Multi-scale ship detection using high-resolution aerial imagery

1st Thomas Kraft
Institute of Space Research (WR)
German Aerospace Center (DLR)
Berlin, Germany
thomas.kraft@dlr.de

2nd Cristina Melnic
Institute of Space Research (WR)
German Aerospace Center (DLR)
Berlin, Germany
cristina.melnic@dlr.de

3rd Daniel Hein
Institute of Space Research (WR)
German Aerospace Center (DLR)
Berlin, Germany
daniel.hein@dlr.de

4th Björn Piltz
Institute of Space Research (WR)
German Aerospace Center (DLR)
Berlin, Germany
bjoern.piltz@dlr.de

5th Matthias Geßner
Institute of Space Research (WR)
German Aerospace Center (DLR)
Berlin, Germany
matthias.gessner@dlr.de

6th Ralf Berger
Institute of Space Research (WR)
German Aerospace Center (DLR)
Berlin, Germany
ralf.berger@dlr.de

***Abstract*—High-resolution airborne imagery (ground sampling distance, GSD, of 3–10 cm) from DLR's Modular Aerial Camera System (MACS) enables real-time maritime situational awareness, yet the extreme variation in scale of ship instances — spanning several orders of magnitude — poses substantial challenges for automatic detection. We investigate YOLOv8 and YOLO11 detectors under two regimes: (a) slice-based training with sliding-window inference and (b) slice-trained models further fine-tuned on resized full images and evaluated under full-image inference. A controlled ablation over slicing overlap and merging thresholds reveals a consistent precision–recall trade-off across object scales that cannot be resolved through postprocessing tuning alone. Fine-tuning on resized full images yields a more robust operating point, achieving up to 91% recall under full-image inference while running at least 7.5× faster than sliced inference.**

***Index Terms*—ship detection, aerial imagery, MACS**

I. Introduction

Automatic ship detection in airborne imagery is a key capability for maritime surveillance, traffic monitoring and emergency response. In this context, airborne optical sensors offer a viable option for capturing large-area, high-resolution data. DLR's Modular Aerial Camera System (MACS) [1] is one such system, enabling real-time mapping of overflown areas [2]. It can integrate sensors from different spectral ranges [3], and has been designed for efficient use on aerial missions [4]. Building on this capability, two scientific flight campaigns conducted in 2023 and 2024 resulted in a comprehensive dataset of high-resolution imagery. The data cover complex maritime environments, including the Elbe river and the port of Hamburg, and include a wide range of vessel types and sizes. The images were captured using a multi-head imaging configuration to enable wide-area coverage and a ground sampling distance (GSD) of 3 to 10 cm per pixel.

Automatic ship detection in this setting is complicated by several co-occurring challenges. Ship sizes span an extreme range: from 2 m dinghies to 400 m cargo vessels, corresponding to bounding-box areas between ∼0.004% and 36% of image area — a scale variation of approximately 10^4. Full-resolution images reach 8 424×6 032 pixels, exceeding the typical input of standard detectors. GSD variation across the images further complicates training, as downscaling changes the apparent object scale and introduces a training-to-inference distribution mismatch. Ship counts per image are highly skewed: the majority of images contain no ships, a smaller number contain a handful, and dense harbour scenes are rare. Finally, sliding-window approaches introduce a post-processing challenge: overlap and merging threshold jointly govern the precision–recall trade-off, and no single global setting performs well across the full scale range.

To address these challenges, we integrate deep convolutional neural networks into MACS's real-time workflow.

II. Materials and Methods

A. Modular Aerial Camera System (MACS)

MACS, developed by DLR's Institute of Space Research, is an airborne imaging system for real-time data acquisition and processing. The system comprises an embedded computing unit, a multi-frequency Global Navigation Satellite System (GNSS) receiver, an industrial-grade Inertial Measurement Unit (IMU), and multiple communication interfaces. Its modular architecture enables simultaneous integration of sensors from visible range up to thermal infrared, with all sensors radiometrically and geometrically calibrated to ensure consistent and physically meaningful data [5]. During operation, each captured image is preprocessed, georeferenced via tight coupling of GNSS and IMU, and stored at full resolution. A comprehensive software suite is available for data access, visualization, and export, enabling seamless integration into downstream processing pipelines [6].

