IEEE Transactions on Field Robotics

Received 12 June 2025; revised 5 September 2025; accepted 6 October 2025.
Date of publication 27 October 2025; date of current version 13 November 2025.
This article was recommended by Executive Editor Timothy Barfoot.

*Digital Object Identifier 10.1109/TFR.2025.3625905*

# NOVA: Navigation via Object-Centric Visual Autonomy for High-Speed Target Tracking in Unstructured GPS-Denied Environments

**ALESSANDRO SAVIOLO[1] AND GIUSEPPE LOIANNO[2] (Member, IEEE)**

[1]Tandon School of Engineering, New York University, Brooklyn, NY 11201 USA
[2]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, CA 94720 USA
CORRESPONDING AUTHOR: ALESSANDRO SAVIOLO (alessandro.saviolo@nyu.edu)

This work was supported in part by NSF CAREER Award 2546659 and in part by Defense Advanced Research Project Agency (DARPA) Young Faculty Award (YFA) under Grant D22AP00156-00.

This article has supplementary downloadable material available at https://doi.org/10.1109/TFR.2025.3625905, provided by the authors.

*(Regular Article)*

**ABSTRACT** Autonomous aerial target tracking in unstructured and global position system (GPS)-denied environments remains a fundamental challenge in robotics. Many existing methods rely on motion capture systems, premapped scenes, or feature-based localization to ensure safety and control, limiting their deployment in real-world conditions. We introduce **NOVA**, a fully onboard, object-centric framework that enables robust target tracking and collision-aware navigation using only a stereo camera and an inertial measurement unit (IMU). Rather than constructing a global map or relying on absolute localization, NOVA formulates perception, estimation, and control entirely in the target's reference frame. A tightly integrated stack combines a lightweight object detector with stereo depth completion, followed by histogram-based filtering to infer robust target distances under occlusion and noise. These measurements feed a visual-inertial state estimator that recovers the full 6-DoF pose of the robot relative to the target. A nonlinear model predictive controller (NMPC) plans dynamically feasible trajectories in the target frame. To ensure safety, high-order control barrier functions (CBFs) are constructed online from a compact set of high-risk collision points extracted from depth, enabling real-time obstacle avoidance without maps or dense representations. We validate NOVA across challenging real-world scenarios, including urban mazes, forest trails, and repeated transitions through buildings with intermittent GPS loss and severe lighting changes that disrupt feature-based localization. Each experiment is repeated multiple times under similar conditions to assess resilience, showing consistent and reliable performance. NOVA achieves agile target following at speeds exceeding 50 km/h. These results show that high-speed, vision-based tracking is possible in the wild using only onboard sensing, with no reliance on external localization or assumptions on the environment structure.

**INDEX TERMS** Aerial robotics, control barrier functions (CBFs), depth completion, global position system (GPS)-denied environments, model predictive control, object detection, target tracking, vision-based navigation, visual-inertial odometry (VIO).

## I. INTRODUCTION

TARGET tracking is the ability of an autonomous system to detect, follow, and predict the motion of an object of interest over time. For unmanned aerial vehicles (UAVs), this capability is critical in dynamic and unstructured environments where premapped routes or fixed waypoints are insufficient. In such scenarios, robust tracking allows the UAV to remain persistently coupled to targets such as people, vehicles, or infrastructure, despite challenges like scene changes, occlusions, or unpredictable target motion.

This capability is key to diverse applications. In search and rescue, tracking allows UAVs to follow moving individuals through unstructured terrain, maintaining proximity without relying on maps or global position system (GPS) [1]. In human–robot interaction, it enables drones to interpret and respond to human motion in real time, facilitating

adaptive behavior [2], [3]. In inspection and surveillance, tracking ensures continuous observation of mobile assets, even under occlusion or changing viewpoints [4], [5], [6]. Moreover, tracking is foundational in multirobot systems, where relative positioning between agents is essential. UAVs in leader–follower formations or aerial swarms must track teammates to maintain coordinated motion [7], [8], [9]. In autonomous landing, a UAV must descend onto a moving platform, requiring precise tracking [10], [11], [12].

Visual tracking methods for UAVs traditionally fall into two categories: image-based visual servoing (IBVS) and position-based visual servoing (PBVS) [13], [14], [15], [16], [17]. IBVS operates directly on image-plane measurements, allowing low-latency control and robustness to calibration errors. However, its effectiveness is limited by the camera's field of view and sensitivity to target occlusion or rapid motion. PBVS instead reconstructs the target's 3-D position, typically in a globally referenced frame, and uses it to guide control. While PBVS mitigates many of IBVS's limitations, it introduces a new dependency: accurate consistent global localization.

In GPS-denied settings such as forests, urban mazes, indoor spaces, or extraterrestrial terrain, global localization typically relies on visual-inertial odometry (VIO) [18], [19], [20] or simultaneous localization and mapping (SLAM) [21], [22], [23], which estimate the robot's pose by tracking visual features across image sequences. These methods assume sufficient texture, consistent lighting, and a relatively static environment to maintain reliable feature correspondences. However, unstructured environments often violate these assumptions due to cluttered geometry, dynamic elements, ambiguous semantics, and degraded visibility. As a result, feature-based localization quickly becomes unreliable [24].

A notable illustration of these limitations is NASA's Ingenuity Mars Helicopter [25], [26], [27], which suffered a navigation failure while flying over smooth, featureless sand ripples [28]. Without distinctive landmarks, its visual-inertial system produced erroneous velocity estimates, leading to loss of control. This incident highlights a broader principle central to our work: in feature-sparse environments, localizing relative to a known, observable target (such as the perseverance rover in that case) can provide a more reliable reference frame than the surrounding terrain.

This article introduces **NOVA**, a unified framework for *navigation via object-centric visual autonomy*. NOVA enables real-time aerial target tracking and obstacle-aware navigation using only onboard sensing, with no dependence on GPS, external maps, or motion capture. Unlike traditional VIO or SLAM-based methods that require globally consistent localization, NOVA formulates perception, estimation, and control directly in the target's frame of reference. This object-centric approach enables tracking in environments where localization is unreliable or unavailable.

At the core of NOVA is a tightly integrated perception-to-control stack that couples a custom object detector, depth completion module, and visual-inertial state estimator. The object detector operates in real time on low-resolution inputs using an adaptive zoom strategy to maintain tracking at long ranges. Depth estimates are computed using a fusion of stereo and monocular cues, then processed via histogram-based filtering to produce robust target distance estimates, even under occlusion or noise. From this, a full 6-DoF state of the robot relative to the target is inferred in real time. Crucially, this perception system is designed to operate under realistic sensing conditions, such as degraded lighting, partial occlusions, and unstructured terrain.

These target-centric estimates are passed to a nonlinear model predictive controller (NMPC) that optimizes dynamically feasible trajectories directly in the target's reference frame. To ensure safe operation in cluttered environments, obstacle avoidance is enforced online using high-order control barrier functions (CBFs), derived from a compact set of high-risk collision points identified in the depth map. This eliminates the need for dense mapping or environmental priors, enabling real-time, map-free collision avoidance.

We validate NOVA across a suite of challenging real-world experiments, illustrated in Fig. 1. Each scenario targets a specific failure mode common to traditional visual tracking pipelines. In the forest trail mission, the UAV follows a moving target at speeds exceeding 50 km/h over more than 1 km, inducing severe motion blur and visual degradation from airborne dust. The building transition scenario involves abrupt lighting changes and full GPS loss as the target moves from an open parking lot into an enclosed space and back outdoors. In the height-offset trial, the UAV maintains a vertical height of 10 m, challenging feature matching due to reduced parallax and resolution.

Across all experiments, NOVA maintains stable target lock, plans feasible trajectories, and avoids obstacles in real time. To evaluate repeatability, each mission is conducted multiple times under similar conditions, with consistent performance observed throughout. All sensing and computation are performed fully onboard using only a stereo camera and an inertial measurement unit (IMU). Unlike prior methods that rely on motion capture, artificial landmarks, or prebuilt maps, NOVA operates autonomously in unstructured environments with no external infrastructure. To the best of our knowledge, this is the first demonstration of real-time, high-speed object-centric navigation that operates fully onboard and in-the-wild, without external infrastructure, explicit mapping, or dependency on structured priors.

## II. RELATED WORKS

### A. TRADITIONAL VISUAL SERVOING FOR TARGET TRACKING

Visual servoing provides a well-established framework for controlling robotic motion with respect to a target, typically categorized into IBVS and PBVS [13], [14], [15], [16], [17]. In IBVS, control commands are computed by minimizing the error between observed and desired image-space

