

Ocean Voyager: A Staged Thermal-to-Vision Perception Architecture for Autonomous Marine Debris Monitoring

Advaith Rajagopalan
BASIS Independent Silicon Valley
San Jose, California, USA
rajagopalan.advaith@gmail.com

Abstract—Marine plastics fragment into microplastics that are difficult to retrieve; effective mitigation therefore relies on locating and characterizing macro-debris before degradation. There is currently a gap between wide-area satellite remote sensing and fixed interceptors in the spatial scale at which intermediate-scale localization is required. We present Ocean Voyager, an autonomous surface vehicle (ASV) that is a catamaran built around a low-cost thermal-to-vision perception architecture, the central feature of which is that a servo-mounted thermal camera proposes candidate bearings for vision inference based on a surface thermal anomaly. This reduces the cost of vision inference and allows the use of Raspberry Pi-class hardware for long-term operations. We present the implementation of the platform and discuss our rationale for staging perception on embedded hardware. We demonstrate this perception architecture and the resulting reduced vision-inference duty cycle in a hardware-in-the-loop experiment in which the entire pipeline, from thermal sweep through to RGB confirmation and waste/non-waste classification, executes autonomously on the deployed platform’s embedded computer.

Index Terms—ASV, marine debris, thermal imaging, multi-camera sensing, embedded perception, staged sensing

I. INTRODUCTION

Plastics in the ocean fragment into microplastics under UV and mechanical degradation and are therefore extremely difficult to retrieve after dispersion in the water column [1]–[3]. Mitigation efforts therefore rely on the localization of floating macro-debris before degradation. This means that part of the marine plastic problem can be recast as a sensing and localization task: debris must be located while it still remains large enough to locate.

Efforts currently exist at two extremes of the spatial spectrum. Wide-area remote sensing can be done using satellites that can detect aggregated floating debris, but does not provide the meter-scale localization needed to intercept, and is limited in revisit and resolution [4]. Fixed-location interceptors such as river-mouth interceptors and offshore collection systems operate only at fixed locations. Intermediate-scale localization, on the order of kilometers to tens of kilometers, would require a mobile platform to localize dispersed debris and provide actionable coordinates [11]. Small ASVs are natural choices for this purpose due to their mobility, low unit cost, and ability to revisit dispersed targets.

However, intermediate-scale monitoring is required to be both persistent and low cost, and this requirement clashes with onboard perception. Continuous perception using modern convolutional detectors across multiple cameras would quickly exhaust the compute and energy budget of a low-cost ASV. The main architectural question this paper addresses is therefore: how can perception be arranged so that a resource-constrained ASV can conduct persistent search without doing continuous expensive vision inference?

Our answer is to stage the perception pipeline on a low-cost ASV using thermal sensing followed by vision inference. Specifically, a servo-mounted thermal camera first sweeps to propose candidate bearings where there is a surface thermal anomaly that suggests a possible floating object. It is then only at the proposed bearings that vision inference takes place: three front-mounted RGB cameras capture frames, an embedded detector classifies them, and multi-view voting confirms or rejects the candidates. Vision inference is gated by the thermal proposal rather than run continuously, allowing a marked reduction in vision-inference duty cycle and long-term operation using Raspberry Pi-class compute.

This paper makes the following contributions:

- a staged thermal-to-vision perception architecture on a low-cost ASV in which cheap thermal proposals gate expensive multi-camera vision inference (Section IV);
- a multi-view voting mechanism designed to suppress viewpoint-dependent false positives such as reflections and biological matter (Section IV-B);
- an analysis of duty-cycle reduction that estimates the advantage gained by staging at 20–60× over continuous scanning (Section IV-C); and
- a hardware-in-the-loop experimental validation of the entire thermal-to-vision pipeline operating on an actual ASV (Section V).

The platform kinematics, autonomy controller, detector evaluation, and resource budget are treated in companion papers [6]–[8]. Here we assume these are established components and focus on the perception architecture and its integration into an actual system.

