Perception, Mapping and Autonomous Guidance

Atılım UAV Team · Atılım University — Software Team Technical Documentation · SUAS 2026

ABSTRACT

The Atılım UAV perception and mapping stack supports the 2026 Student Unmanned Aerial Systems (SUAS) Search–Detect–Deliver and Risk Mapping tasks. A SIYI A8 Mini gimbal camera supplies stabilized nadir imagery to a Jetson Orin NX companion computer, where a YOLO11m detector processes the most recent RTSP frame to avoid queue latency. The detector is trained on 16,015 merged and deduplicated images and achieves 0.950 mAP50 and 0.745 mAP50–95 on a validation set containing 31 real aerial images and 121 labeled instances. At the 50 m mission altitude, a 3x digital zoom setting reduces the ground sampling distance to 14.8 mm/px for detection and terminal guidance, while Risk Mapping is flown at unity zoom and processed by the in-house FastMosaic pipeline. Field tests detected human stand-ins at 49 m AGL with confidence values between 0.60 and 0.80. End-to-end missions at approximately 20 m and 50 m placed both payloads inside the 50 ft scoring radius using stored target geolocations. The image-based visual-servoing controller is implemented, while full closed-loop release validation on the live image remains the next flight-test milestone.

1 DESIGN CHOICES

This document describes the imaging hardware, object detection pipeline, target geolocation method, image-based terminal-guidance controller, Risk Mapping pipeline, simulation environment, and field validation used by the Atılım UAV Team.

Two competition tasks drive the architecture. Search–Detect–Deliver requires the aircraft to locate a mannequin and a tent, calculate their coordinates, approach the selected target, and command the correct payload delivery. Risk Mapping requires a single stitched image of the Search Boundary to be generated on site in the submission-ready image format. These requirements determined the camera, mission altitude, zoom settings, onboard compute, and processing pipeline.

2 IMAGING HARDWARE AND GROUND SAMPLING DISTANCE

The SIYI A8 Mini was selected because its three-axis gimbal holds the camera at nadir and absorbs aircraft motion during survey and target-approach maneuvers. The camera provides an Ethernet RTSP stream that can be processed onboard and monitored through the SIYI HM30 digital link.

Figure 1. The SIYI A8 Mini three-axis gimbal camera used for stabilized nadir imaging.

2.1 Ground Sampling Distance at Mission Altitude

For a frame of width N_h pixels at altitude h above the target plane, the ground footprint and ground sampling distance are

$$W_{\text{ground}} = 2h \tan\left(\frac{\text{HFOV}}{2}\right), \qquad \text{GSD} = \frac{W_{\text{ground}}}{N_h}. \tag{1}$$

The mission is flown at 50 m (164 ft) AGL, which keeps the aircraft above the competition's 150 ft AGL minimum altitude fence. With the 81° horizontal field of view of the SIYI A8 Mini at unity zoom and a horizontal frame width of 1920 px, Equation 1 gives 44.5 mm/px at that altitude, which is too coarse for reliable detection of a supine mannequin. Cropping the 4K sensor readout at 3x narrows the field of view to 31.8°, raises the effective focal length to $f = 3372$ px, and reduces the ground sampling distance to 14.8 mm/px. At the selected zoom level, the mannequin occupies sufficient image area for reliable detection after resizing to the inference resolution. The target's short-axis footprint is the limiting factor and therefore determines the operational zoom setting at the mission altitude.

The 3x setting is used for target detection and terminal guidance. The mapping survey pass is flown at unity zoom, where the wider field of view covers approximately 85 m of ground per frame at 50 m and keeps the full-boundary survey to approximately 15–20 frames.

2.2 Compute, Navigation, and Video Link

Object detection is performed onboard by a Jetson Orin NX. Live video from the SIYI A8 Mini is received through RTSP, and only the most recent frame is processed to eliminate queue latency. The SIYI HM30 carries the MAVLink stream and HD video, allowing the operator to monitor the same mission while onboard processing continues. The Cube Orange Plus supplies aircraft attitude and heading, while the CUAV Neo3 Pro Global Navigation Satellite System (GNSS) receiver supplies the position used by the geolocation pipeline.