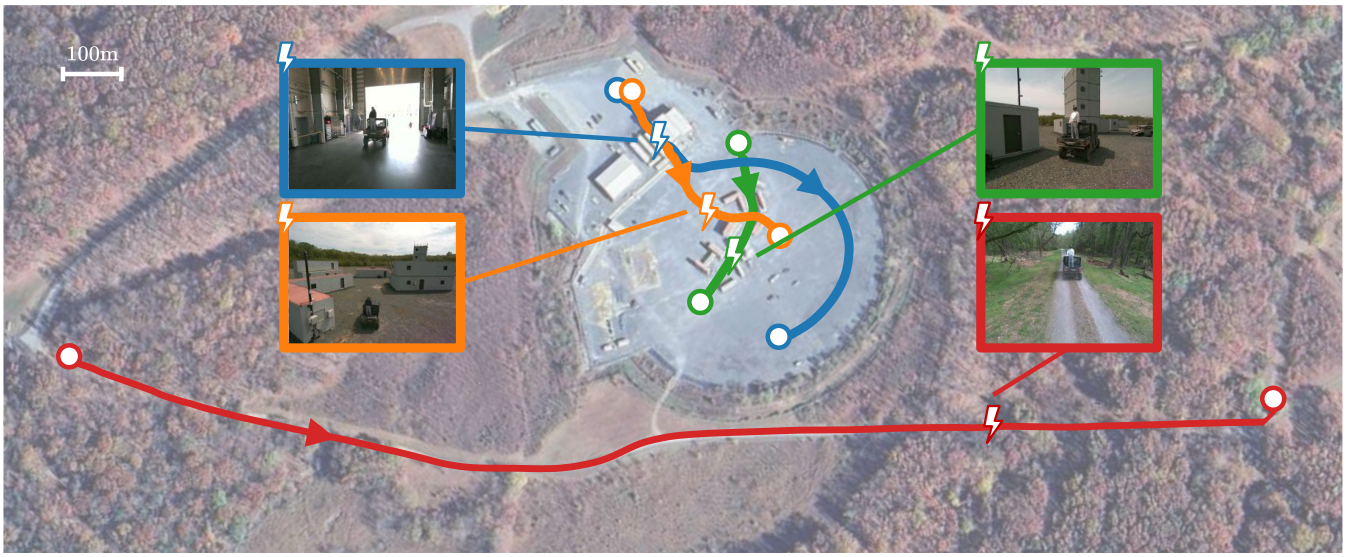



**FIGURE 1.** Satellite imagery of representative NOVA flight missions in real-world environments. Each overlaid UAV trajectory illustrates a distinct experimental scenario designed to test specific failure modes in visual-inertial target tracking and control. Red: High-speed tracking in a forest trail over 1 km, with target speeds exceeding 50 km/h and motion blur. Blue: Indoor–outdoor transition with complete GPS loss, exposure collapse, and perceptual ambiguity due to structural occlusions and distractors. Orange: Elevated tracking from a 10-m offset, challenging stereo depth perception, feature consistency, and planning geometry. Green: Urban container maze with narrow corridors, cluttered turns, and minimal texture. Insets show raw onboard RGB frames during each flight. NOVA maintains real-time target lock, depth estimation, obstacle avoidance, and closed-loop control using only a stereo camera and IMU. GPS is not used by NOVA and appears here only for visualization of flight paths.

features. These features are extracted from the camera view and mapped to robot motion using an interaction matrix that links image variation to movement. This method is efficient and effective for tasks where the target remains in view and image-space motion corresponds smoothly to inputs.

PBVS, on the other hand, estimates the full 3-D pose of the target using stereo vision, depth sensors, or geometric priors, and computes control actions in Euclidean space. It aims to minimize the spatial error between the current and desired poses of the robot relative to the target. This spatial reasoning improves robustness when the target leaves the field of view. However, PBVS depends on accurate global pose estimates of both the robot and the target, requiring consistent global localization across frames. In GPS-denied or visually degraded environments, this dependency often becomes a critical point of failure.

While both paradigms are effective in structured, static environments [29], [30], [31], [32], [33], their assumptions often break down in aerial target tracking. UAVs operate under rapid motion and wide viewpoint changes, which distort target appearance and invalidate interaction models. During aggressive maneuvers, UAVs experience frequent occlusions, motion blur, and target scale variation, causing visual features to become unreliable. IBVS fails when the target exits the frame, undergoes large appearance shifts, or when image-space motion no longer correlates predictably with control. PBVS struggles in GPS-denied environments or texture-sparse scenes where global pose estimation is not available or stable.

These challenges are further exacerbated by the limited computational resources available onboard UAVs and the stringent real-time requirements for flight control. In practice, neither classical IBVS nor PBVS alone can reliably handle the demands of high-speed, long-range target tracking in unstructured, dynamic environments.

### B. DATA-DRIVEN AERIAL TARGET TRACKING

Recent approaches in aerial target tracking have leveraged the power of deep learning to enhance perception and decision-making under challenging conditions. One prominent direction is the use of end-to-end learning, where a neural network directly maps raw sensory inputs to low-level control actions. These models, often trained via reinforcement learning or imitation learning, have demonstrated fast reaction times and adaptability to perceptual noise [34], [35], [36], [37]. However, this tight coupling of perception and control sacrifices interpretability, and poses significant challenges in generalization to unseen environments. Moreover, training such models requires extensive data collected in diverse, unstructured settings, an effort that is not only costly but also difficult to scale. Unstructured outdoor environments, in particular, exhibit extreme variability in appearance, geometry, and semantics, making it impractical to fully cover their distribution during training [38].

To address these limitations, a more modular paradigm has emerged where perception and control are treated as distinct but tightly integrated subsystems. In this setting, the perception module is responsible for detecting and localizing the

target, while the control module translates this information into safe and efficient motion commands. Extensive research has been devoted to each side of the pipeline. On the perception front, the computer vision community has produced a range of approaches for object detection and tracking, from compact architectures optimized for real-time inference on embedded platforms [39], [40], [41], [42], to large-scale foundation models capable of open-set recognition and zero-shot generalization [43], [44], [45], [46]. Techniques including temporal attention, depth prediction, and visual transformers have improved robustness to occlusion and visual degradation.

Alongside advances in perception, control design has emerged as a key focus in enabling robust aerial target following. While model-free approaches such as reinforcement learning have demonstrated flexibility and generalization [47], [48], [49], [50], they often struggle to provide consistent safety or performance guarantees, particularly under distribution shifts or in dynamic settings. In contrast, model-based methods, especially those based on NMPC [51], [52], [53], [54], offer a structured way to embed physical constraints, safety margins, and task-specific objectives directly into the control policy. By explicitly optimizing trajectories over a finite horizon, NMPC enables predictive and context-aware behaviors, which are critical for maintaining target visibility and avoiding obstacles during high-speed, reactive flight [55].

This separation of perception and control not only improves system modularity and interpretability, but also allows for leveraging the complementary strengths of data-driven learning and analytical planning. However, the efficacy of such hybrid systems depends critically on how well the two modules are coupled.

### C. OBJECT-CENTRIC VISUAL AUTONOMY

Traditional navigation frameworks rely on global localization through GPS, VIO, or SLAM, followed by waypoint-based trajectory generation [56], [57], [58]. While effective in structured environments, these methods often fail in GPS-denied or visually degraded settings, where feature tracking and map consistency degrade. Object-centric navigation offers a promising alternative: instead of referencing a global frame, control objectives are defined relative to specific scene objects, such as people or vehicles. This allows robots to maintain situational awareness and track targets even when global localization is unreliable or unavailable.

Several recent works have pursued this idea, though typically under restrictive assumptions that limit applicability in high-speed or unstructured environments.

An early example is [59], which formulates control directly in the target's frame of reference. While conceptually aligned with object-centric reasoning, the system relies entirely on motion capture to provide perfect state estimates for both robot and obstacle. No onboard sensing is used, and full target velocity is assumed known. This removes perception from the equation, making the setup impractical outside controlled environments.

More onboard-focused designs have emerged, such as [60], which uses a monocular camera and a blob detector to track a single spherical obstacle. The framework integrates perception and control onboard, but the simplicity of the sensing pipeline imposes its own constraints. Obstacle avoidance is implemented as a soft cost, tuned manually against the tracking objective, and the system assumes only one obstacle is visible at a time. The drone must be launched manually, and no mechanism is provided for adapting to clutter or dynamic changes. While a step forward, the system is difficult to generalize to real-world scenarios involving dense geometry or multiple obstacles.

Multiagent frameworks like CoNi-MPC [61] take a different approach, where a UAV follows a ground robot by optimizing a relative objective. This eliminates the need for global SLAM, but again depends on full-state telemetry from the target and is tested under motion capture. The absence of onboard perception and low operational speeds (under 0.4 m/s) further limits real-world relevance.

Later extensions add LiDAR-based obstacle sensing [62], improving situational awareness. However, core limitations persist. The system still relies on external tracking for the target and operates in structured, static environments at low speeds (below 1.5 m/s). Visual perception remains absent, and the setup assumes favorable sensing conditions.

Together, these efforts reflect growing interest in object-relative control, but also reveal a gap: current systems are not designed for fast, reactive tracking in complex environments using only onboard sensing. Our framework is built to address exactly that gap. By combining real-time object detection, depth-completed state estimation, and constraint-aware planning in a fully onboard pipeline, we enable agile, target-relative control in GPS-denied, cluttered environment, without reliance on external infrastructure or prior maps.

## III. METHODOLOGY

### A. SYSTEM OVERVIEW

NOVA is a fully onboard framework for vision-based aerial target tracking and collision avoidance in GPS-denied and visually degraded environments. It requires no maps, infrastructure, or prior tuning, enabling deployment in unknown and cluttered settings. The system is structured around two tightly integrated modules: perception and control (see Fig. 2).

The perception stack processes visual and inertial input to estimate the target's relative state and detect potential collision risks in real time. These estimates feed into an NMPC that dynamically tracks the target while enforcing safety through system dynamics and real-time obstacle constraints. The entire pipeline runs onboard, supporting reactive agile flight with no external dependency.

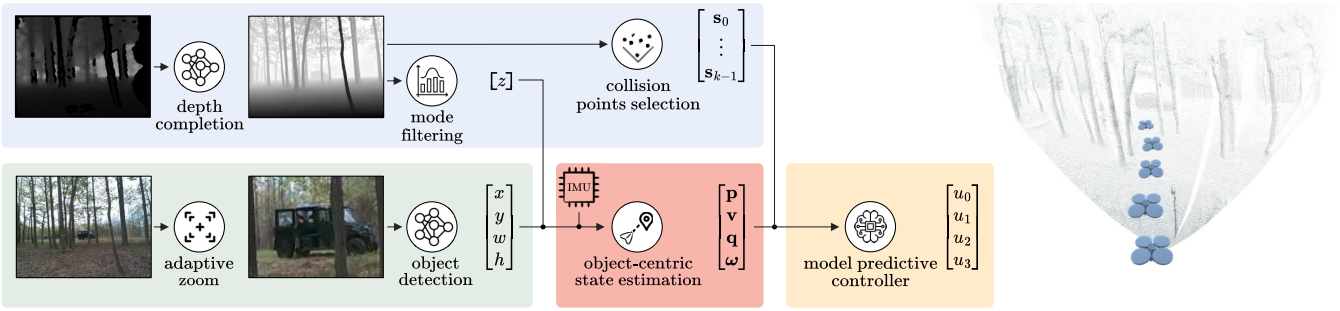



**FIGURE 2. System architecture overview. The framework consists of two tightly integrated modules: perception and control. The perception stack detects the target using a lightweight object detector, estimates its depth via stereo-monocular disparity alignment, and processes obstacle information for collision avoidance. These outputs are fused with inertial data (omitted from the figure for simplicity) to produce a smooth, high-frequency target-centric odometry. The control module uses NMPC, augmented with high-order CBFs, to generate safe and agile thrust and angular rate commands for tracking the target.**

### B. PERCEPTION

The perception pipeline estimates the target's position and near obstacles in real time using onboard visual and inertial sensors. At each frame, the system detects the target, estimates depth, and fuses this information with IMU data to maintain a full relative state. Simultaneously, it identifies high-risk collision points from completed depth maps. These estimates form the basis for downstream motion control.

