessing

Video streams obtained from ornithological monitoring systems are typically captured at high resolutions, such as Full HD (1920×1080) or 4K (3840×2160), enabling the preservation of fine details in the scene. However, using full-resolution images during model training introduces several limitations, including increased memory consumption, reduced processing speed, and a substantial growth in final model size. A common solution to these issues is image downscaling. Nevertheless,

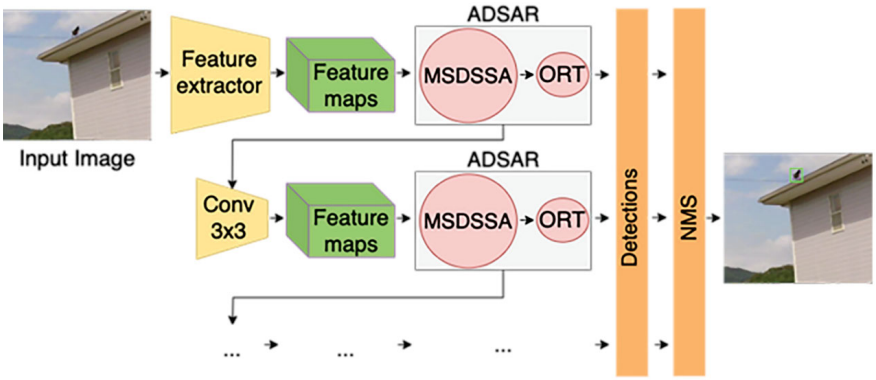

Fig. 1. The scheme of a convolutional neural network SSD-ADSAR

in small object detection tasks, such preprocessing is often counterproductive, as it leads to the loss of critical fine-grained visual features necessary for accurate localization.

To train the detection model, a dataset of 10000 original images was collected from various imaging devices. These images were preprocessed by splitting them into fixed-size patches of 640×640 pixels, a format commonly used in SSD-based architectures. During the cropping process, care was taken to avoid partial inclusion of objects of interest at patch boundaries. The annotation data served as a guide, and patches were generated in a way that centered around target objects. As a result of this preprocessing, the dataset was expanded to 23061 training images.

C. Dataset Challenges and Augmentation

Analysis of the initial training dataset revealed two key challenges. First, there was a pronounced imbalance in background contexts: the majority of frames depicted uniform scenes (e.g., low-detail skies or homogeneous forest areas), whereas scenes containing urban environments, man-made structures, or mixed landscapes were significantly underrepresented. Second, the target objects (birds) exhibited high within-class variability in terms of shape, spatial positioning, and especially size — ranging from just a few pixels to several dozen pixels in projected length. These factors reduced the model's generalization capability and necessitated additional measures in constructing the training dataset.

Another challenge was the limited availability of high-quality annotated datasets suitable for training models on small object detection. For example, although the AirBirds dataset [1] is widely used, it suffers from several common issues seen in open-source corpora: inconsistent annotation quality, inaccuracies in object boundary placement, and overly large bounding boxes around targets — issues that are particularly critical when working with small objects. In scenarios where precise annotation is essential, such limitations significantly constrain the usability of existing public datasets.

To increase the representativeness of the training data and improve model robustness, synthetic image generation was employed. Synthetic samples were generated using the Stable Diffusion XL model [25], fine-tuned via the LoRA method [26] (rank 128) on a small set of real bird images from the original dataset. Fig. 2 shows an example of a generated image with visually plausible content and an annotated region of interest. To quantitatively assess the quality of the synthetic data, the Frechet Inception Distance (FID) metric was computed, resulting in a score of 19.44, which corresponds to a good, near-realistic level of image generation.

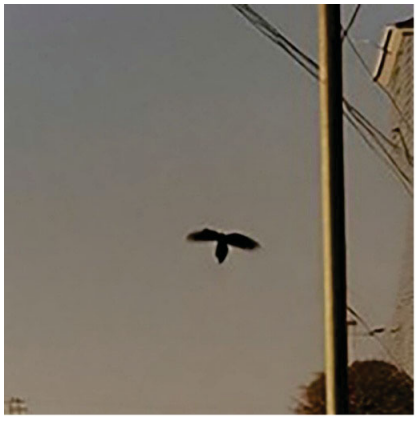

Fig. 2. Example of a synthetic bird image generated with Stable Diffusion XL and LoRA, used to augment the training dataset and balance background types

The integration of synthetic data helped to mitigate the background distribution imbalance and improve the generalization ability of the model. As a result, the final training dataset was expanded to 28061 images.