II. RELATED WORK

Sensing of marine debris across scales. Remote sensing of floating plastic from satellite platforms has rapidly advanced, to the point where multispectral sensing can distinguish floating plastic from water under suitable conditions [4]. While powerful for wide-area surveys, these methods are limited by revisit cadence, pixel resolution relative to debris size, and atmospheric and sea-state conditions, and are incapable of providing the meter-scale localization needed for interception. Environmental characterization of the plastic journey—from land-based plastic entering the ocean, aggregating in convergence zones, and fragmenting [1]–[3], [11]—has driven the development of early interception at the intermediate scale.

Low-cost ASVs for environmental monitoring. Catamaran and other twin-hull ASVs have been widely used for water-quality and limnological surveys, from small autonomous catamarans for coastal monitoring [9] to low-cost, Raspberry Pi-based platforms aimed at reproducibility and citizen science [10]. These establish the catamaran-plus-electric-thruster configuration as a robust basis for surface sensing, but many of them target in-situ water-quality measurements and almost none consider the onboard-compute costs of continuous vision inference.

Staged and attention-driven perception. The idea of a cheap, broad cue gating expensive, narrow inference is prevalent in the literature on resource-efficient perception, where the low-cost proposal stage narrows the computation to a promising region of the image. Our contribution is in applying this approach cross-modally on an ASV: thermal sensing is the cheap proposal cue and RGB vision is the expensive confirmation stage, selected to fit the compute and energy budget of a low-cost ASV. The controller that consumes the confirmations and handles re-observation of ambiguous candidates is a separate contribution published elsewhere [7].

III. SYSTEM OVERVIEW

Ocean Voyager is a catamaran ASV constructed from commercial-off-the-shelf components as a low-cost monitoring platform. The twin-pontoon hull provides sensor stability: its wide base and low center of gravity reduce roll and sway compared to a monohull, maintaining the pointing stability of the bow-mounted RGB cameras and the validity of thermal bearings. Two Blue Robotics T200 thrusters, mounted one under each pontoon ahead of the stern, allow differential-thrust steering with thrust vectors near the center of mass.

The central electronics bay, a watertight enclosure between the pontoons, houses a Raspberry Pi-class single-board computer, three RGB cameras mounted at the bow, and an MLX90640 thermal sensor mounted on a servo allowing a full 360° panoramic sweep, which reverses at its limit to prevent cable wrap. One of the key features of the design is the explicit segregation of the high-current propulsion power domain and the low-current compute-and-sensor domain: the compute and sensors are powered from a regulated rail, preventing thruster-generated voltage transients from browning out the controller

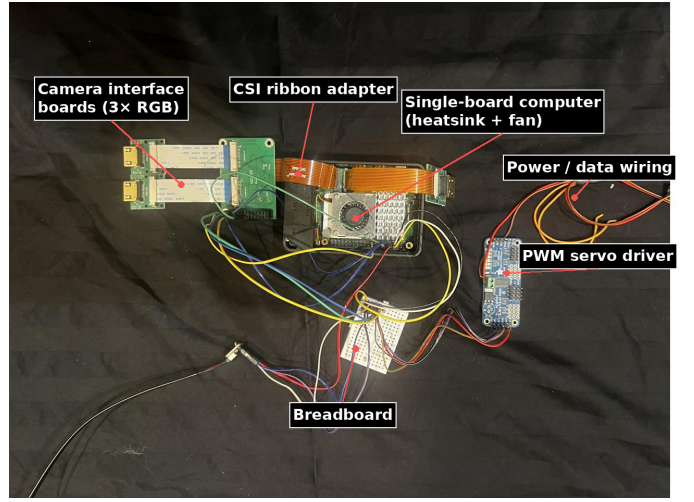


Fig. 1. Control-system and wiring layout of the assembled platform: the single-board computer with the CSI ribbon adapter and camera interface boards for the three RGB cameras, the PWM driver for the servo-mounted thermal camera, a prototyping breadboard, and the power/data harness. The compute-and-sensor supply is regulated separately from the high-current propulsion rail so that thruster transients do not disturb the controller or the sensors [8].

or injecting noise into the sensor power [8]. The power-domain segmentation, along with layered epoxy-and-silicone waterproofing and vibration isolation of the imagers from the hull, allows the perception subsystem to run reliably under propulsion load. Fig. 1 shows the control circuitry of the assembled electronics.