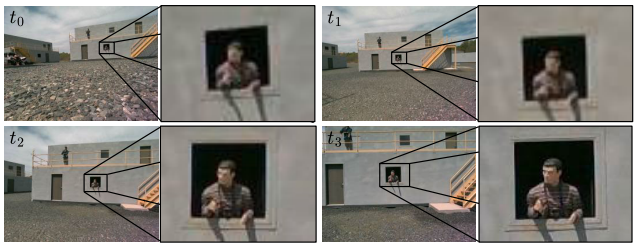



**FIGURE 3. Adaptive zoom strategy. Consecutive frames during a tracking sequence are shown. The UAV follows a target that emerges only partially from a window (raw images). For each frame, the system dynamically crops and rescales the region around the target (black bounding boxes, zoom-in views), which is then passed to the object detector. This adaptive zoom-in view maintains high detection confidence even when the target is small or partially occluded, by reducing irrelevant context and limiting false positives.**

#### 1) TARGET DETECTION

A central challenge in target tracking lies in balancing detection range with computational efficiency. High-resolution inputs, such as 640 × 480 pixels, support target detection at distances beyond 40 m but exceed the processing limits of embedded platforms. In contrast, lower resolution inputs like 320 × 240 pixels are more computationally efficient but constrain detection to shorter ranges, often under 10 m. To resolve this tradeoff, we introduce an adaptive zooming mechanism that dynamically crops and rescales the input image around a region of interest (see Fig. 3).

The zoom module takes the full-resolution RGB frame $\mathbf{I}^t \in \mathbb{R}^{H\times W\times 3}$ and a bounding box $\mathbf{b}^{t-1} = [x, y, w, h]$, where $(x, y)$ denotes the center coordinate and $(w, h)$ the width and height, from the previous detection. This box is enlarged by a factor $\alpha$ to introduce a margin that accounts for motion and detection uncertainty. A crop window is then centered on the enlarged box and shaped to match the detector's aspect ratio. The zoomed-in region is extracted from $\mathbf{I}^t$ and resized to the detector's input resolution, yielding $\mathbf{I}^t_{\text{crop}}$.

If detection fails at time $t$, the crop size is progressively increased to widen the search region. Once the target is reacquired, the zoom readjusts to tightly frame the detection. This mechanism preserves long-range detection capabilities at lower resolutions while improving robustness. By narrowing the detector's receptive field to a localized region, the system avoids distractions from background clutter and allows the network to focus on the target, increasing confidence and accuracy under occlusion and motion.

We employ a lightweight, custom-trained object detector based on the YOLOv11-small architecture to enable onboard real-time detection with limited computational resources. The input to the detector is the zoomed-in view crop $\mathbf{I}^t_{\text{crop}}$, and the output is a 2-D bounding box $\mathbf{b}^t$ in crop coordinates, which is then projected back into the full-resolution frame.

The object detector is trained from scratch on the COCO dataset [63] with domain-specific augmentations designed to mirror the visual effects introduced by our zooming strategy. In particular, randomized zooming and resizing simulate the pixelation and scale variations that occur during aggressive cropping, ensuring the detector remains effective even when targets appear coarse or low-resolution. Additional augmentations include pitch and roll rotations and lighting variations to improve robustness under aerial tracking conditions.

#### 2) DEPTH ESTIMATION

Accurate depth perception is essential for estimating the target distance as well as ensuring safe navigation. While stereo cameras provide absolute depth maps $\mathbf{D}^t_{\text{abs}} \in \mathbb{R}^{H\times W}$, these measurements are often sparse or noisy in areas with low texture, specular surfaces, or insufficient baseline [55]. To overcome these limitations, we employ a learning-based

depth completion approach that fuses monocular and stereo cues via a disparity-domain alignment (see Fig. 4).

We compute the absolute disparity from stereo depth as

$$\Delta^t_{\text{abs}} = \frac{f \cdot B}{\mathbf{D}^t_{\text{abs}} + \epsilon} \tag{1}$$

where $f$ represents the camera's focal length, $B$ is the stereo baseline, and $\epsilon$ is a small constant used for numerical stability. In parallel, a monocular depth estimation neural network $\mathcal{N}_{\text{mde}}(\mathbf{I}^t)$ [64] predicts a relative disparity map $\Delta^t_{\text{rel}}$.

To align scales, we fit a second-order polynomial between the two disparity maps using valid stereo pixels defined by a binary mask $\mathbf{M}^t \in \{0,1\}^{H\times W}$. Let $n$ be the number of valid pixels in $\mathbf{M}^t$. We construct a vector $\mathbf{y} \in \mathbb{R}^n$ containing the absolute disparities $\Delta^t_{\text{abs}}(i,j)$ at valid pixel locations. We also construct a design matrix $\mathbf{X} \in \mathbb{R}^{n\times 3}$, where each row $\mathbf{X}_k = [\Delta^t_{\text{rel}}(i_k,j_k)^2, \Delta^t_{\text{rel}}(i_k,j_k), 1]$. The vector of polynomial coefficients is denoted $\boldsymbol{\theta} = [a,b,c]^\top$.

The polynomial fit is defined as a least squares problem

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^3} |\mathbf{X}\boldsymbol{\theta} - \mathbf{y}|^2 \tag{2}$$

with the closed-form solution

$$\boldsymbol{\theta}^* = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}. \tag{3}$$

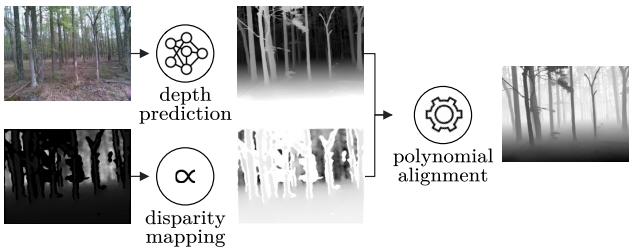



**FIGURE 4. Depth completion pipeline. Monocular (top) and stereo (bottom) disparity maps are fused to produce a dense, absolute-depth estimate. Monocular predictions capture relative structure, while stereo provides scale in textured regions. A polynomial alignment module merges the two in disparity space for accurate, absolute depth.**

The resulting polynomial parameters are applied across the full relative disparity map to generate a completed disparity estimate

$$\Delta^t_{\text{com}} = a\cdot(\Delta^t_{\text{rel}})^2 + b\cdot\Delta^t_{\text{rel}} + c \tag{4}$$

and converted back into a dense completed depth map

$$\mathbf{D}^t_{\text{com}} = \frac{f\cdot B}{\Delta^t_{\text{com}} + \epsilon}. \tag{5}$$

To estimate the target's depth, a naive approach might extract a single value from the center of the detected bounding box $\mathbf{b}^t$. However, this is prone to failure under partial occlusions. For instance, the target may remain visible and correctly detected, while the center pixel falls on a background object or an occluding surface, yielding an incorrect depth estimate.

To address this, we apply histogram-based mode filtering. We extract all valid depth values within $\mathbf{b}^t$ from $\mathbf{D}^t_{\text{com}}$, bin them into fixed-width depth intervals, and select the most frequent bin as the representative estimate, denoted $z$. This approach leverages the observation that the dominant mode in the depth distribution corresponds to the visible target. If this was not the case, the detector would likely fail to localize the target reliably. As illustrated in Fig. 5, this mode-based filtering improves robustness to noise, clutter, and occlusion without requiring pixel-level semantic segmentation.

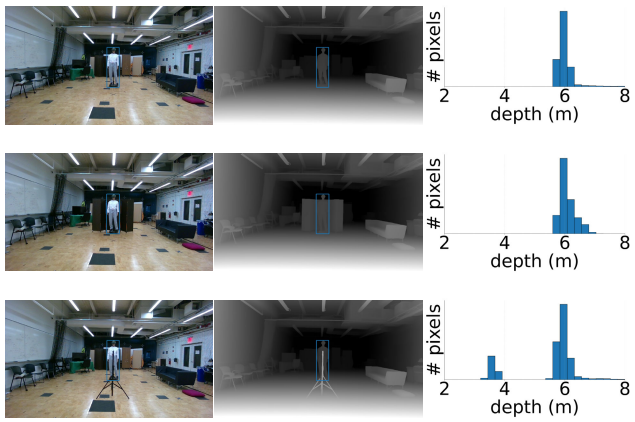



**FIGURE 5. Histogram-based mode filtering for target depth estimation. Depth values inside the detected bounding box are collected and binned. The most frequent bin is selected to estimate the target distance, improving robustness to noise, occlusion, and background clutter.**

### 3) OBJECT-CENTRIC STATE ESTIMATION

To enable smooth and reactive control, we estimate the full 6-DoF relative state of the quadrotor with respect to the target by fusing low-rate visual detections with high-rate inertial measurements. Each visual update provides the 2-D center coordinate $(x, y)$ of the detected bounding box and the target depth $z$ computed via histogram-based filtering.

As illustrated in Fig. 6, the detected 2-D center coordinate is back-projected into a 3-D point in the camera frame $\mathcal{C}$ using the known camera intrinsics $(f_x, f_y, c_x, c_y)$ obtained through camera calibration [65]

$$\mathbf{p}_{\mathcal{C}} = \begin{bmatrix} (x-c_x)z/f_x \\ (y-c_y)z/f_y \\ z \end{bmatrix}. \tag{6}$$

The resulting 3-D point is transformed into the body frame $\mathcal{B}$ using the static extrinsics between the camera and IMU

$$\mathbf{p}_{\mathcal{B}} = \mathbf{q}_{\mathcal{BC}} \odot \mathbf{p}_{\mathcal{C}} + \mathbf{t}_{\mathcal{BC}} \tag{7}$$

where $\mathbf{q}_{\mathcal{BC}}$ is the camera-to-body rotation (as a quaternion), $\mathbf{t}_{\mathcal{BC}}$ is the translation, and $\odot$ is the quaternion-vector product.