D. Training Setup and Evaluation Metrics

As in [23], training was performed using an NVIDIA RTX 3090 GPU throughout 100 epochs. The AdamW optimizer was employed, a variant of the well-known Adam optimizer featuring decoupled weight decay regularization, with the following parameters: $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay factor of 0.05. The batch size was set to 16. The initial learning rate was 10^{-8}, and cosine annealing was applied throughout training, gradually decreasing the learning rate to zero.

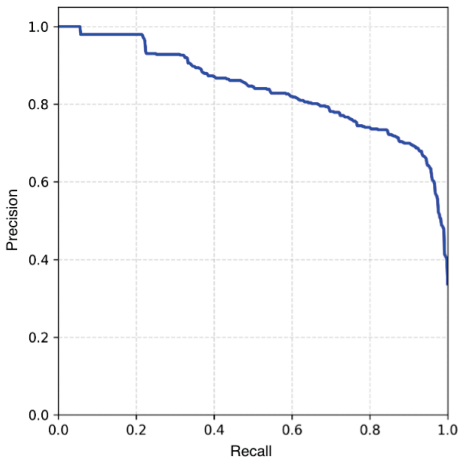

Fig. 3. The precision-recall curve for the model performing the detection

Model evaluation was based on the precision–recall (PR) curve, which accounts for both precision and recall. The resulting area under the PR curve (AUC-PR) was 0.83. The resulting precision–recall curve is shown in Fig. 3.

The achieved detection performance is considered high, especially given that object detection in the context of this study typically serves as an intermediate step rather than a final goal. In monitoring systems, the output of the neural detector is commonly used to initialize automatic object tracking, for example, by generating a strobe region or defining a capture zone for a tracking module.

It should be noted that the reported accuracy metrics were calculated for objects in the training set with a maximum projected size of 20 pixels or greater, which generally corresponds to target parameters for automated bird detection systems. These objects accounted for 87% of the training data, while objects with a maximum projected size of 10 to 20 pixels made up 9%, and those between 5 and 10 pixels (with a minimum of 5×5 pixels) represented 4%. Accuracy metrics were not computed for the latter two categories; these samples were included to diversify the dataset and reflect the real-world conditions in which smaller-than-target projections may also be present and, in some cases, successfully detected. In particular, the third category of objects (with minimal sizes) included high-contrast targets on uniform backgrounds, which enabled reliable detection within the proposed algorithm (see Fig. 4).

E. Experimental Setup

The experimental study was conducted for two detectors: the prototype model from [23] and the modified model proposed in this work. Both models were trained on the same set of real images; however, the second model additionally incorporated synthetic data. As part of the adaptation of SSD-ADSAR to the target task, the attention block was expanded and the layer structure was optimized, while maintaining the total number of parameters. The evaluation was carried out on a separate

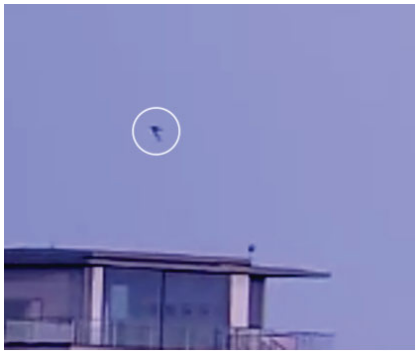

Fig. 4. Illustration of a minimum-size training object (5×5 pixels), representing the lower bound of detectable targets

test set of bird images not used during training. Examples of such images are shown in Fig. 5.

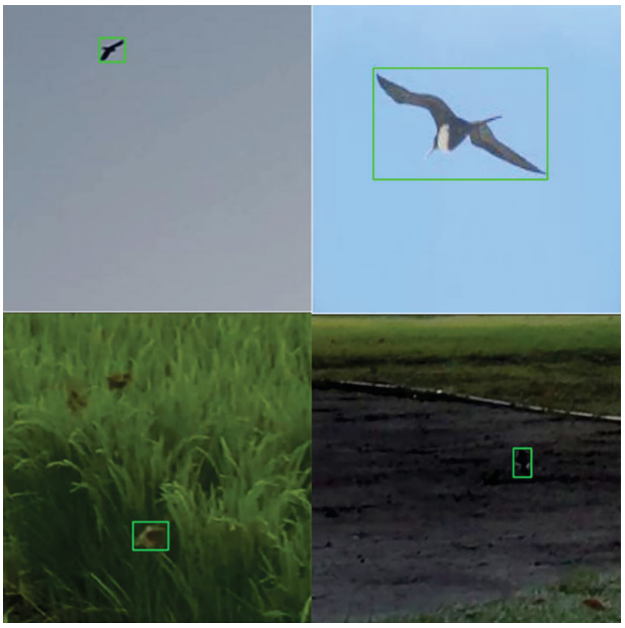

Fig. 5. Sample test images demonstrating varied environmental conditions for model evaluation