Table 1. Perception hardware chain and the role of each component.

Component	Selection	Role
Camera and gimbal	SIYI A8 Mini	Stabilized nadir RTSP imagery
Video and telemetry link	SIYI HM30	Live operator video and MAVLink transport
Onboard compute	Jetson Orin NX	YOLO11m inference, geolocation, and guidance logic
Flight controller	Cube Orange Plus	Attitude, heading, navigation, and velocity setpoints
GNSS receiver	CUAV Neo3 Pro	Aircraft position for target geolocation
Ground Control Station	Mission Planner	Mission upload, telemetry, and operator supervision

3 OBJECT DETECTION

YOLO11m was selected because development experiments showed that dataset diversity, rather than network capacity, was the main performance constraint. Smaller variants gave up accuracy needed for the two-class problem, while larger variants consumed onboard throughput without sufficient operational benefit. A P2 detection head and a dual-model ensemble were also evaluated and rejected because their added computational cost was not justified.

3.1 Dataset Generation

The dataset contains 16,015 images collected from multiple Roboflow projects. Duplicate images were removed using filename and perceptual-hash comparison, and fifteen source classes were consolidated into the two competition classes: mannequin and tent. Vehicle-related classes were discarded, and the class imbalance was reduced through oversampling of tent images. Training examples include aerial scenes, debris and grass backgrounds, hard-negative frames, and partially occluded targets.

3.2 Training and Operational Inference

Training data are generated using overlapping 1280×1280 image tiles and hard-negative background samples from debris and grass environments. YOLO11m is initialized from COCO- pretrained weights and trained at a 1280 px input resolution with scaling, flipping, rotation, mosaic augmentation, and mixup.
The operational flight configuration performs full-frame inference at 640 px with a 0.25 confidence threshold. The threshold is intentionally low because release logic depends on persistence across successive frames rather than a single detection. A recurring moderate-confidence detection is more valuable than an isolated high-confidence result, and the temporal requirement filters false positives. TensorRT-accelerated tiled inference remains available as a contingency configuration for smaller or lower-confidence targets.

3.3 Validation Results

Validation was performed exclusively on real aerial imagery captured using a separate survey drone. The validation set contains 31 images and 121 labeled instances. The model achieved an overall mAP50 of 0.950

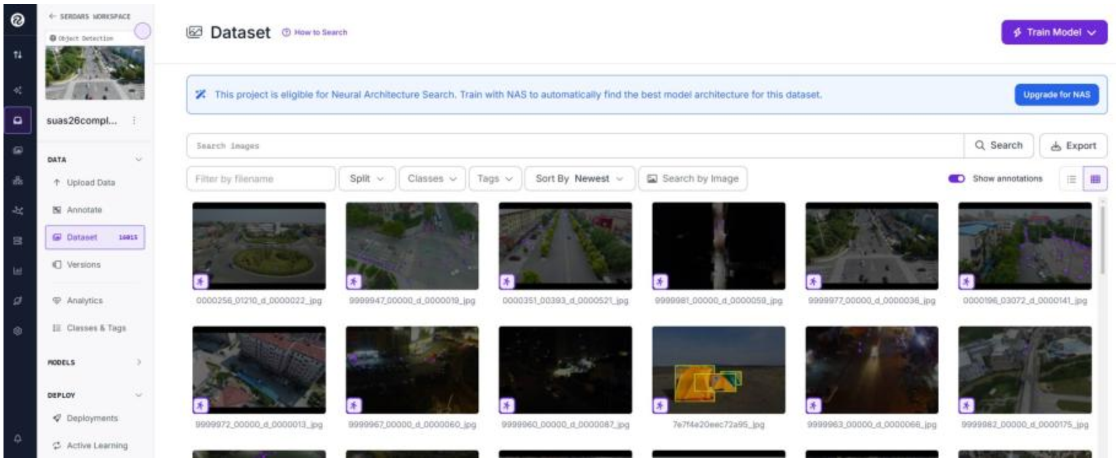

Figure 2. The Roboflow dataset containing 16,015 images across the mannequin and tent classes.

and mAP50–95 of 0.745. Precision and recall are defined as $P = TP/(TP + FP)$ and $R = TP/(TP + FN)$.