We define a moving target-centric frame $\mathcal{T}$, centered on the target and aligned with the ground plane. Although this frame may translate over time, we assume it remains parallel to the ground, and we model its orientation using the IMU-derived

attitude of the robot $\mathbf{q}_{\mathcal{TB}}$. The relative position of the robot in this target frame is given by

$$\mathbf{p}_{\mathcal{T}} = -(\mathbf{q}_{\mathcal{TB}} \odot \mathbf{p}_{\mathcal{B}}). \tag{8}$$

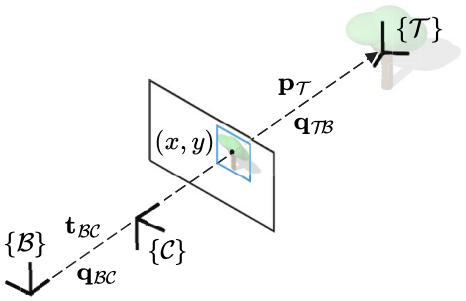



**FIGURE 6. Coordinate frame convention. The object-centric frame $\mathcal{T}$ is defined at the target location, with its orientation fixed and aligned with gravity using the IMU. A 2-D detection $(x, y)$ in the camera frame $\mathcal{C}$ is back-projected into 3-D and transformed into the body frame $\mathcal{B}$ using known extrinsics $(\mathbf{q}_{\mathcal{BC}}, \mathbf{t}_{\mathcal{BC}})$. The resulting position is then rotated into the target frame $\mathcal{T}$ using the IMU-derived attitude $\mathbf{q}_{\mathcal{TB}}$. This defines the relative pose $\mathbf{p}_{\mathcal{T}}$, which serves as the reference for target-centric perception and control.**

The resulting vector $\mathbf{p}_{\mathcal{T}}$ expresses the quadrotor's position relative to the target. This signal is tracked over time using an unscented Kalman filter (UKF) [66]. The UKF incorporates only the IMU's angular velocity as process input, while visual measurements provide asynchronous updates. Linear acceleration is intentionally excluded, since in the target-centric formulation it would implicitly assume that the target is nonaccelerating; fusing such measurements across frames would introduce inconsistencies and degrade filter stability. The full target-centric state consists of position $\mathbf{p}_{\mathcal{T}}$, velocity $\mathbf{v}_{\mathcal{T}}$, orientation $\mathbf{q}_{\mathcal{TB}}$, and angular velocity $\boldsymbol{\omega}_{\mathcal{B}}$.

This approach eliminates the need for any environment-specific information, such as visual texture or persistent features, for state estimation. Unlike traditional VIO or SLAM systems, our method relies solely on the detection and tracking of the target to estimate the full relative state of the quadrotor. This enables robust performance even in textureless, dynamic, and unknown environments, making it especially well-suited for agile target tracking.

#### 4) HIGH-RISK COLLISION POINTS SELECTION

To support safe target tracking in cluttered environments, we extract a sparse set of high-risk obstacle points from the completed depth map $\mathbf{D}^t_{\text{com}}$ for use in downstream model predictive control. These points are selected based on a time-to-collision (TTC) criterion computed with respect to the robot's instantaneous velocity $\mathbf{v}_{\mathcal{T}}$.

We first generate a TTC map $\mathbf{T}^t \in \mathbb{R}^{H\times W}$ by projecting robot's velocity vector $\mathbf{v}^t_{\mathcal{C}}$, expressed in the camera frame, along the viewing ray direction of each pixel. These ray directions $\mathbf{R}^t(i,j)$ are unit-length vectors pointing from the camera center through each pixel $(i,j)$, and are computed from the intrinsic calibration parameters. The TTC map is then computed as

$$\mathbf{T}^t(i,j) = \frac{\mathbf{D}^t_{\text{com}}(i,j)}{\left\|\mathbf{v}^t_{\mathcal{C}} \cdot \mathbf{R}^t(i,j)\right\|}. \tag{9}$$

To reduce the dimensionality of $\mathbf{T}^t$, we divide it into nonoverlapping grid cells $\mathcal{P}_{u,v}$ of size $P \times P$, and select the pixel in each cell with the minimum TTC

$$\mathcal{S}^t_{\text{ttc}} = \left\{ (i,j) \in \mathbf{T}^t \mid (i,j) = \underset{(i,j)\in\mathcal{P}_{u,v}}{\arg\min}\, \mathbf{T}^t(i,j)\ \forall\, u, v \right\}. \tag{10}$$

We then apply a filtering step to retain only those points likely to intersect the quadrotor's projected dimensions at their corresponding depths

$$\mathcal{S}^t_{\text{filt}} = \left\{ (i,j) \in \mathcal{S}^t_{\text{ttc}} \;\middle|\; \begin{matrix} |i - c_x| \le Q_x f_x / [2\mathbf{D}^t_{\text{com}}(i,j)] \\ |j - c_y| \le Q_y f_y / [2\mathbf{D}^t_{\text{com}}(i,j)] \end{matrix} \right\} \tag{11}$$

where $Q_x$ and $Q_y$ represent the robot's width and height.

From this filtered set, we extract the top-$K$ most critical collision points

$$\mathcal{S}^t_{\text{top-}K} = \underset{\mathcal{S}'^t_{\text{filt}}\subseteq\mathcal{S}^t_{\text{filt}}, |\mathcal{S}'^t_{\text{filt}}|=K}{\arg\min} \sum_{(i,j)\in\mathcal{S}'^t_{\text{filt}}} \mathbf{T}^t(i,j). \tag{12}$$

The selected pixel coordinates are back-projected into 3-D points in the camera frame $\mathcal{C}$

$$\mathcal{S}^t_{\mathcal{C}} = \left\{ \left( \frac{(i-c_x)z}{f_x}, \frac{(j-c_y)z}{f_y}, z \right) \;\middle|\; \begin{matrix} z = \mathbf{D}^t_{\text{com}}(i,j) \\ (i,j) \in \mathcal{S}^t_{\text{top-}K} \end{matrix} \right\}. \tag{13}$$

Finally, these 3-D points are transformed into the body frame $\mathcal{B}$ and then to the target-centric frame $\mathcal{T}$ using the known extrinsics

$$\mathcal{S}^t_{\mathcal{T}} = \left\{ \mathbf{q}_{\mathcal{TB}} \cdot (\mathbf{q}_{\mathcal{BC}} \odot \mathbf{s}_{\mathcal{C}} + \mathbf{t}_{\mathcal{BC}}) \;\middle|\; \mathbf{s}_{\mathcal{C}} \in \mathcal{S}^t_{\mathcal{C}} \right\}. \tag{14}$$

These target-centric obstacle points $\mathcal{S}^t_{\mathcal{T}}$ are published at each planning cycle and used to construct hard safety constraints within our NMPC formulation.

### C. PLANNING AND CONTROL

We adopt an NMPC framework to compute optimal control commands that enable fast, smooth, and reactive flight while tracking the target and avoiding obstacles. At each time step, the NMPC solves an optimal control problem over a finite horizon, minimizing deviations from a dynamically updated target reference while ensuring dynamic feasibility and safety.

#### 1) OBJECT-CENTRIC MODELING

We define the object-centric state of the quadrotor relative to the moving target as

$$\mathbf{x} = \begin{bmatrix} \mathbf{p}_{\mathcal{T}} \\ \mathbf{v}_{\mathcal{T}} \\ \mathbf{q}_{\mathcal{TB}} \\ \boldsymbol{\omega}_{\mathcal{B}} \end{bmatrix} \in \mathbb{R}^{13} \tag{15}$$

and the control input $\mathbf{u} \in \mathbb{R}^4$ as the vector of motor thrusts. Thus, the quadrotor's continuous-time dynamics evolve as

$$\dot{\mathbf{x}} = \begin{bmatrix} \dot{\mathbf{p}}_{\mathcal{T}} \\ \dot{\mathbf{v}}_{\mathcal{T}} \\ \dot{\mathbf{q}}_{\mathcal{TB}} \\ \dot{\boldsymbol{\omega}}_{\mathcal{B}} \end{bmatrix} = \begin{bmatrix} \mathbf{v}_{\mathcal{T}} \\ (\mathbf{q}_{\mathcal{TB}} \odot \boldsymbol{\tau})/m + \boldsymbol{g}_{\mathcal{T}} \\ (\mathbf{q}_{\mathcal{TB}} \odot \boldsymbol{\omega}_{\mathcal{B}})/2 \\ \mathbf{J}^{-1}(\boldsymbol{\mu} - \boldsymbol{\omega}_{\mathcal{B}} \times \mathbf{J}\boldsymbol{\omega}_{\mathcal{B}}) \end{bmatrix} \tag{16}$$

where $m$ is the quadrotor mass, $\mathbf{g}_{\mathcal{T}} = \begin{bmatrix} 0, & 0, & -9.81 \end{bmatrix}^{\top}$ is gravity in the target frame, $\mathbf{J} = \text{diag}(J_{xx}, J_{yy}, J_{zz})$ is the diagonal moment of inertia matrix, and the collective thrust $\boldsymbol{\tau}$ and torque $\boldsymbol{\mu}$ of the quadrotor are defined as

$$\boldsymbol{\tau} = k_\tau \sum_{i=0}^{3} u_i^2, \quad \boldsymbol{\mu} = \begin{bmatrix} k_\tau l(u_0^2 + u_1^2 - u_2^2 - u_3^2) \\ k_\tau l(-u_0^2 + u_1^2 + u_2^2 - u_3^2) \\ k_\mu(u_0^2 - u_1^2 + u_2^2 - u_3^2) \end{bmatrix} \tag{17}$$

where $k_\tau$ is the rotor thrust constant, $k_\mu$ is the rotor torque constant, and $l$ is the length of the quadrotor arm.