During the experiments, the position of each object in every image was determined automatically. The accuracy of both models was assessed using standard metrics: precision, recall, F1-score, as well as mean average precision at two thresholds, mAP@0.5 and mAP@0.5:0.95. The evaluation results are presented in Figs. 6, 7, and 8.

F. Results

The proposed detector model demonstrates a significant improvement in recall, showing a 9.2% increase compared to the prototype. This led to a 4.8% gain in the F1-score. The mAP values increased by 6.4% and 2.8%, respectively, confirming the improved object detection performance of the new model. Despite the gain in recall, precision remained nearly unchanged; the proposed detector misses fewer objects without compromising precision.

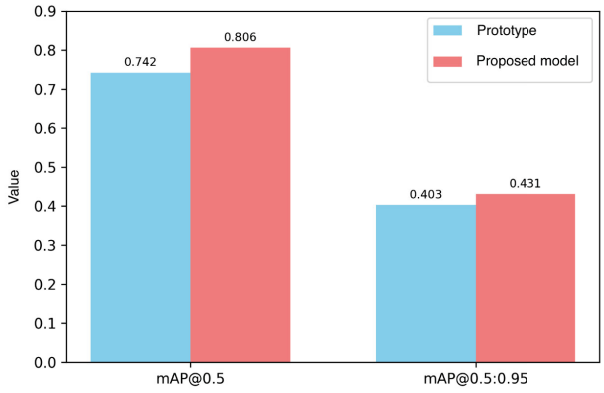

Fig. 6. mAP comparison between the prototype and the proposed model

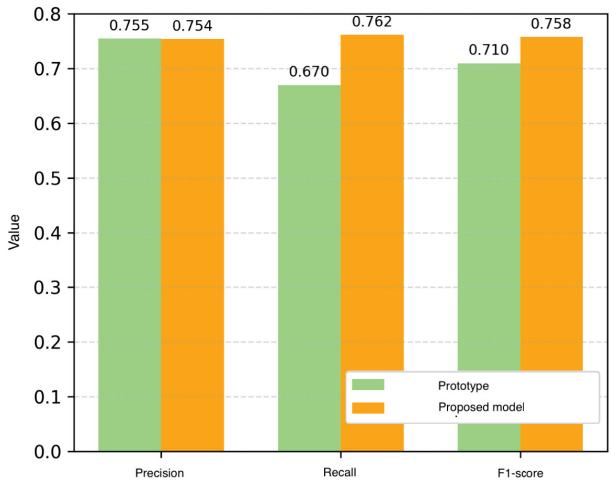

Fig. 7. Precision, recall and F1-score comparison between the prototype and the proposed model

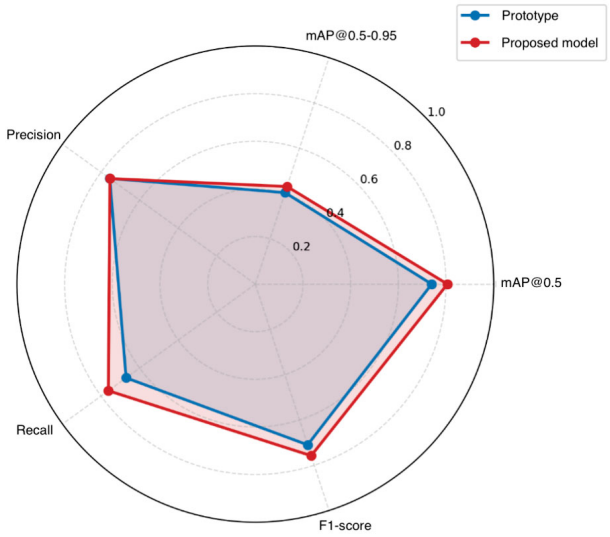

Fig. 8. Comparison of metrics for the prototype and the proposed model using a radar chart