Table 2. Validation performance of the retrained YOLO11m model by class.

Class	P	R	mAP50	Val. instances
Mannequin	0.902	0.801	0.906	115
Tent	0.942	1.000	0.995	6
All	—	—	0.950	121

The tent class includes only six validation instances, so the result is reported without claiming statistical confidence beyond the available sample size. Tiled inference was tested on two held-out aerial images for which full-frame inference produced no detections. It recovered five detections on one image and one detection on the other, with a peak confidence of 0.74. However, onboard processing performance fell short of the rate required by autonomous guidance, so full-frame inference remains the flight configuration and tiling is retained only as a contingency.

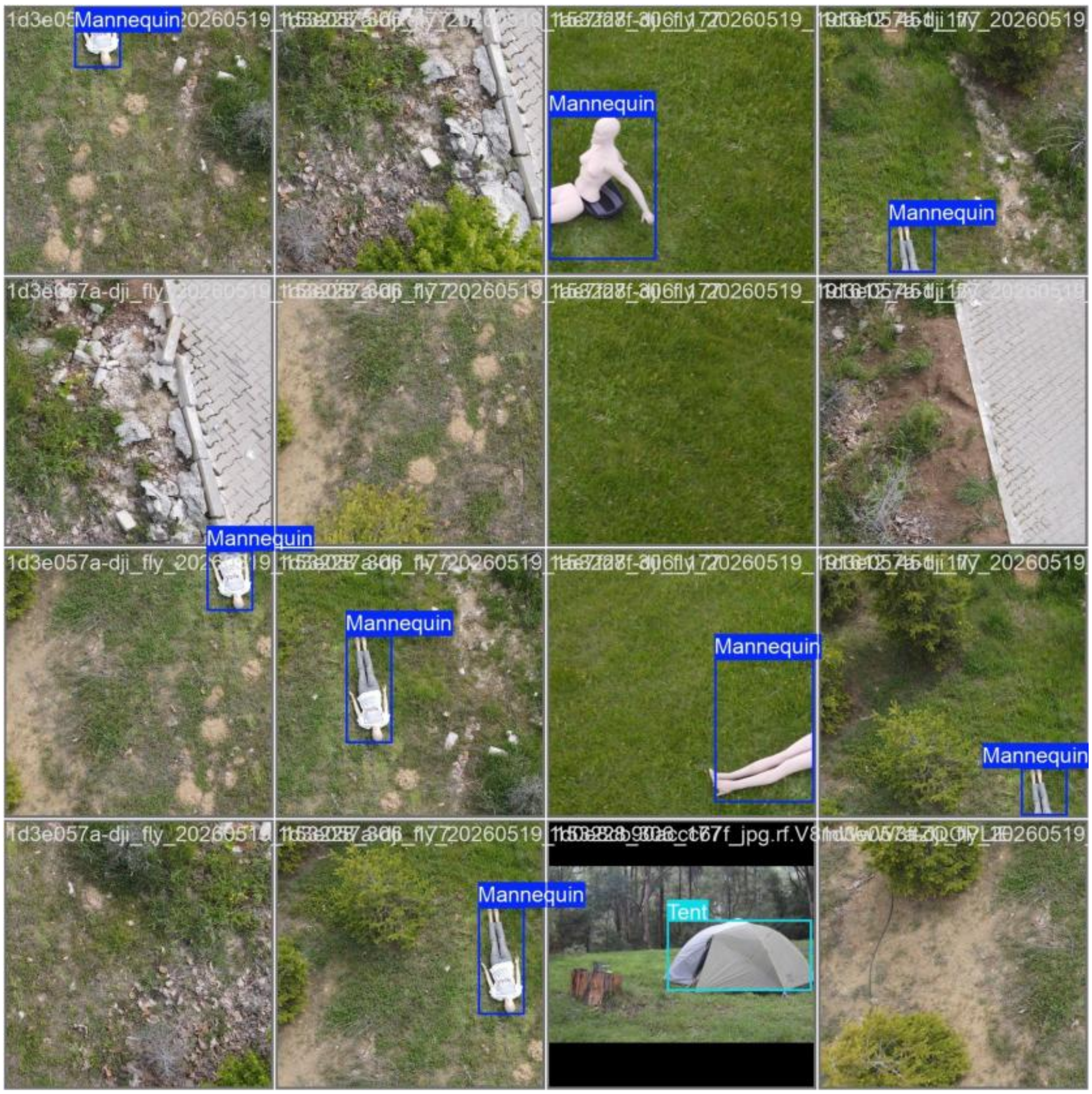

Figure 3. Labeled mannequin and tent instances in real aerial imagery, including hard-negative backgrounds and partial-occlusion cases.

Figure 4. Representative YOLO11m validation predictions across exposed, partially framed, and difficult aerial targets.

4 RISK MAPPING

The mapping subsystem converts imagery collected during the autonomous waypoint tour into one stitched map of the Search Boundary. The map is generated on site immediately after the survey and written directly in the submission-ready PNG format. FastMosaic is optimized for rapid field turnaround, a small survey set of approximately 15–20 frames, and visually consistent seams.

FastMosaic operates in four stages. GPS-assisted frame selection orders frames using the flight log and rejects frames captured during heading changes greater than 30°. The remaining frames are matched against a sliding buffer using ORB features with CLAHE contrast normalization. A 4-degree-of-freedom similarity transform is estimated by RANSAC,

$$\mathbf{x}' = s\,\mathbf{R}(\theta)\,\mathbf{x} + \mathbf{t}. \tag{2}$$

Distance-transform feathering blends each frame into the canvas, which is cropped and written as the final PNG. GPS supplies selection and sequencing, while visual features provide pixel-level placement.

FastMosaic was developed in house rather than using an external photogrammetry suite. This keeps the mapping chain inside the same Python toolchain as the rest of the mission software, removes an external field-processing dependency, and supports a real-time preview from the A8 Mini RTSP stream.

Figure 5. An early FastMosaic stitching test over a scale model.

The result is a visual mosaic rather than a rigorously orthorectified product. Similarity transforms do not model terrain relief or full perspective effects, and incremental placement can accumulate drift. Quality is therefore controlled through nadir imagery, overlap-rich survey paths, maneuver filtering, and visual comparison against reference imagery. Camera intrinsics currently use nominal datasheet-derived values, and lens distortion remains uncorrected.

Mapping was verified at component and mission levels. RTSP ingestion was tested before flight, frames captured during sharp heading changes were filtered automatically, and frames that failed feature matching were discarded rather than forced into the final mosaic. Acceptance criteria included complete area coverage, seamless image alignment, consistent exposure, and visual agreement with reference imagery.