B. Aerial Image Data

Data was acquired during two flight campaigns using a three-camera setup consisting of one nadir-looking and two

oblique-looking RGB cameras. The oblique cameras were oriented to the left and right side of the aircraft with a fixed viewing angle of 34° relative to nadir, enabling an extended ground coverage beyond the central viewing direction. All cameras were industrial-grade sensors (SVS-Vistek HR51CXGE) equipped with high-quality, mechanically modified lenses from the Leica APO-Summicron series. The nadir camera was fitted with a 50 mm lens, while the oblique cameras used 75 mm lenses to compensate for the increased object distance. Data acquisition was performed over the Elbe river and the port of Hamburg at flight altitudes of 1,000 ft and 2,000 ft above ground level, resulting in a swath width of up to 1.3 km. The campaigns were conducted on September 27, 2023, and October 12, 2024, respectively, under stable daylight conditions with clear weather. The resulting dataset comprises high-resolution images capturing ships of various types and sizes under realistic operating conditions. An incomplete overview of the ships is given in figure 1. For the subsequent ship-detection experiments, the proprietary MACS raw images were color-corrected, de-vignetted, and converted to radiometrically normalized 8-bit JPEG images.



Fig. 1: Exemplary overview of ship types in the aerial images

C. Training Data

SeaDronesSee (pretraining only): We initialised models using weights that we pretrained on the publicly available SeaDronesSee Object Detection v2 dataset [7], which contains high-quality maritime drone imagery. The classes *boat* and *jetski* were merged into a single *watercraft* class to align with our single-class detection objective.

Full Images: A proprietary dataset of 1 362 directly georeferenced MACS images was curated for ship detection. Images have a typical resolution of 8 416 × 6 032 pixels at a GSD predominantly around 3 cm per pixel. In total 4 234 ships have been labelled using the open-source software 'Computer Vision Annotation Tool' (CVAT) [8]. Bounding-box areas span over $\sim 0.004\%$–36% of image area, i.e. with a scale variation of roughly 10^4. Following a standard clustering algorithm , the relative area thresholds of 0.052% and 0.369% were chosen to define small (S), medium (M), and large (L) ships. The categorized distributions represented in figure 2 emphasize the increased tendency of the distribution towards a long-tail as the size progresses from small to large. The category boundaries are not sharp by nature — at the transition between adjacent size classes, both the physical appearance of ships and their bounding-box dimensions overlap, making a clear separation, and thus a per-size detection model, difficult.

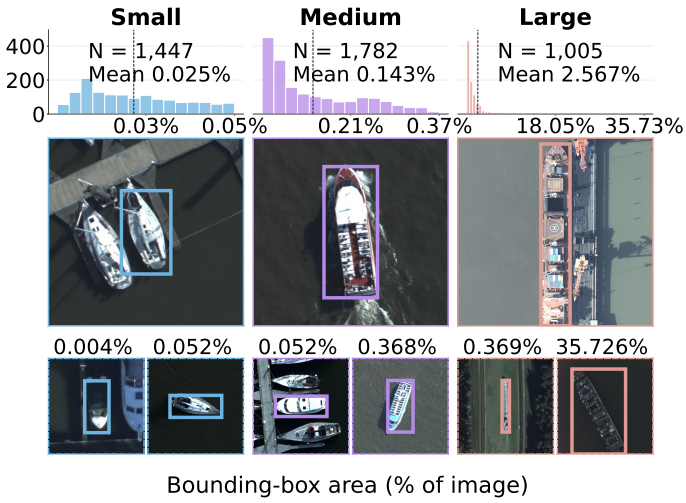


Fig. 2: The top panel shows the count distributions of the small, medium, and large ships by bounding box relative to the full image area. The middle panel shows examples of full images of ships with mean bounding box areas. The bottom thumbnails show ships with boxes with the minimum and maximum areas.

Sliced images: The full-resolution images were sliced into 1 280×1 280 pixel crops with 20% overlap to support tile-based training and to serve as the baseline for sliding-window inference experiments. After balancing the dataset with regard to the number of negatives, the full dataset had 2 978 training samples with 5 144 vessel instances, - 910 more vessels than the initial dataset arising from ships at tile boundaries appearing in two overlapping crops.