Fig. 9. Example from the COCO dataset [27], illustrating differences between generic training data and ornithological scenes

To further assess the proposed solution, a comparison was conducted with the results of the standard pre-trained YOLO11n model on the same validation dataset. This model was selected as a baseline for comparative evaluation due to the following reasons:

- YOLO11n is widely considered an informal industry standard for object detection tasks, making it a suitable benchmark for assessing the complexity of the specialized detection problem addressed in this study;
- the model was originally trained on public datasets that include the "bird" class, although under conditions not specifically tailored to ornithological monitoring systems. Nevertheless, the training data included scenes generally similar to those encountered in bird monitoring systems (see Fig. 9).

YOLO11n achieved the following results on the target data: F1-score = 0.33, mAP@0.5 = 0.346. These results are understandable, as the pre-trained YOLO11n model was trained on datasets and designed for tasks with entirely different characteristics. The purpose of reporting these results is not to criticize a well-known architecture, but to highlight the fact that the task of detecting objects of interest in monitoring systems has distinct data and scene-specific challenges. These challenges make it impractical to rely solely on off-the-shelf models for such applications.

Experiments showed that the proposed method performs reliably in both simple and complex scenes. It ensures sufficient processing speed and delivers high detection accuracy across a wide range of conditions, including challenging scenarios.

V. Discussion and Conclusion

This study presents a method for detecting objects of interest, specifically designed for the analysis of video data in ornithological monitoring tasks. To detect objects of interest, a convolutional architecture, SSD-ADSAR, with a dual-stream self-attention module, is proposed. To better adapt the detector architecture to the specific task, we enhanced the attention mechanism by increasing the number of attention heads and adjusted the network depth-to-width ratio to deepen feature representations while preserving the overall parameter count. These structural improvements contributed to more accurate

detection of both small and large objects in complex natural scenes without increasing computational costs. The integration of synthetic images into the training dataset, generated using Stable Diffusion XL with LoRA fine-tuning, allowed for the mitigation of dataset imbalance and enhanced the generalization ability of the model.

The proposed model achieved an mAP@0.5 of 0.806, which is 6.4% higher than the baseline, and an mAP@0.5:0.95 of 0.431, representing a 2.8% improvement. The overall F1 score improved by 4.8%. A comparison with the industrial detector YOLO11n, which achieved F1 = 0.33 and mAP@0.5 = 0.346 on the same validation data, confirms the relevance of developing an architecture adapted to the detection of small-sized objects. The results of the experimental evaluation demonstrate that the proposed method provides accurate detection of small objects (mAP@0.5 = 0.876 and mAP@0.5:0.95 = 0.645) and delivers high runtime performance, close to real-time (over 40 frames per second at a resolution of 1920×1080 using an NVIDIA RTX 3090 GPU).

While the proposed model achieves real-time performance on modern GPUs, its deployment on edge devices and low-power platforms remains challenging. In resource-constrained environments, there is an inherent trade-off between maintaining high detection accuracy and reducing computational load to meet strict latency and energy requirements. Future work will therefore focus on model compression, pruning, and quantization strategies to achieve a more favorable balance between accuracy and efficiency.

Current experiments rely solely on visual input. However, ornithological monitoring often requires robustness under poor visibility, occlusion, or clutter. A promising direction is the integration of complementary modalities, such as radar or infrared sensors, which could enhance detection reliability in low-light or dense habitats. Exploring multimodal architectures will broaden the applicability of the approach across a wider range of ecological scenarios.

In addition, the use of generative models to augment training datasets offers benefits in terms of diversity and scale, but it also introduces risks. Synthetic images may embed hidden biases, overrepresent certain visual patterns, or reduce ecological validity compared to real-world observations. To mitigate these risks, future efforts should emphasize rigorous evaluation of synthetic data quality and ensure balanced integration with real datasets. A combined strategy, leveraging both synthetic and naturalistic data, appears essential for maintaining ecological fidelity and avoiding systematic biases in monitoring outcomes.