To obtain a discrete-time model, we define the dynamics function $f(\mathbf{x}, \mathbf{u}, \delta t)$ as the result of forward integration of the continuous-time dynamics over a fixed time step $\delta t$. Specifically, the next state is computed as

$$\mathbf{x}^{t+1} = f(\mathbf{x}^t, \mathbf{u}^t, \delta t) \tag{18}$$

where, in practice, this integration can be approximated using numerical methods such as Euler or Runge–Kutta [67].

This state-action-dynamics formulation is commonly used for quadrotors and provides a flexible foundation that supports sophisticated dynamic models, including aerodynamics, vibrations, motor interactions, and drag forces [68], [69], [70], [71]. By adopting this formulation for visual target tracking, we can leverage the extensive prior work on dynamics modeling developed for traditional trajectory tracking tasks [72].

### 2) REFERENCE GENERATION

Directly commanding the quadrotor to reach the target's estimated position can lead to unsafe behavior, particularly under noisy observations or occlusion. To mitigate this, we implement a goal-shifting strategy that maintains a minimum horizontal safety margin, denoted $d_{\text{safe}}$, in the xy-plane. This creates a cylindrical exclusion zone around the target, allowing for vertical flexibility (e.g., during takeoff or hovering) while maintaining lateral separation.

The shifted reference is computed along the horizontal line-of-sight vector from the target to the current quadrotor position

$$\bar{\mathbf{p}}_{\mathcal{T}}^t = d_{\text{safe}} \cdot \frac{[p_{\mathcal{T},x}, p_{\mathcal{T},y}, 0]^{\top}}{\sqrt{p_{\mathcal{T},x}^2 + p_{\mathcal{T},y}^2}} \tag{19}$$

ensuring that the reference remains directionally aligned in the plane while enforcing a safe lateral offset.

To avoid discontinuities during tracking and to facilitate safe takeoff behavior, this shifted goal is blended with the robot's current position using a time-varying ramp function

$$\bar{\mathbf{p}}_{\mathcal{T}}^t = \alpha(t) \cdot \mathbf{p}_{\mathcal{T}} + (1 - \alpha(t)) \cdot \bar{\mathbf{p}}_{\mathcal{T}}^t \tag{20}$$

where $\alpha(t)$ is a monotonically decreasing ramp function with quadratic decay in the horizontal plane and linear decay in altitude. Initially, $\alpha(0) = 1$, anchoring the reference to the current position. As $\alpha(t) \to 0$, the reference converges to the shifted goal, ensuring smooth takeoff and tracking.

The reference trajectory passed to the NMPC is generated by replicating the same desired state over the prediction horizon of length $N$

$$\bar{\mathbf{x}}_k^t = \begin{bmatrix} \bar{\mathbf{p}}_{\mathcal{T}}^t \\ \mathbf{0} \\ \bar{\mathbf{q}}_{\text{yaw}}^t \\ \mathbf{0} \end{bmatrix} \quad \forall k \in [0, N) \tag{21}$$

where $\bar{\mathbf{q}}_{\text{yaw}}^t$ is the desired yaw quaternion. The desired yaw angle $\bar{\psi}^t$ is obtained by correcting the current vehicle yaw $\psi^t$, extracted from the onboard quaternion $\mathbf{q}_{\mathcal{TB}}^t$, with the yaw offset that would center the target in the image

$$\bar{\psi}^t = \psi^t - \tan^{-1}\left(\frac{x - c_x}{f_x}\right). \tag{22}$$

This yaw angle is then converted to a quaternion, assuming zero roll and pitch

$$\bar{\mathbf{q}}_{\text{yaw}}^t = \left[\cos\left(\bar{\psi}^t/2\right)\ 0\ 0\ \sin\left(\bar{\psi}^t/2\right)\right]^{\top}. \tag{23}$$

The reference control inputs are simply set to zero for regularizing the control effort, $\bar{\mathbf{u}}_k = \mathbf{0}\ \forall k \in [0, N)$.

### 3) COST FUNCTION

The NMPC cost is designed to promote accurate target tracking, smooth control, and stable forward-facing flight. At each time step $k$, the stage cost is defined as

$$J(\mathbf{x}, \mathbf{u}) = \underbrace{\|\mathbf{x} - \bar{\mathbf{x}}_k\|_{\mathbf{Q}_x}^2}_{\text{State cost}} + \underbrace{\|\mathbf{u} - \bar{\mathbf{u}}_k\|_{\mathbf{Q}_u}^2}_{\text{Input cost}} + \underbrace{\left\|v_y^{\mathcal{B}}\right\|_{Q_{\text{o}}}^2}_{\text{Orbiting cost}} \tag{24}$$

where $\mathbf{Q}_x$ and $\mathbf{Q}_u$ are positive-definite weighting matrices, and $Q_{\text{o}}$ is a scalar that penalizes lateral motion in the body frame. The term $v_y^{\mathcal{B}}$ is the lateral component of the quadrotor's velocity in its own body frame.

The first two terms are standard quadratic costs that drive the system toward the reference state while regularizing actuation effort. The third term addresses a fundamental observability limitation inherent to visual target tracking.

Specifically, when only a single target is detected, the quadrotor's position relative to the target is observable, but its yaw orientation remains underconstrained. This creates ambiguity in the rotation about the vertical axis, permitting orbiting or lateral drift behaviors that preserve the perceived target location but reduce stability and situational awareness.

While previous methods [60] resolve this by tracking multiple targets, we argue that such assumptions are impractical in cluttered, dynamic, or degraded visual environments where sustaining visibility of multiple objects is unreliable.

Instead, we introduce a soft regularization on $v_y^{\mathcal{B}}$, exploiting the fact that unnecessary orbiting correlates with sustained lateral body velocity. This regularization does not remove lateral motion entirely (since velocity estimates naturally jitter with detection noise and fast maneuvers) but it significantly reduces its magnitude and prevents persistent orbiting. The result is a forward-facing, more stable flight without additional perceptual burden.

#### 4) CBFs FOR OBSTACLE AVOIDANCE

To ensure safety during target tracking, we incorporate second-order CBFs into the NMPC to enforce minimum distance constraints with respect to perceived obstacles. Each obstacle point $\mathbf{s}_{\mathcal{T}}^t \in \mathcal{S}_{\mathcal{T}}^t$ is derived from depth perception and expressed in the target frame $\mathcal{T}$, consistent with object-centric control formulations proposed in recent works [55].

For each high-risk point, we define a safety function

$$h_k(\mathbf{x}^t) = \|\mathbf{s}_{\mathcal{T}}^t - \mathbf{p}_{\mathcal{T}}^t\|^2 - Q_{\max}^2 \tag{25}$$

where $Q_{\max} = \max(Q_x, Q_y)$ is the minimum allowable distance that accounts for the robot's physical dimensions.

Safety is enforced by requiring the CBF condition

$$\ddot{h}_k(\mathbf{x}^t, \mathbf{u}^t) + 2\lambda \dot{h}_k(\mathbf{x}^t) + \lambda^2 h_k(\mathbf{x}^t) \geq 0 \tag{26}$$

which guarantees forward invariance of the safe set defined by $h_k(\mathbf{x}) \geq 0$. Here, $\lambda > 0$ controls the rate of enforcement.

The first derivative of the safety function is

$$\dot{h}_k(\mathbf{x}^t) = 2(\mathbf{s}_{\mathcal{T}}^t - \mathbf{p}_{\mathcal{T}}^t)^\top \mathbf{v}_{\mathcal{T}}^t \tag{27}$$

and the second derivative, capturing the effect of control inputs, is given by

$$\ddot{h}_k(\mathbf{x}^t, \mathbf{u}^t) = 2\|\mathbf{v}_{\mathcal{T}}^t\|^2 + 2(\mathbf{s}_{\mathcal{T}}^t - \mathbf{p}_{\mathcal{T}}^t)^\top \dot{\mathbf{v}}_{\mathcal{T}}(\mathbf{u}) \tag{28}$$

where $\dot{\mathbf{v}}_{\mathcal{T}}(\mathbf{u})$ is the acceleration of the robot in the target frame, obtained from the dynamics model $f(\mathbf{x}^t, \mathbf{u}^t, \delta t)$.

Each of these constraints is imposed at runtime for every selected obstacle point $\mathbf{s}_{\mathcal{T}}^t$, ensuring that the NMPC respects proximity constraints while pursuing the tracking objective. This formulation allows us to encode collision avoidance in a computationally efficient and differentiable form compatible with real-time optimization.

#### 5) OPTIMIZATION PROBLEM

At each planning step, the NMPC solves a constrained optimal control problem to compute a dynamically feasible and safe control sequence over a finite horizon $N$. The objective is to minimize a cumulative cost that promotes accurate tracking, smooth control, and orbit-free motion

$$\min_{\substack{\mathbf{x}^t,\ldots,\mathbf{x}^{t+N} \\ \mathbf{u}^t,\ldots,\mathbf{u}^{t+N-1}}} \sum_{j=0}^{N-1} J(\mathbf{x}^{t+j}, \mathbf{u}^{t+j}) + J(\mathbf{x}^{t+N}, \mathbf{0}) \tag{29}$$

where the terminal cost is evaluated with zero control input.

The optimization is subject to the following constraints for all predictions $j \in [0, N]$ and safety constraints $k \in [0, K)$:

$$\mathbf{x}^t = \hat{\mathbf{x}}^t \tag{30a}$$

$$\mathbf{x}^{t+1+j} = f(\mathbf{x}^{t+j}, \mathbf{u}^{t+j}) \tag{30b}$$

$$\mathbf{x}_{\min} \leq \mathbf{x}^{t+j} \leq \mathbf{x}_{\max} \tag{30c}$$

$$\mathbf{u}_{\min} \leq \mathbf{u}^{t+j} \leq \mathbf{u}_{\max} \tag{30d}$$

$$\ddot{h}_k(\mathbf{x}^{t+j}, \mathbf{u}^{t+j}) + 2\lambda \dot{h}_k(\mathbf{x}^{t+j}) + \lambda^2 h_k(\mathbf{x}^{t+j}) \geq 0. \tag{30e}$$

This formulation allows the NMPC to reason over future trajectories while embedding safety constraints and actuation limits into a unified reactive control framework.