D. YOLO

Ship detection experiments used the Ultralytics implementations of YOLOv8 [9] and YOLO11 [10]. Both architectures incorporate feature pyramid structures that support multi-scale representation and are well suited for handling extreme size variation [11]. We evaluated nano (n), small (s), medium (m), and large (l) model variants. Training was performed on two NVIDIA RTX 6000 Ada GPUs (48 GB VRAM each). Unless otherwise stated, Ultralytics default settings were used with multi-scale training enabled. The chosen evaluation metrics on the validation sets were: Precision (P), Recall (R), F_1-score (harmonic mean of P and R), mean Average Precision (mAP) at an Intersection over Union (IoU) threshold of 0.50 (mAP@50) and averaged over IoU 0.50–0.95 (mAP@50:95).

E. Training Protocol

Stage 1 – Tile-Based Training: Models were trained for 100 epochs on 1 280×1 280 pixel MACS tiles, initialised from SeaDronesSee-pretrained weights [7]. The slice size was chosen as a practical compromise between inference computation cost and the need to capture the largest vessels: a

TABLE I: Validation performance metrics [%] from Ultralytics training. Best values per metric within each block are in bold.

Model	P	R	F_1	mAP@50	mAP@50:95
Stage 1: Slices of 1280×1280 px					
YOLO11s	91.0	93.5	92.2	**96.9**	**78.9**
YOLOv8n	89.7	**95.1**	**92.3**	96.8	78.3
YOLO11l	90.0	92.8	91.4	96.5	77.3
YOLO11m	**91.3**	91.2	91.2	96.6	76.2
Stage 2: Full images, scaled to 1280×1280 px					
YOLO11s	**92.8**	**92.6**	**92.7**	**95.2**	**74.1**
YOLOv8n	90.7	91.1	90.9	94.2	71.3

1 280×1 280 tile covers approximately 1.6 Mpx, slightly below the maximum single-ship bounding-box area of 1.8 Mpx.

Stage 2 – Full-Image Scaling: Stage 1 weights were further fine-tuned for 100 epochs using full images resized to 1 280 pixels on both sides. This regime biases the model toward global context and reduces reliance on post-hoc merging of sliced predictions.

F. Sliding-Window Inference (SAHI)

Sliding-window inference was performed following the SAHI framework [12]. Images were sliced with 10%, 25%, and 50% overlap (OV), and predictions from all tiles were merged using the GREEDYNMM postprocessing algorithm with an Intersection Over Smaller area (IOS) matching metric. The Matching Threshold (MT) was varied over $\{0.2, 0.5, 0.8\}$ to explore the precision–recall performance at different scales.

III. Results

A. Stage 1: Training on sliced images

Models trained on 1 280×1 280 pixel slices achieved strong validation performance: YOLO11s reached mAP@50 = 96.9% and mAP@50:95 = 78.9%, while YOLOv8n achieved the highest recall at 95.1% (Table I). Given their favourable performance-compute trade-off relative to the larger variants, subsequent experiments focused on these two models.

B. Stage 2: Training on rescaled full images

Despite strong tile-validation metrics, Stage 1 models applied to full-resolution images via sliding-window inference produced substantially degraded results, with precision and recall falling to approximately 50% or below (see row 'Stage 1+SAHI', column 'All' in Table II). To address this, Stage 1 weights were further fine-tuned on resized full images. This improved robustness substantially: YOLO11s achieved mAP@50 = 95.2% and recall of 92.6% on the validation set (Table I). This is the only model in our evaluation to reach $P, R \gtrsim 90\%$, which we take as a practical deployment threshold.

Training dynamics for both stages are shown in fig. 3, confirming steady convergence without overfitting, and suggesting achievable gains with further training.