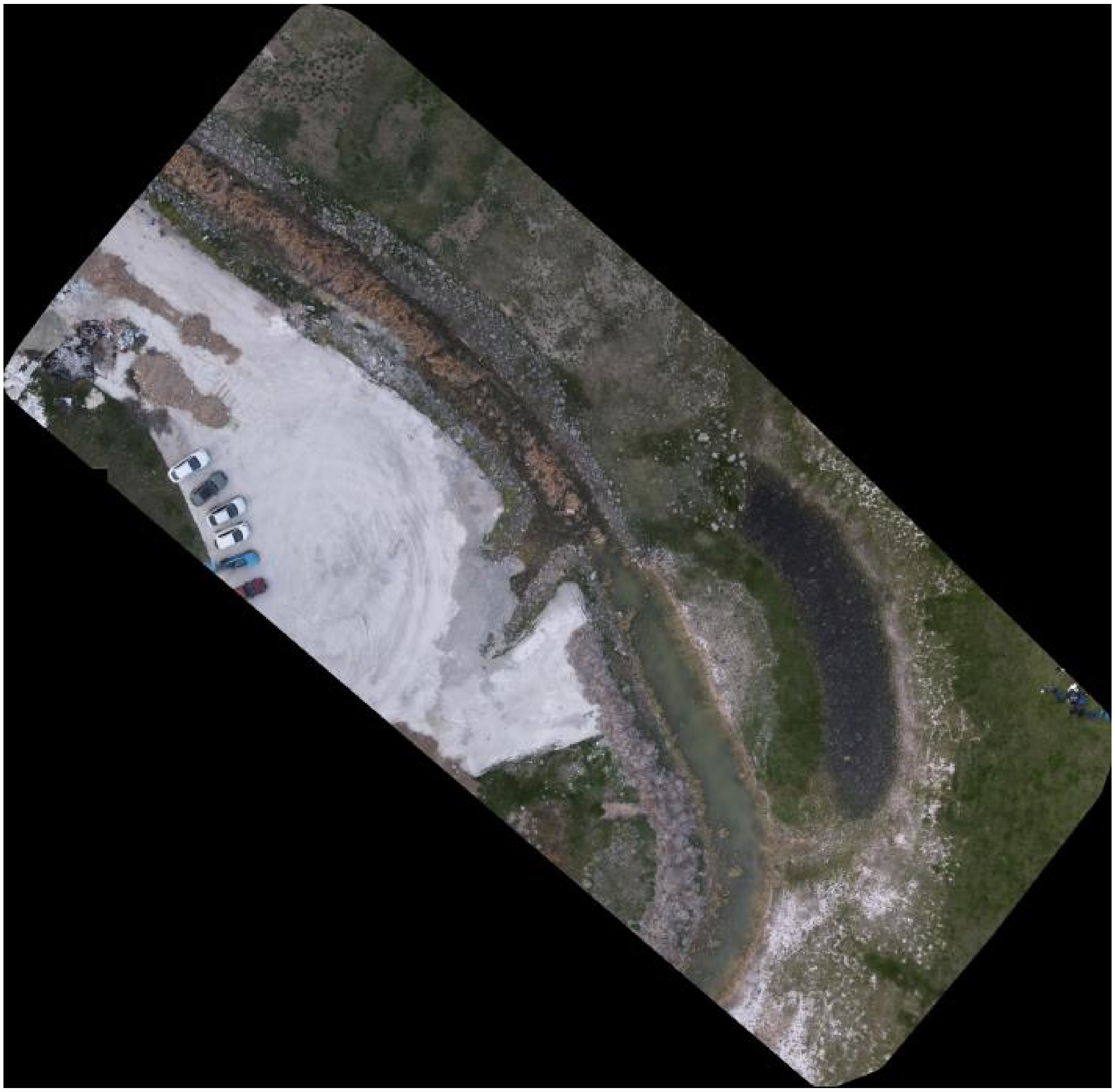

Figure 6. A FastMosaic stitching test over the flight-test field using natural ground textures.

5 SIMULATION-BASED TESTING

Every new capability is exercised in simulation before hardware flight testing. The environment combines Gazebo, ArduPilot SITL, ROS 2, pymavlink, and Mission Planner. The flight-dynamics simulator, GCS, and Jetson companion-computer roles operate as separate hosts on a shared Ethernet network, reproducing the communication paths used during flight. Camera imagery reaches the Jetson through RTSP so the complete image-processing chain can be tested under representative conditions.

SITL and Gazebo are used for autonomous takeoff, waypoint navigation, landing, and Return-to- Launch scenarios. Survey missions are generated with the Simple Grid tool. PID tuning and GPS-lock logic are validated in simulation using the actual UAV's weight and propulsion characteristics before hardware testing. Simulated targets with known positions and orientations allow detection, bounding-box alignment, processing latency, and geolocation decisions to be checked against ground truth. GPS/EKF degradation, communication loss, and low-battery scenarios are also introduced to verify LOITER, landing, and Return-to-Launch behavior.