The ASV operates as an integrated pipeline. At each search station, the thermal camera performs the thermal sweep and proposes candidate bearings; the ASV approaches the bearing, captures RGB frames, and votes to confirm or reject; confirmed detections are logged with coordinates and search resumes. The finite-state-machine controller that executes this decision logic and its bounded-retry re-observation policy is the subject of a companion paper [7], and is discussed here only insofar as it explains how the perception stages interact with one another.

IV. PERCEPTION PIPELINE

The core architectural premise of our approach is that continuous full-frame inference over all cameras is infeasible on embedded marine hardware over long periods. Perception is therefore divided into two stages: a cheaper, continuous thermal proposal stage and an expensive, on-demand vision confirmation stage.

A. Stage 1: Thermal Anomaly Proposal

A floating object, with thermal properties different from those of the surrounding water, can present a surface-temperature difference when under thermal load from the environment. Floating debris is likely made of materials with differing emissivity or thermal mass compared to water, and the floating object and the surrounding surface do not necessarily have to be at the same temperature; depending on the

material and the environmental conditions, the object may be warmer or cooler than the water. The proposal stage exploits the temperature difference in either direction, not a fixed sign. The servo performs the thermal sweep by rotating through fixed increments and sampling the thermal field at the specified bearing. Candidate anomalies are selected according to their temperature difference and consistency and stored with their bearings. Importantly, no inference takes place in this stage: the thermal sweep is a cheap, broad proposal mechanism whose job is to indicate where the expensive vision stage should search. The thermal stage detects possible anomalies; it does not by itself establish whether a candidate is debris, and, as shown in the results, temperature difference alone is insufficient to select which anomalies are debris, so that determination is left to the vision stage.

B. Stage 2: RGB Confirmation with Multi-View Voting

As the ASV approaches a proposed bearing, the three RGB cameras capture frames and the embedded detector classifies them. The training, evaluation, and performance of the detector are characterized in a companion study [6]; here the detector is treated as an established component that feeds the voting rule with per-camera output.

A candidate is recognized as debris only if it is recognized by at least two of the three cameras. The motivation for voting is based on the geometry of the dominant failure mode: the most common false positives of the detector are viewpoint-dependent, such as biological-matter clusters and specular glare, which may look like debris from one camera but not from the other two. Since the cameras view from different bow angles, a false positive at one camera's viewpoint is less likely to occur at all three. Multi-view voting is aimed at suppressing viewpoint-dependent false positives in a way that cannot be done by post-processing a single-frame image, because the disambiguating information—the other two viewpoints—is not necessarily contained in a single frame. This mechanism is therefore a property of the system geometry, and thus a component of the perception pipeline rather than of detector tuning.

It is worth noting that a controlled ablation of two-of-three voting against single-camera recognition is not performed in this paper; the viewpoint-dependence of the underlying false positives is established in [6], and quantifying the suppression that voting achieves is left to future work.

There are two important caveats to this mechanism. First, the three viewpoints are correlated, so their errors are not statistically independent; a deceptive appearance that is visible from one bow camera is likely to also be visible from the adjacent one, and an independence assumption would overstate the effect of voting. Thus, voting is expected to reduce, but not eliminate, viewpoint-dependent false positives. Second, ambiguity—when the vote is split—poses a control problem: the ASV must decide whether to re-observe from another viewpoint or move on. That re-observation policy and its bounded-retry design are the topic of a companion controller paper [7] and are not claimed here.

C. Compute Argument

The key architectural advantage of staging is the significant decrease in the number of expensive vision-stage invocations. The argument is formulated here at the design level in terms of inference calls; the measured latency, total compute and power budget, and mission-phase scheduling are reported in the companion resource-architecture paper [8] and are not repeated here.