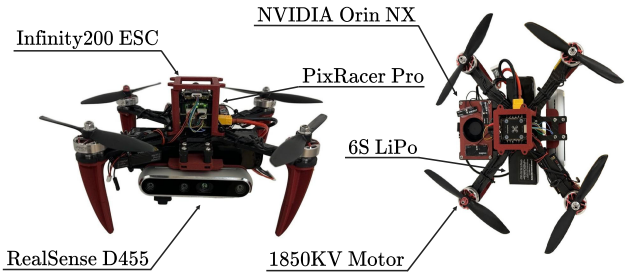



**FIGURE 7. Aerial robot platform. Custom quadrotor used in experiments, equipped with an onboard computer, a stereo RGB-D camera, and a PX4 flight controller.**

## IV. EXPERIMENTAL SETUP

### A. AERIAL SYSTEM

Our quadrotor platform, shown in Fig. 7, weighs 1.3 kg and has a motor span of 25 cm, with a thrust-to-weight ratio of approximately 4 to 1. It is powered by a 6S LiPo battery and uses 1850-KV motors paired with a NewBeeDrone Infinity200 V2 4IN1 ESC 55A to enable agile flight.

The system integrates an Intel RealSense D455 stereo camera for visual and depth sensing and an onboard IMU. The camera is mounted front-facing, and its intrinsics are $(f_x, f_y, c_x, c_y) = (325.2, 430.9, 323.1, 246.9)$, obtained from standard calibration [65]. The camera-to-body transformation is defined by the static quaternion $\mathbf{q}_{\mathcal{BC}} = [-0.5, 0.5, -0.5, 0.5]^\top$ and translation $\mathbf{t}_{\mathcal{BC}} = [0.061, 0.047, -0.065]^\top$ m. RGB-D frames are captured at 60 Hz with a resolution of 640 $\times$ 480.

Onboard processing is handled by an NVIDIA Jetson Orin NX (16 GB), while low-level stabilization is managed by a PixRacer Pro flight controller running PX4 v1.14. Communication between the Jetson and PixRacer is established via MAVLink over UART. Although the dynamics are expressed at the per-motor thrust level for completeness, in practice, the NMPC outputs collective thrust and body-rate commands. This design is motivated by two factors: i) hardware constraints, since the PixRacer permits reading but not commanding individual motor RPMs; and ii) prior evidence that collective thrust and body-rate policies yield greater robustness than motor-level ones [73]. The PixRacer executes the

inner-loop attitude–thrust control and motor mixing through its onboard PID stabilization.


**FIGURE 8. Targets used for tracking. Experiments use either a 1.8-m mannequin on an ATV or a 0.8-m reflective stop sign. The mannequin simulates a human presence, while the stop sign adds glare and reflection challenges. The ATV moves the targets dynamically at speeds up to 50 km/h.**

### B. TARGET DETECTOR

We employ a YOLOv11-small model for object detection, trained from scratch on the COCO dataset with heavy augmentation. Images are randomly zoomed within a scale range of [0.6, 1.4], rotated within $\pm 15°$ in pitch and roll, and augmented with synthetic motion blur (30% probability) and brightness shifts of $\pm 20\%$. Crops are resized to $320 \times 256$ for inference. The model is trained using Adam with a learning rate of $2 \times 10^{-4}$, batch size 32, weight decay $10^{-5}$, and for 300 epochs. The inference runs at 92 Hz using TensorRT on the Orin NX with floating 16 precision.

To initialize tracking, we use a one-time prompting stage that leverages a large, high-capacity object detector. Specifically, a YOLOv11x model [74] is run once on the full-resolution RGB image to acquire an initial bounding box for the target. This model is not trained or fine-tuned within our system. It is used solely for inference. Among all detections returned, the bounding box with the highest confidence score is selected. This bounding box seeds the zooming module for subsequent frames. After this initialization, the large YOLOv11x model is terminated to eliminate unnecessary computational overhead. From that point onward, tracking continues using only the lightweight onboard detector. Importantly, this prompting mechanism is modular. Any object detector capable of producing a coarse bounding box on the initial frame can be substituted without modifying the downstream pipeline. The sole requirement is that it provides an early spatial prior for focusing the zoomed window.

In our experiments, the detector is applied to human-shaped mannequins and stop signs mounted on an ATV (see Fig. 8). These objects were selected to enable safe, repeatable evaluation of high-speed ground target tracking while ensuring that the targets remain within standard COCO categories for reliable detection. Although these objects are used for validation, the framework itself is class-agnostic: any detector trained on the desired class can be integrated without altering the downstream pipeline.

### C. DEPTH COMPLETION

Depth maps from the RealSense D455 (in HighAccuracy mode [75]) are completed using DepthAnythingV2 [64] with a ViT-S backbone, producing per-frame estimates in 31.6 ms. A histogram bin width of 0.15 m is used for mode filtering.

### D. OBSTACLE SELECTION

We divide the image into $17 \times 17$ nonoverlapping cells and select $K = 10$ obstacle points with the lowest TTC. The drone's projected dimensions are $Q_x = 0.4$ m, $Q_y = 0.2$ m.

### E. STATE ESTIMATION

Target-relative state is tracked using an UKF that fuses asynchronous visual detections with high-rate inertial measurements. The filter estimates the full 6-DoF state along with accelerometer and gyroscope biases.

The UKF process noise standard deviations are defined as follows: linear acceleration noise is $(0.1, 0.1, 1.0)$ m/s$^2$, angular velocity noise is $(0.2, 0.2, 0.2)$ rad/s, accelerometer bias noise is $(0.1, 0.1, 0.1)$ m/s$^2$, and gyroscope bias noise is $(0.1, 0.1, 0.1)$ rad/s. Measurement noise standard deviations for the visual update are 0.01 m in $(x, y)$ and 0.001 m in $z$ for the position estimate, and 0.0001 for all four quaternion components in the orientation update.

The IMU provides linear acceleration and angular velocity at 200 Hz. The UKF is also executed at this rate and handles bias estimation online without requiring separate precalibration. All position and orientation estimates are computed in the target-centric frame described in Section III-B.3.

For outdoor flights, a u-Blox Neo-M9N GPS module was fused with IMU data using an extended Kalman filter (EKF), yielding position updates at 100 Hz. This global localization information was purely used for benchmarking. The proposed system operates **independently** of this information during all the flight experiments.

### F. PLANNING AND CONTROL

We formulate an NMPC problem with a prediction horizon of $N = 10$ steps over 2 s and a discrete time step of $\delta t = 0.2$ s. The dynamics are integrated using Runge–Kutta. The goal reference is shifted using a lateral safety margin of $d_{\text{safe}}$ chosen based on the experiment, and blended with a ramp function $\alpha(t)$ with quadratic horizontal decay over 1 s. The desired yaw angle is computed to center the target in the image, and converted to a quaternion with zero roll pitch.

The NMPC cost function includes a state deviation penalty $\mathbf{Q}_x = \text{diag}(150, 100, 150, 15, 15, 15, 50, 15, 15, 50, 5, 5, 5)$, input cost $\mathbf{Q}_u = \text{diag}(1, 1, 1, 1)$, and an orbiting penalty $Q_o = 20$ on lateral body velocity. State constraints are enforced with element-wise bounds $\mathbf{x} \in [-999, 999]^3 \times [-25, 25]^3 \times [-10, 10]^4 \times [-40, 40]^3$, and control inputs are constrained to $\mathbf{u} \in [0.05, 8.0]^4$. Safety constraints are enforced using second-order CBFs with a safety radius $Q_{\max} = 0.3$ m and gain $\lambda = 2$. Up to

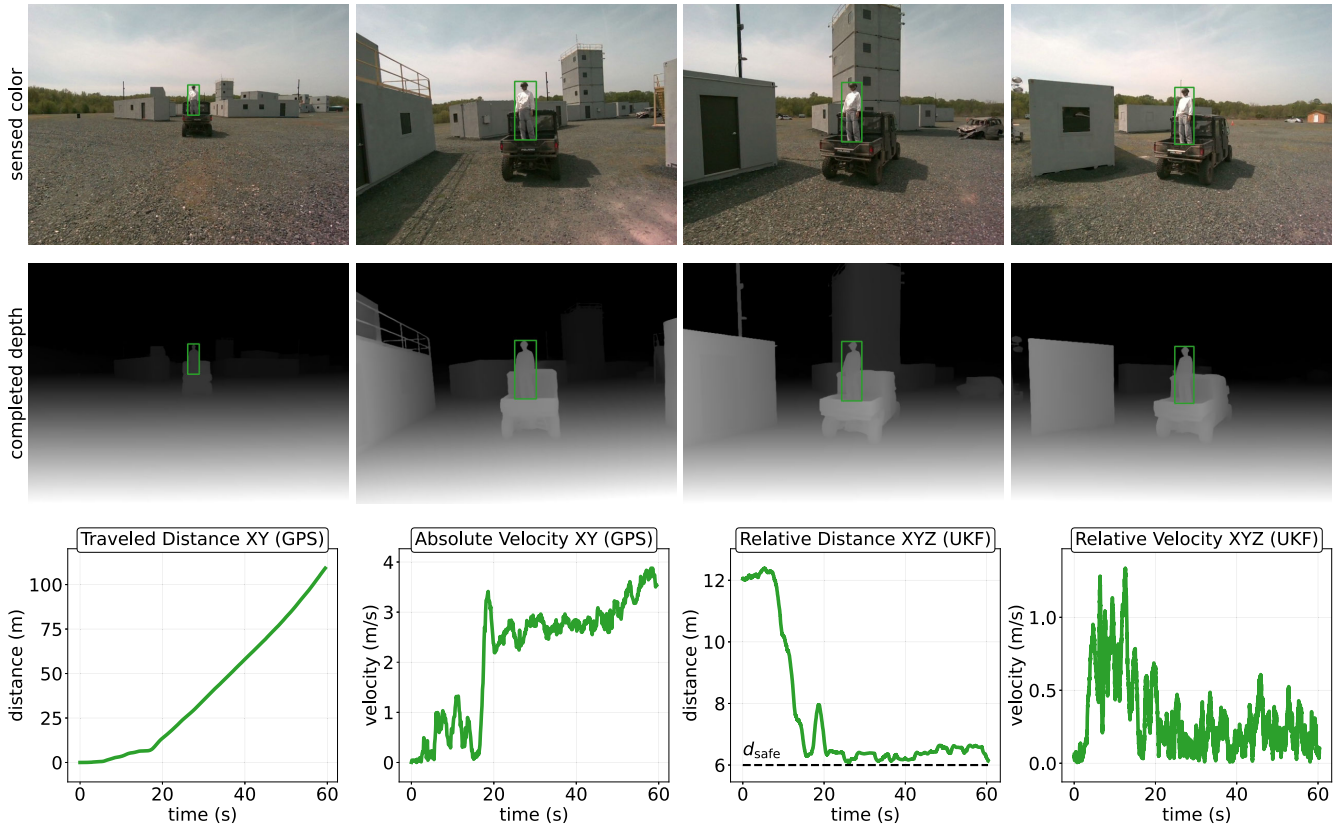



**FIGURE 9. Tracking performance in a structured urban maze. NOVA follows the target through narrow corridors maintaining stable depth estimation and obstacle awareness despite minimal texture. The bottom shows consistent velocity regulation and relative distance control, confirming safe and responsive tracking throughout the mission.**

10 constraints are active per solve. The optimization problem is solved with `acados` [76], using SQP with Gauss–Newton Hessian approximation and Levenberg–Marquardt regularization equal to $10^{-2}$, with a max iteration limit of 20 and a feasibility tolerance of $10^{-4}$.