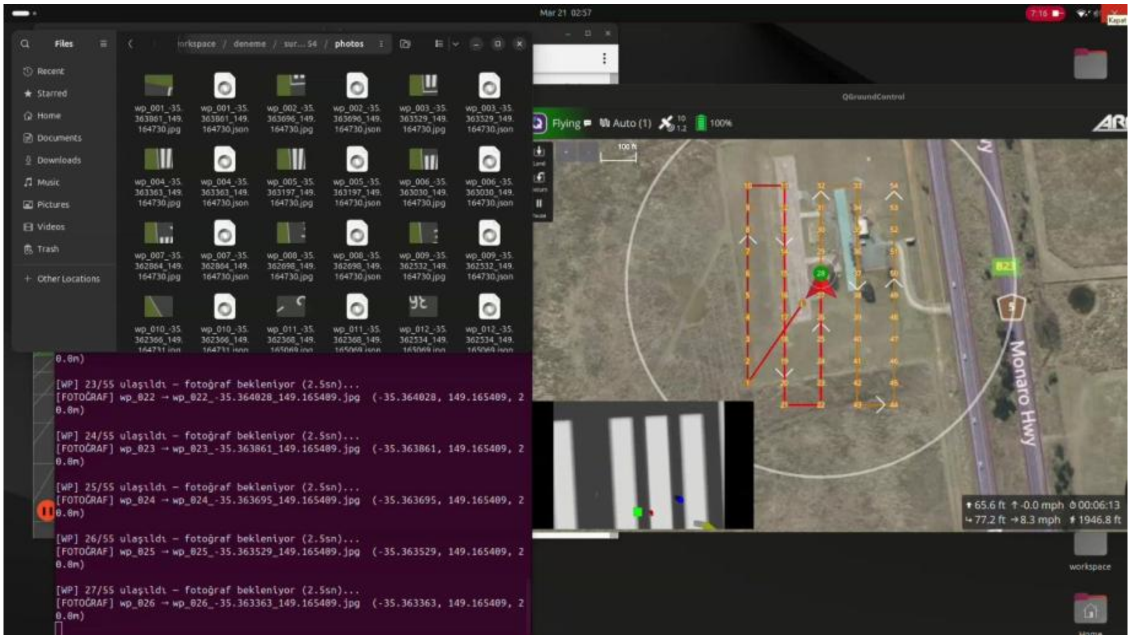

Figure 7. A Gazebo and ArduPilot SITL survey test used to verify waypoint execution, simulated image capture, and the networked mission architecture before hardware flight testing.

Latency-critical measurements are rechecked on the Jetson Orin NX because desktop-GPU inference and simulation real-time factor do not directly predict performance on the target hardware.

6 TARGET GEOLOCATION

A detector output is a bounding box in image coordinates, while the guidance system requires a target latitude and longitude. With the gimbal at nadir, the target pixel (u, v) in a $W \times H$ frame defines a ray in a locally level forward-right-down frame. The aircraft heading ψ rotates this ray into the local NED frame,

$$\begin{aligned}\mathbf{r}_c &= \left[-\frac{v - H/2}{H} 2\tan\left(\frac{\text{VFOV}}{2}\right), \frac{u - W/2}{W} 2\tan\left(\frac{\text{HFOV}}{2}\right), 1\right]^T, \\ \mathbf{r}_n &= \mathbf{R}(\psi)\,\mathbf{r}_c.\end{aligned} \tag{3}$$

Intersecting the ray with the ground plane at AGL altitude h yields the north and east offsets,

$$\begin{aligned}[\Delta N, \Delta E]^T &= \frac{h}{r_D}[r_N, r_E]^T, \\ \text{lat}_t &= \text{lat} + \frac{\Delta N}{R_E}, \\ \text{lon}_t &= \text{lon} + \frac{\Delta E}{R_E \cos(\text{lat})}.\end{aligned} \tag{4}$$

Equations 1 and 3 use the same pinhole model. The effective focal length is approximately 1124 px at unity zoom and 3372 px at 3x zoom. Detection sizing and geolocation therefore use the same zoom-matched camera parameters.