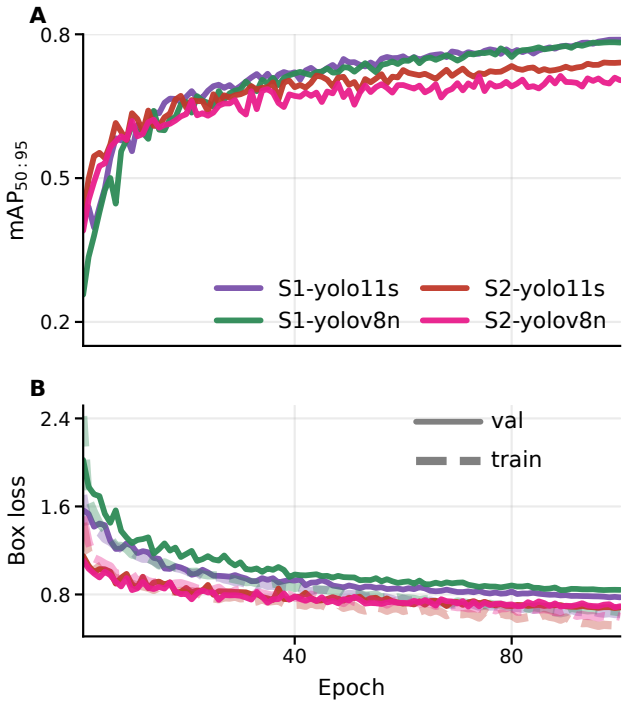


Fig. 3: Training dynamics for Stage 1 (tile-based, cool colours: purple/green) and Stage 2 (full-image scaling, warm colours: red/pink). Panel A shows validation $\text{mAP}_{50:95}$ over epochs; both stages converge steadily without plateauing at epoch 100, suggesting potential gains with extended training. Panel B shows box localisation loss for validation (solid) and training (dashed); the close alignment of both tracks confirms the absence of overfitting in either regime.

C. Sliding-Window Inference Analysis

Per-size results are summarised in Table II and Fig. 4. Full-image Stage 2 achieves the strongest overall performance across scales, with no SAHI configuration matching its precision-recall balance.

Stage 1 nearly fails under full-image inference, recovering only a single small-ship detection. SAHI restores partial recall for small and medium vessels (+38.8 and +8.4 pp over full-image inference), but for large ships recall improves only marginally (+3.2 pp) while precision collapses from 61.0% to 21.9%. Panel (d) reveals why: the best SAHI run inflates large-ship predictions to 224 against only 95 ground-truth instances, far exceeding the full-image count of 75 without recovering proportionate true positives. The resulting precision and recall are insufficient for practical deployment at any configuration.

Stage 2 exhibits the opposite failure. Full-image inference produces prediction counts close to ground truth at every scale, whereas SAHI suppresses these to well below ground truth across all size classes (panel (e)). The tight clustering of SAHI markers in panels (a)–(c) confirms this suppression is invariant to OV and MT configuration: the model loses confidence on cropped context it was not trained on, and postprocessing offers no remedy. Where tiled inference is

TABLE II: Per-size detection performance (IoU=0.5). Values shown as Precision / Recall (%). SAHI rows use best F_1 score per size (parameters in parentheses).

Model	Small (206)	Medium (190)	Large (95)	All (491)
Stage 1	100.0 / 0.5	76.0 / 40.0	61.3 / 48.4	74.4 / 26.7
+SAHI	69.8 / 39.3	68.7 / 48.4	21.9 / 51.6	49.4 / 45.2
	(ov10, mt50)	(ov25, mt50)	(ov25, mt20)	(ov10, mt20)
Stage 2	79.2 / **88.8**	**90.3 / 88.4**	**94.4 / 89.5**	**88.2 / 91.0**
+SAHI	80.6 / 48.5	91.5 / 62.6	83.3 / 57.9	85.9 / 57.0
	(ov25, mt80)	(ov10, mt80)	(ov10, mt80)	(ov25, mt80)

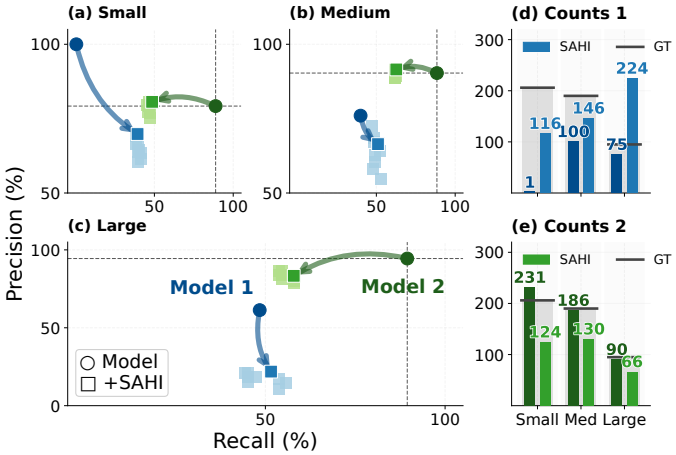


Fig. 4: (a)–(c) Precision–recall operating points for full-image inference (circles) and all SAHI configurations (squares) for small, medium, and large ships. Dashed crosshairs mark the best-performing configuration overall. (d)–(e) Detection counts under full-image and SAHI inference for Stage 1 and Stage 2 respectively, compared to ground-truth instance counts (GT).

unavoidable, Stage 2 with SAHI remains a usable fallback, at the cost of 25–40 pp recall relative to the full-image baseline.