## V. EXPERIMENTAL RESULTS

We evaluate NOVA in real-world missions that test its performance under fast motion, degraded perception, and minimal prior knowledge. Unlike isolated benchmarks, these flights require full onboard planning, perception, and control in real time. Our aim is to assess not just whether NOVA can track a moving target, but how it holds up when each subsystem is stressed.

We ask three questions: i) Can it maintain visual lock on fast targets across varied terrain? ii) Does it generalize across urban, forested, and indoor–outdoor transitions without retuning? iii) Is its behavior consistent under repeated trials, trajectory changes, and viewpoint shifts?

To probe this, we deploy NOVA in progressively harder environments that tax specific stack components: tracking under occlusion and blur, depth under lighting collapse, and planning under tight geometry. The system runs with one fixed configuration—no parameter tuning, map registration, or external localization. All decisions derive from raw onboard observations in real time.

### A. CONTAINER MAZE NAVIGATION

We evaluate NOVA in a structured but constrained setting: a container maze built from stacked shipping units. The layout resembles an urban canyon with narrow corridors, sharp turns, and limited escape paths. Additional obstacles include poles, fencing, and a wrecked vehicle, while blind corners disrupt the line of sight. The flat gray container walls provide minimal texture.

Fig. 9 shows representative sequences from one flight. The RGB images capture moments during the tracking performance, while the depth maps demonstrate consistent geometry extraction despite weak visual cues. As the UAV advances, it continuously modulates velocity and acceleration. The bottom plots quantify this behavior: relative distance remains within a safe range, and the UAV's speed closely matches the ATV's even through sudden turns.

### B. FOREST TRAIL PURSUIT

We next evaluate NOVA in unstructured natural terrain, where sensing and control are challenged by long-range motion, high speeds, and unpredictable conditions. The 1 km forest

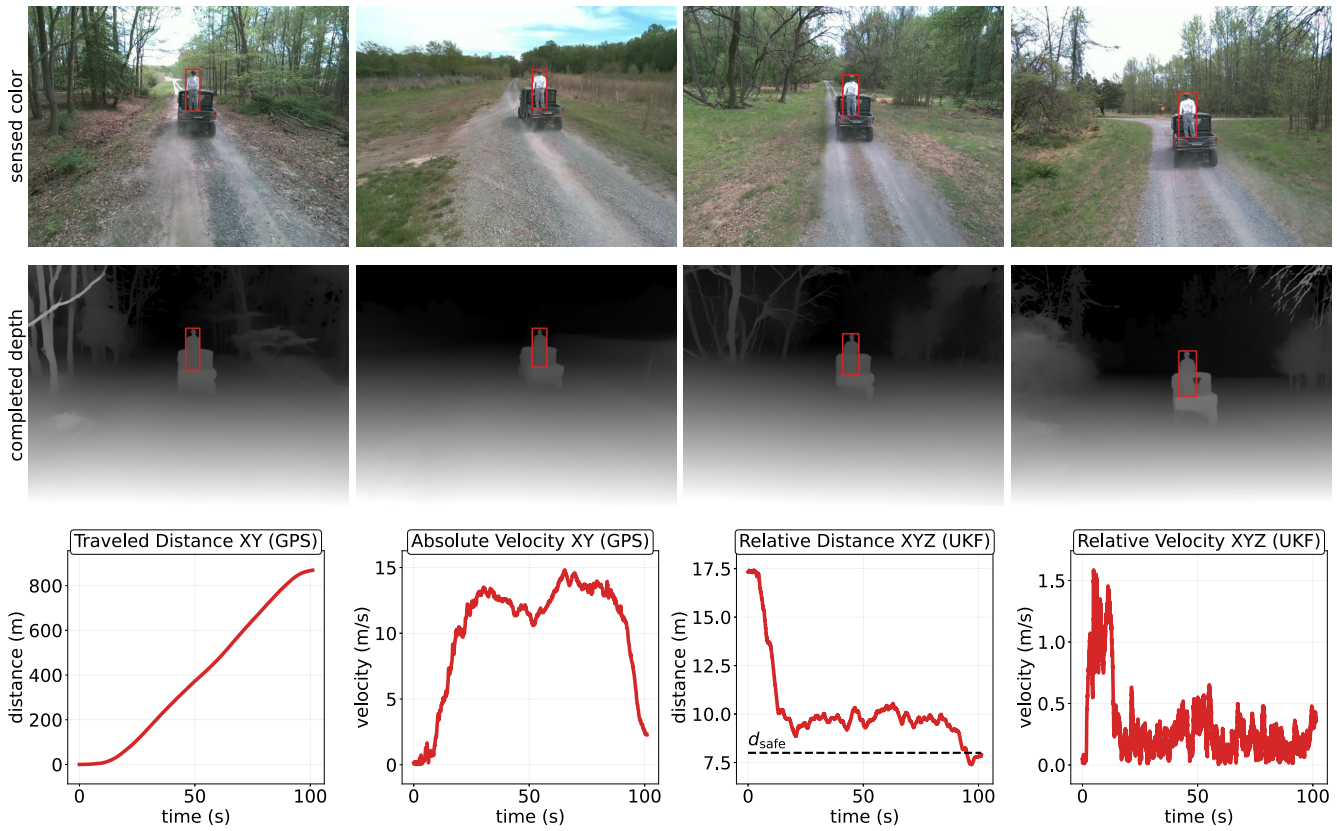



**FIGURE 10. High-speed pursuit in a rugged forest trail. NOVA tracks the target over a 1 km unstructured path with potholes, dust, and vegetation. The run induces strong lighting transitions, airborne dust, motion blur, and occasional visual occlusion by branches. While these degradations are not always visible in the static frames, they are clearly observed in the video and reflected in the quantitative plots. The bottom reports traveled distance, target velocity, and relative distance/velocity across the mission, confirming that the UAV sustains target speed while maintaining a safe distance.**

trail alternates between dense canopy and open clearings, producing rapid lighting transitions. The uneven ground jolts the ATV, dust adds visual noise, and overhanging branches with narrow passageways demand fast obstacle avoidance. The ATV reaches up to 50 km/h, forcing the UAV to sustain high-speed flight while keeping visual lock, estimating depth, and planning safe paths in real time. Without structured boundaries or predictable layouts, NOVA relies solely on onboard sensing for adaptive navigation.

As shown in Fig. 10, the UAV maintains close formation throughout. Despite motion blur and exposure shifts, the system consistently detects the target and produces dense depth maps with usable geometry. Velocity plots show near-zero relative speed, confirming high temporal tracking fidelity. The UAV regulates distance smoothly and navigates cluttered vegetation without hesitation or instability.

### C. HANGAR TRANSITION WITH VISUAL AMBIGUITY

We evaluate NOVA in a scenario designed to stress perception to its limits. The mission begins in an open gravel lot, transitions into a tall metallic hangar, and exits back outdoors. This path produces cascading challenges: GPS dropout inside the hangar, abrupt lighting shifts from overexposure to near-darkness, and a cluttered interior with reflective surfaces, structural obstacles, and narrow doorways. To further raise the difficulty, two additional mannequin targets are placed near one exit. These decoys closely resemble the mannequin on the ATV, forcing the system to preserve target identity during degraded visibility and unstable lighting. This explicitly tests robustness to ambiguity, occlusion, and potential target switching.

Such environments typically break conventional tracking pipelines. Feature-based methods struggle with low texture, stereo matching fails under reflections, and bright-to-dark transitions overwhelm standard exposure control. GPS-based localization is unavailable inside the structure, and visual odometry is easily corrupted.