6.1 Image-Based Terminal Guidance

A stored coordinate carries the error of a single geolocation estimate, and the aircraft must still be centered over the target at release. After navigating to the stored coordinate, the image-based visual-servoing controller uses the target offset from the image center, $\mathbf{e} = (e_u, e_v) = (u - W/2, v - H/2)$, to generate horizontal velocity commands. The offset is smoothed with an exponential moving average using $\alpha = 0.15$. A discrete PID controller runs at 20 Hz and rotates body-frame commands into NED using the current yaw,

$$\begin{aligned}\mathbf{v}_b[k] &= K_p \bar{\mathbf{e}}_b[k] + K_i \Delta t \sum_{j=0}^{k} \bar{\mathbf{e}}_b[j] + K_d \left(\bar{\mathbf{e}}_b[k] - \bar{\mathbf{e}}_b[k-1] \right), \\ \bar{\mathbf{e}}_b &= [-\bar{e}_v, \bar{e}_u]^T, \\ [v_N, v_E]^T &= \mathbf{R}(\psi) \mathbf{v}_b.\end{aligned} \tag{5}$$

The gains are $K_p = 0.003$, $K_i = 0.0001$, and $K_d = 0.004$, command magnitude is limited to 0.8 m/s, and the integral term is clamped to prevent wind-up. Release is armed only after the target remains within 20 px of the full 1920 × 1080 frame center for a sustained period. At the 50 m mission altitude and 3x zoom, this corresponds to approximately 0.30 m on the ground. A sustained target loss aborts the approach and returns the aircraft to the waypoint mission.

The controller uses a fixed-depth approximation tuned for the release altitude. The 14.8 mm/px mission configuration is within 17% of the 17.8 mm/px scale used during tuning, allowing the gains to transfer without rescheduling. The field releases reported below were commanded from the stored geolocation alone; full closed-loop terminal-guidance validation remains pending.

Table 3. Detection-to-delivery chain and current validation status.

Stage	Input	Output	Status
Capture	A8 Mini RTSP frame	Frame and timestamp	Working
Detection	Resized frame	Class and bounding box	Field-validated
Geolocation	Box, pose, and camera model	Target latitude and longitude	Field-validated
Approach	Stored target coordinate	Autonomous navigation and release	Field-validated
Terminal guidance	Live image offset	Closed-loop velocity correction	Implemented; flight validation pending

7 FIELD VALIDATION

Field validation proceeded in three stages. First, detection robustness was evaluated under low-light conditions using human stand-ins on concrete surfaces. Across six trials, confidence values remained approximately 0.75 from distances of 20–25 m. Second, detection was verified at the mission altitude: with the aircraft at 49 m AGL and the camera at 3x zoom, a human stand-in was detected with confidence values between 0.60 and 0.80, above the 0.25 operational threshold.

Finally, end-to-end autonomous mission tests were performed. The first flight was conducted at approximately 20 m AGL to keep the release within visual range. Both competition targets were detected, their positions were transferred to the guidance system, and the aircraft navigated to the stored release coordinates. The strobing beacon landed on the tent target, while the water bottle landed approximately 5 m from the mannequin. The sequence was repeated at the 50 m mission altitude with comparable detection confidence and a delivery offset of the same order. Both tests placed the payloads within the 50 ft (15.2 m) scoring radius.

8 ONGOING WORK

The current next steps are:

- Validate the image-based terminal-guidance loop in a repeat flight at the 50 m mission altitude.
- Calibrate the SIYI A8 Mini intrinsics and lens distortion for both geolocation and mapping.
- Increase the tent-class validation set beyond the current six instances.
- Add coordinate-accurate georeferencing to the FastMosaic output while preserving rapid field turnaround.

REFERENCES

[1] Ultralytics, "YOLO11 Documentation." [Online]. Available: `https://docs.ultralytics.com`

[2] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *2011 International Conference on Computer Vision*, 2011, pp. 2564–2571.

[3] OpenDroneMap Authors, "ODM: A command line toolkit to generate maps, point clouds, 3D models and DEMs from drone, balloon or kite images," `https://github.com/OpenDroneMap/ODM`, 2020.

[4] MAVLink, "MAVLink Developer Guide." [Online]. Available: `https://mavlink.io`

[5] SIYI Technology, "A8 Mini User Manual," 2024.