Full-image inference also carries a substantial computational advantage. The reported 7.5× speedup (0.525 vs 3.932 s/image) is measured at minimum overlap (OV = 10%). For the OV = 25% and OV = 50% settings, the ones that yield best F_1 for Stage 1, this factor grows to approximately 15× and 30×, respectively. Aggressive downscaling introduces a detection floor for the smallest vessels, a limitation at very fine GSD.

IV. Conclusion and Outlook

We studied ship detection in high-resolution aerial imagery of approximately 8k×6k pixels, with ship scales spanning ∼0.004% to 36% of image area. While Stage 1 training (SeaDronesSee pretraining followed by slice-based training) produced strong validation metrics on sliced data, these gains did not transfer to deployment on full-resolution images with sliding-window inference. Even the best Stage 1 SAHI configuration reached only 49.4% precision and 45.2% recall overall.

Extending Stage 1 with fine-tuning on resized full images proved substantially more effective. The best Stage 2 model (YOLO11s) operationally achieved 88.2% precision and 91.0% recall under single-pass full-image inference and outperformed all SAHI configurations across all ship sizes. In contrast, SAHI reduced recall to at most 57.9%, despite maintaining comparatively high precision, indicating that this gap cannot be closed by overlap and merging-threshold tuning alone.

The SAHI ablation further revealed distinct failure modes: Stage 1 models tend to overproduce detections, particularly for large ships, whereas Stage 2 models systematically underdetect when applied to tiles. Full-image inference was also substantially faster, with a speedup of at least 7.5× and up to ∼30× at higher overlap settings.

Overall, the most robust solution is the two-stage training regime with full-image inference. Future work will focus on improving tiled inference for cases where full-image processing is not feasible, including size-adaptive merging, better handling of tile-border instances, and generalization to cross-mission data.

References

[1] DLR Institute of Space Research, "Modular Aerial Camera System (MACS)," Feb. 2026. [Online]. Available: https://macs.dlr.de

[2] D. Hein and R. Berger, "Terrain aware image clipping for real-time aerial mapping," in *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, ser. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, B. Jutzi, M. Weinmann, and S. Hinz, Eds., vol. IV, no. 1. Copernicus Publications, Oktober 2018, pp. 61–68. [Online]. Available: https://elib.dlr.de/122068/

[3] J. Brauchle, S. Bayer, D. Hein, R. Berger, and S. Pless, "Macs-mar: a real-time remote sensing system for maritime security applications," *CEAS Space Journal*, pp. 1–10, April 2018. [Online]. Available: https://elib.dlr.de/121263/

[4] J. Brauchle, T. Bucher, D. Hein, R. Berger, M. Gessner, and K. Stebner, "Capabilities and applications of macs aerial camera systems for environmental research," in *EGU General Assembly 2021*, April 2021. [Online]. Available: https://elib.dlr.de/142914/

[5] D. Hein, T. Kraft, J. Brauchle, and R. Berger, "Integrated uav-based real-time mapping for security applications," *ISPRS International Journal of Geo-Information*, vol. 8, no. 5, pp. 1–16, Mai 2019. [Online]. Available: https://elib.dlr.de/127382/

[6] DLR Institute of Space Research, "MACS-Box," Feb. 2026. [Online]. Available: https://macs.dlr.de/box

[7] L. A. Varga, B. Kiefer, M. Messmer, and A. Zell, "SeaDronesSee: A maritime benchmark for detecting humans in open water," 2021. [Online]. Available: https://arxiv.org/abs/2105.01922

[8] CVAT.ai Corporation, "Open Data Annotation Platform," Feb. 2026. [Online]. Available: https://www.cvat.ai/

[9] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8," 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[10] G. Jocher and J. Qiu, "Ultralytics YOLO11," 2024. [Online]. Available: https://github.com/ultralytics/ultralytics

[11] N. Habash, A. A. Alqumsan, and T. Zhou, "Recent real-time aerial object detection approaches, performance, optimization, and efficient design trends for onboard performance: A survey," *Sensors*, vol. 25, no. 24, 2025. [Online]. Available: https://www.mdpi.com/1424-8220/25/24/7563

[12] F. C. Akyon, S. Onur Altinuc, and A. Temizel, "Slicing aided hyper inference and fine-tuning for small object detection," in *2022 IEEE International Conference on Image Processing (ICIP)*, 2022, pp. 966–970.