For a baseline of continuous inference at $f = 10$ frames/s on all three cameras, over a search cycle of T seconds,

$$N_{\text{cont}} = 3 f T \quad (1)$$

inference calls are performed—for T in the range 30 to 60 s, that is between 900 and 1800 calls per cycle. With staging, the detector fires only when a thermal anomaly is proposed. In a cycle, approximately one to three candidates are detected, each analyzed by the three cameras over about five approach frames, yielding

$$N_{\text{staged}} \approx 15\text{--}45 \quad (2)$$

inference calls per cycle. The reduction ratio is therefore $N_{\text{cont}}/N_{\text{staged}}$: with a conservative configuration (short $T = 30$ s cycle, dense $N_{\text{staged}} = 45$) this is $900/45 \approx 20\times$, whereas with a favorable configuration (long $T = 60$ s cycle, sparse $N_{\text{staged}} = 30$) it is $1800/30 = 60\times$. Equivalently, using the approximate 11 ms per-frame inference latency measured previously [6], [8], the vision-inference duty cycle decreases to around 0.5–1.5% of the ASV's time, compared to a continuous load that would take a substantial fraction of the embedded budget before any navigation, logging, or control overhead. This 20–60 $\times$ reduction forms the key architectural argument: it transforms an otherwise compute-bound perception workload into a regime that leaves room for other tasks, enabling long-duration, solar-assisted operation on Raspberry Pi-class hardware. The exact multiplier depends on scene density and sweep parameters, so a range is reported rather than a single value; confirmation of this effect under real mission conditions is the goal of future work.

V. SYSTEM INTEGRATION AND EXPERIMENTAL VALIDATION

The perception architecture has been implemented on a physical prototype of the ASV and experimentally validated using a hardware-in-the-loop (HIL) experiment to check that the thermal-to-vision pipeline works as intended. This experiment complements the dataset-based detector evaluation [6] and the software-in-the-loop controller validation [7], and is focused on the integration of the perception pipeline as an autonomous whole on the real compute and sensor hardware.

A. Setup

A physical prototype of the platform equipped with the MLX90640 thermal camera and the RGB cameras on the single-board computer was operated in an 8×8 ft test area designed to emulate the sensing task in terms of optics and thermal contrast. Four physical objects were placed in the

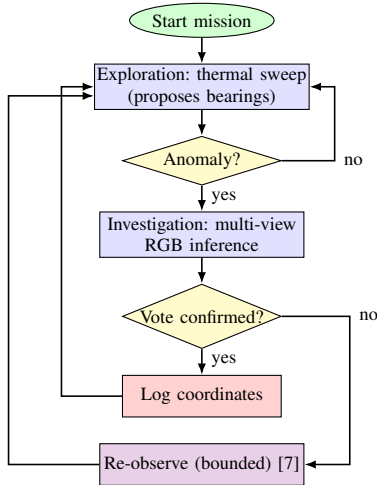


Fig. 2. Workflow of the staged pipeline: a continuous, cheap thermal sweep proposes candidate bearings; expensive multi-view RGB inference runs only on proposed candidates; ambiguous verifications are handled by a bounded re-observation policy described in the companion controller paper [7].

TABLE I
PER-OBJECT OUTCOMES, 8 × 8 FT HIL RUN

Object	Thermal ΔT (°C)	RGB conf.	Decision
Waste 1	5.25	0.79	confirmed
Waste 2	6.72	0.74	confirmed
Waste 3	7.72	0.88	confirmed
Non-waste	17.96	— (negative)	rejected

test area: three to emulate floating waste and one non-waste object acting as a deliberate distractor. The platform was manually repositioned along a serpentine coverage path—propulsion and closed-loop navigation were not tested in this run—while the entire perception and decision software stack ran autonomously on the onboard computer: thermal sweeping, candidate proposal, RGB acquisition, detector inference, the confirmation vote, and the resulting log/duplicate/boundary decisions were all performed on the deployed hardware. This isolates the perception pipeline and its integration for evaluation, while autonomous navigation—handled by the controller in the companion study [7]—is left to future on-water work. During the experiment, the platform visited 47 coverage-scan stations, at 8 of which the thermal stage proposed a candidate; the synchronized thermal, RGB, and decision streams were logged for analysis.