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] H. Sun, Y. Wang, X. Cai, P. Wang, Z. Huang, D. Li, Y. Shao, and S. Wang, “Airbirds: A large-scale challenging dataset for bird strike prevention in real-world airports,” 2023. [Online]. Available: https://arxiv.org/abs/2304.11662

[2] Z. Sun, Z. Hua, H. Li, Z. Qi, X. Li, Y. Li, and J. Zhang, “Fbd-sv-2024: Flying bird object detection dataset in surveillance video,” *Scientific Data*, vol. 12, no. 1, p. 530, 2025. [Online]. Available: https://doi.org/10.1038/s41597-025-04872-6

[3] 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

[4] 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

[5] 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

[6] DHI Group, “Bat and bird monitoring guidance,” https://www.dhigroup.com/upload/publications/2023/bat-and-bird-monitoring-guidance.pdf, 2023, accessed: 2025-08-07.

[7] Vattenfall Wind Power Ltd, “Aowfl aberdeen offshore wind farm: Final seabird study report,” https://group.vattenfall.com/uk/contentassets/1b23f720f2694bd1906c007effe2c85a/aowfl_aberdeen_seabird_study_final_report_20_february_2023.pdf, Vattenfall, Tech. Rep., 2023, accessed: 2025-08-07.

[8] DHI Group, “Muse multi-sensor bird detection application,” https://cms.dhigroup.com/media/ginnqtvs/muse-whitepaper-mar-2024.pdf, Tech. Rep., 2024, accessed: 2025-08-07.

[9] G. Norevik, A. Hedenström, and S. Åkesson, “Radar monitoring of nocturnal bird migration: assessing altitude and density over time,” *Animals*, vol. 14, no. 23, p. 3353, 2024, accessed: 2025-08-07. [Online]. Available: https://www.mdpi.com/2076-2615/14/23/3353

[10] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2. IEEE, 2001, pp. 511–518.

[11] R. Lienhart and J. Maydt, “An extended set of haar-like features for rapid object detection,” in *Proceedings of the International Conference on Image Processing (ICIP)*, vol. 1. Rochester, NY, USA: IEEE, 2002, pp. 900–903.

[12] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[13] H. Bay, T. Tuytelaars, and L. Van Gool, “SURF: Speeded up robust features,” in *Computer Vision – ECCV 2006*, ser. Lecture Notes in Computer Science, A. Leonardis, H. Bischof, and A. Pinz, Eds., vol. 3951. Berlin, Heidelberg: Springer, 2006, pp. 404–417.

[14] 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.

[15] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision – ECCV 2016*, ser. Lecture Notes in Computer Science, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., vol. 9905. Springer, 2016, pp. 21–37.

[16] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Computer Vision – ECCV 2014*, ser. Lecture Notes in Computer Science, vol. 8693. Springer, 2014, pp. 740–755.

[17] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10 778–10 787.

[18] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 7464–7475.

[19] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," *arXiv e-prints*, 2020. [Online]. Available: https://arxiv.org/abs/2005.12872

[20] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021. [Online]. Available: https://arxiv.org/abs/2010.04159

[21] H. Zhang, F. Li, S. Liu, L. Wang, X. Zhang, X. Qi, and P. Luo, "DINO: DETR with improved denoising anchor boxes for end-to-end object detection," *arXiv e-prints*, 2022. [Online]. Available: https://arxiv.org/abs/2203.03605

[22] S. Chen, P. Sun, Y. Song, and P. Luo, "Diffusiondet: Diffusion model for object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 19 773–19 786. [Online]. Available: https://arxiv.org/abs/2211.09788

[23] A. Samarin, A. Savelev, A. Toropov *et al.*, "Adsar: Advanced dual-stream attention and reweighting for small object detection," *Pattern Recognition and Image Analysis*, vol. 35, pp. 211–218, 2025. [Online]. Available: https://doi.org/10.1134/S1054661825700129

[24] D. K. Alqahtani, M. A. Cheema, and A. N. Toosi, "Benchmarking deep learning models for object detection on edge computing devices," in *Service-Oriented Computing. ICSOC 2024*, ser. Lecture Notes in Computer Science, W. Gaaloul, Q. Z. Sheng, Q. Yu, and S. Yangui, Eds., vol. 15404. Singapore: Springer, 2025, pp. 185–202. [Online]. Available: https://doi.org/10.1007/978-981-96-0805-8_11

[25] R. Rombach, A. Blattmann, D. Müller, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 10 674–10 685. [Online]. Available: https://doi.org/10.1109/CVPR52688.2022.01042

[26] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint*, 2021, presented at ICLR 2022. [Online]. Available: https://arxiv.org/abs/2106.09685

[27] "Microsoft coco dataset," https://cocodataset.org, accessed: 2025-07-01.