B. Results

All four placed objects were proposed by the thermal stage and reached a terminal decision. The three waste objects were confirmed by RGB verification, while the one non-waste object was correctly rejected; the vehicle also re-encountered already-logged objects four times, suppressing the duplicates rather than logging them again. Table I summarizes the per-object results.

The run clearly illustrates the separation of the two stages. The thermal stage executed its task as a proposal mechanism, detecting all four objects as of interest. However, thermal

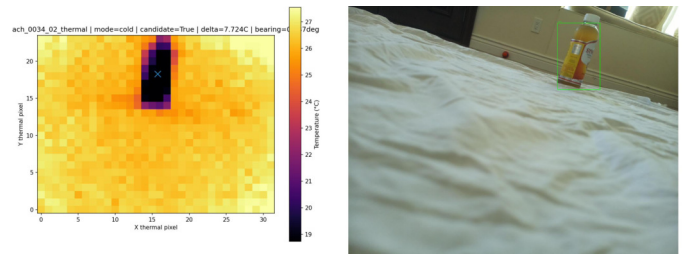


Fig. 3. Example output pair from the staged pipeline during the HIL run (event 34). Left: the MLX90640 thermal image, showing a cold anomaly against a warmer background ($\Delta T = 7.72^\circ\text{C}$); the indicated cell is the proposal. Right: the following RGB image, in which the detector confirms the object with a bounding box. The thermal stage selects a thermally contrasting region, whereas the vision stage decides whether it is waste.

magnitude alone failed to distinguish waste from non-waste: indeed, the non-waste distractor produced the largest thermal contrast in this run ($\Delta T = 17.96^\circ\text{C}$), above any of the three confirmed waste items ($5.25\text{--}7.72^\circ\text{C}$). Had target classification been based on thermal response, the non-waste object would have ranked first. Instead, the RGB confirmation stage rejected it: the detector returned negative on the approach frames, aborting the candidate as non-waste. This illustrates the division of labor on which the architecture is based: thermal proposes broadly, while vision confirms selectively what thermal cannot. Fig. 3 shows an example thermal-and-RGB pair for a confirmed detection.

C. Observed Limitations

The major limitations seen in this run were related to navigation rather than sensing. The vehicle showed lane displacement while navigating through coverage, experienced duplicate re-encounters with already-logged objects, and faced boundary-handling events when a planned maneuver would have taken it out of the test area. These are coverage- and control-layer behaviors—the responsibility of the autonomy controller developed separately [7]—rather than limitations intrinsic to the thermal and vision stages, which did their job properly here. They are reported because they constitute the main limitation observed in this HIL run: the perception pipeline worked as intended, while robust closed-loop navigation is a different problem addressed in the companion work. The single controlled run reported here is a functional demonstration rather than a statistical one; quantifying detection, confirmation, and false-rejection rates across multiple runs under different wave, lighting, and clutter conditions is the next step.

VI. DISCUSSION AND FUTURE WORK

The results confirm the architectural claims of the paper: staging a cheap thermal proposal ahead of expensive multi-view vision enables the operation of a neural perception pipeline on a low-cost ASV without paying for expensive continuous inference. The physical HIL run demonstrated that the pipeline works end-to-end on the actual hardware, and the duty-cycle analysis explains how it is possible.

There are several limitations to this work. The duty-cycle reduction is an architectural estimate, and its exact value depends on scene density and sweep parameters; its confirmation under sustained missions remains future work. The thermal proposal mechanism relies on a surface-temperature contrast whose magnitude varies with material, weather, time of day, and water conditions; its reliability—and its false-alarm rate on warm objects, like the distractor in the HIL run—has to be evaluated in the field. Multi-view voting is designed to reduce but not eliminate viewpoint-dependent false positives, and the magnitude of that reduction has not been quantified. In the HIL run, navigation was the principal limitation, but robust on-water navigation and thermal reliability under field conditions both remain open problems.

Future work follows directly. The immediate priority is field validation of the system on open water, including quantification of end-to-end performance—detection and false-confirmation rates, time-to-localization, and thermal-trigger reliability under different environmental conditions—building on the controlled HIL demonstration reported here. Beyond single-platform operation, intermediate-scale monitoring is a multi-vehicle problem, since the relevant accumulation regions are larger than the coverage of a single platform [11]; coordinating a fleet of low-cost staged-perception ASVs is a logical next step. Energy-efficient operation and endurance extension enabled by the staged computation is another potential direction for future work.

VII. CONCLUSION

In this paper we presented the staged thermal-to-vision perception architecture for resource-constrained autonomous marine debris monitoring embodied by Ocean Voyager. The architecture uses a cheap 360° thermal sweep to propose candidate bearings and gates expensive RGB inference and multi-view voting behind those proposals, using system geometry to suppress viewpoint-dependent false positives and reducing the vision-inference duty cycle by an estimated 20–60× compared to continuous scanning. A physical hardware-in-the-loop demonstration confirmed that the whole pipeline runs end-to-end on the assembled platform, with the thermal stage proposing objects and the vision-plus-vote stage confirming them. Together with the companion papers on the detector [6], the controller [7], and the resource architecture [8], this work constitutes the perception-architecture piece of a low-cost, reproducible system for intermediate-scale marine debris monitoring.

REFERENCES

- [1] J. R. Jambeck *et al.*, “Plastic waste inputs from land into the ocean,” *Science*, vol. 347, no. 6223, pp. 768–771, Feb. 2015, doi: 10.1126/science.1260352.
- [2] L. Lebreton *et al.*, “Evidence that the Great Pacific Garbage Patch is rapidly accumulating plastic,” *Scientific Reports*, vol. 8, Art. no. 4666, Mar. 2018, doi: 10.1038/s41598-018-22939-w.
- [3] A. L. Andrady, “Microplastics in the marine environment,” *Marine Pollution Bulletin*, vol. 62, no. 8, pp. 1596–1605, Aug. 2011, doi: 10.1016/j.marpolbul.2011.05.030.
- [4] L. Biermann, D. Clewley, V. Martinez-Vicente, and K. Topouzelis, “Finding plastic patches in coastal waters using optical satellite data,” *Scientific Reports*, vol. 10, Art. no. 5364, Apr. 2020, doi: 10.1038/s41598-020-62298-z.
- [5] G. Jocher, J. Qiu, and A. Chaurasia, “Ultralytics YOLO,” version 8.0.0, software, Jan. 2023, doi: 10.5281/zenodo.7867787.
- [6] A. Rajagopalan, “Deployment-based hard-negative stress test of marine debris detection using water-dependent negative examples for resource-constrained ASVs,” in *Proc. 2026 IEEE Canadian Atlantic Ocean Symp. (CAOS)*, Halifax, NS, Canada, 2026, to be published.
- [7] A. Rajagopalan, “A two-loop finite-state-machine controller with bounded-retry verification for resource-constrained autonomous marine debris investigation,” in *Proc. 2026 Int. Conf. Control, Automation and Systems (ICCAS)*, 2026, to be published.
- [8] A. Rajagopalan, “Resource-constrained edge computing architecture for multi-sensor autonomous marine monitoring,” in *Proc. 2026 IEEE Sensors and Systems Technology (SST)*, 2026, to be published.
- [9] G. Ferri, A. Manzi, F. Fornai, F. Ciuchi, and C. Laschi, “The HydroNet ASV, a small-sized autonomous catamaran for real-time monitoring of water quality: From design to missions at sea,” *IEEE J. Ocean. Eng.*, vol. 40, no. 3, pp. 710–726, Jul. 2015, doi: 10.1109/JOE.2014.2359361.
- [10] J. H. Ryu, “Prototyping a low-cost open-source autonomous unmanned surface vehicle for real-time water quality monitoring and visualization,” *HardwareX*, vol. 12, Art. no. e00369, Oct. 2022, doi: 10.1016/j.ohx.2022.e00369.
- [11] A. Rajagopalan, “Environmental contamination pathways and the intermediate-scale monitoring gap for marine plastic debris,” in *Proc. 2026 ACEER*, Springer LNCE, 2026, to be published.