Despite these conditions, NOVA maintains continuous tracking. As shown in Fig. 11, RGB and completed depth frames remain usable even under poor lighting and weak geometry. When stereo cues degrade, NOVA falls back on inertial integration and histogram-filtered depth estimates to maintain relative pose. It requires no artificial markers, global maps, or external infrastructure. Trajectory plots confirm

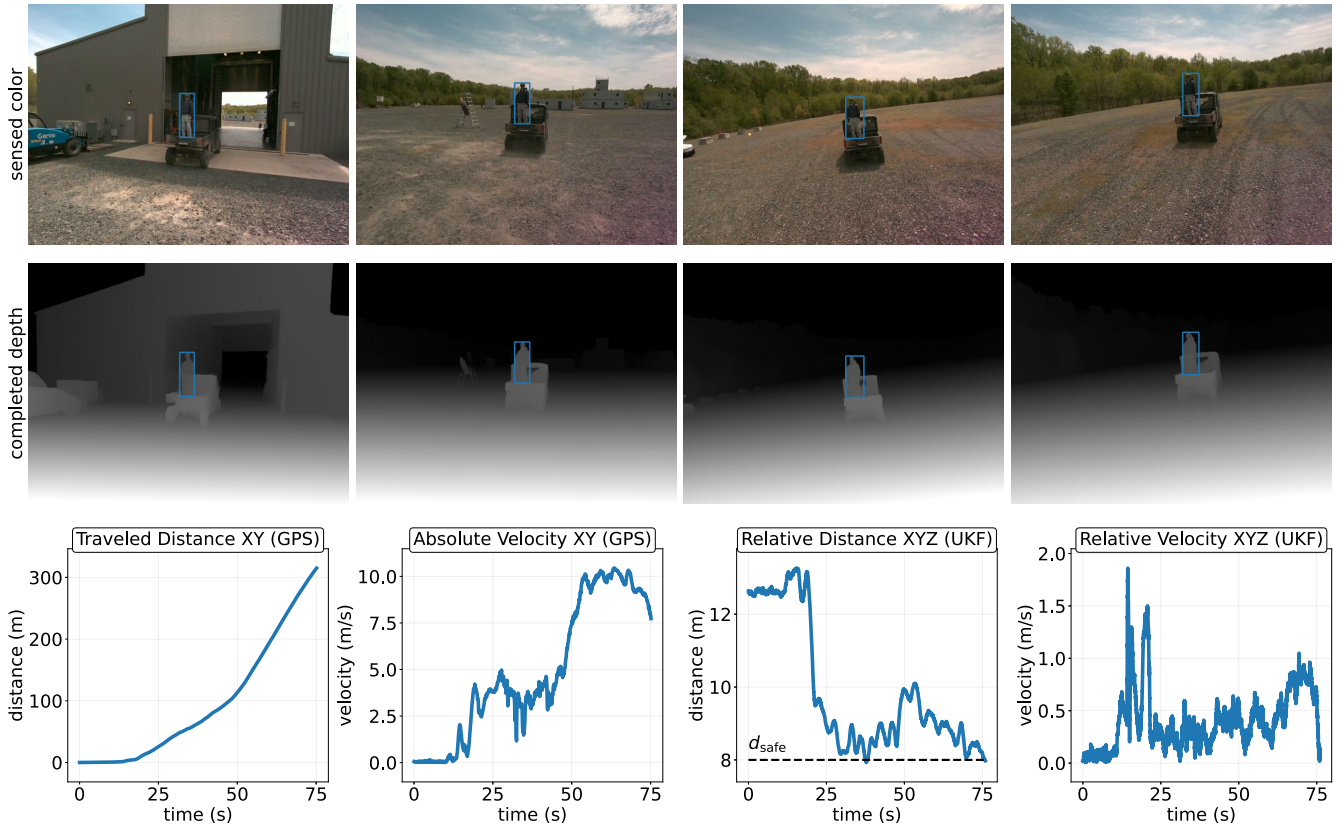



**FIGURE 11. Robust tracking across indoor–outdoor transitions. The UAV follows the target from a gravel lot through a tall metallic hangar and back outside, navigating abrupt lighting changes, GPS dropout, and visual sparsity. Despite exposure shifts and structural occlusions, NOVA maintains a safe distance and continuous target lock throughout the mission. Additional details of the mission, including lighting variation and occlusion events, are more clearly visible in the supplementary video.**

smooth flight: relative distance remains stable, with no abrupt control shifts. The UAV transitions through the building without reinitialization, rejects the decoy mannequins, and sustains accurate lock on the moving target throughout the most challenging phase.

### D. ELEVATED TRACKING ACROSS MIXED TERRAIN

We evaluate NOVA under altered spatial configuration, requiring the UAV to maintain a constant 6-m vertical offset from the target—flying at nearly 10-m altitude—through a continuous mission spanning the parking lot, hangar, gravel yard, and container maze. This setup compounds perception and control challenges. At higher altitude, reduced stereo parallax degrades depth quality, the target occupies fewer pixels, and detector robustness declines. The UAV's field of view also narrows with respect to lateral motion, complicating the anticipation of turns and accelerations. Obstacle avoidance becomes critical around overhanging structures, hangar entrances, and confined maze corridors. The environments amplify these issues: the hangar induces GPS dropout and lighting shifts, while the container maze adds frequent occlusions, low-texture surfaces, and sharp geometry.

From above, NOVA must preserve tight target coupling while flying over clutter. As shown in Fig. 12, NOVA completes the full mission without loss of lock or violation of safety bounds. Despite degraded sensing geometry, it maintains coherent relative estimates, reconstructs dense depth, and regulates distance and velocity smoothly. This experiment shows NOVA generalizes not only across environments but also across spatial configurations that alter perception and planning dynamics. No system modifications are introduced; all modules run unchanged, underscoring adaptability to shifted tracking regimes.

### E. MEASURED PERFORMANCE ACROSS TESTED ENVIRONMENTS

To complement the qualitative demonstrations, we present quantitative metrics from the four representative experiments described earlier. Table 1 summarizes overall system performance, including GPS-based velocity, control effort, pitch dynamics, target distance regulation, and the consistency of object detections across repeated trials.

In the Urban Maze, the UAV operates within narrow corridors and sharp turns, demanding precise obstacle avoidance and frequent reacceleration. NOVA sustains high detection

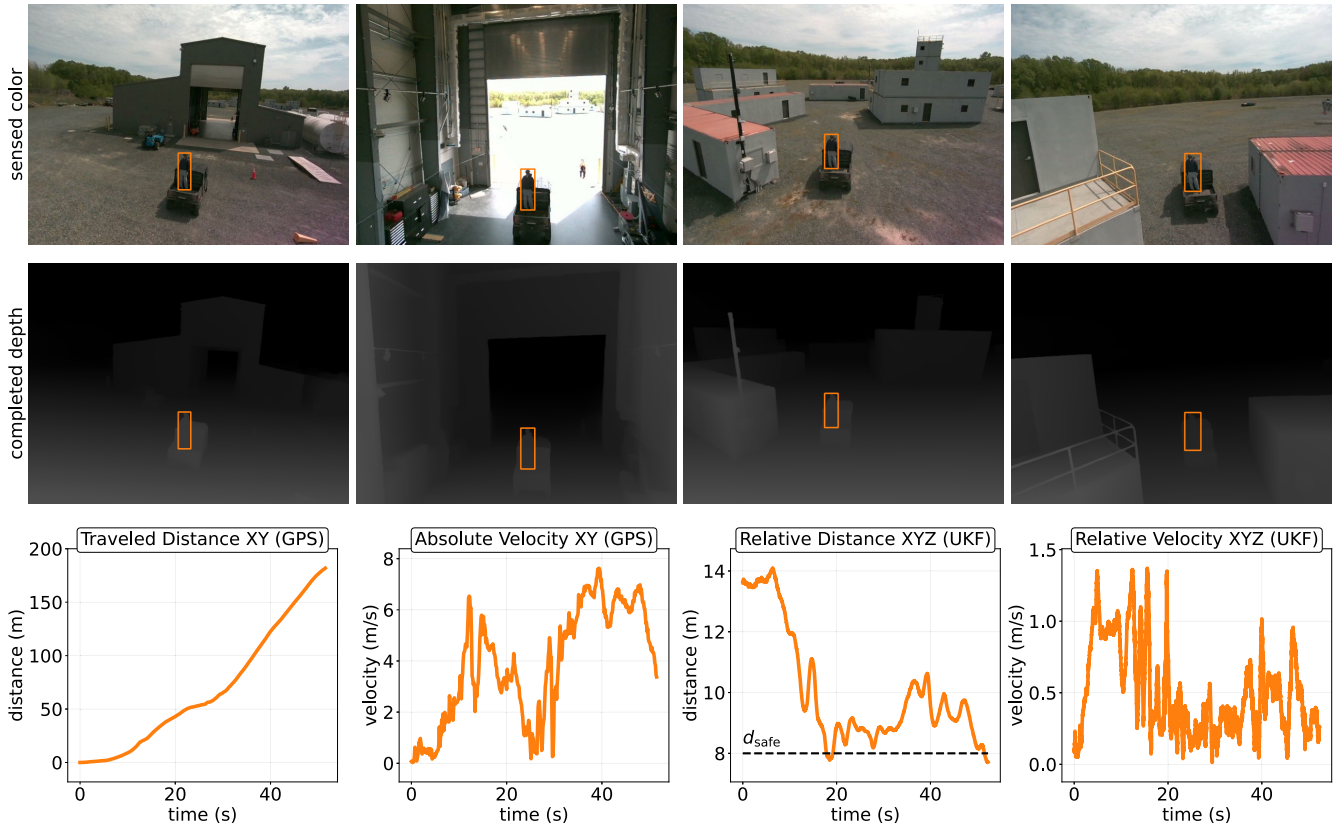



**FIGURE 12. Tracking from elevated viewpoints with forced height offset. The UAV follows the target through a combined indoor–outdoor and urban maze mission while maintaining a 6-m vertical offset. This setup introduces degraded stereo geometry, reduced field of view, and tight spatial constraints. NOVA preserves visual lock, avoids obstacles, and regulates target distance despite elevated flight and complex terrain. See video for full mission execution.**

**TABLE 1. Quantitative performance across the four outdoor tracking missions. We report results from representative trials in diverse environments, each introducing unique challenges in geometry, speed, sensing, and elevation. NOVA consistently maintains safe separation, stable flight, and high detection rates without environment-specific tuning. Reported metrics include UAV velocity (mean and peak), linear acceleration, pitch angle, mean and minimum distance to the target compared to the safety threshold $d_{\text{safe}}$, and overall detection rate.**

| Scenario | GPS Velocity (km/h) | | Control Effort (m/s²) | | Pitch (°) | | Rel. Distance (m) | | | Detections (%) |
|---|---|---|---|---|---|---|---|---|---|---|
| | Mean | Max | Mean | Max | Mean | Max | Mean | Min | $d_{\text{safe}}$ | |
| Urban Maze | 5.7 | 14.0 | 10.2 | 32.4 | 2.3 | 15.0 | 8.2 | 6.2 | 6.0 | 98.6 |
| Forest Trail | 35.0 | 53.3 | 10.2 | 13.9 | 10.1 | 19.6 | 10.0 | 8.4 | 8.0 | 94.5 |
| Building Transition | 17.6 | 37.6 | 10.3 | 22.6 | 5.3 | 16.6 | 9.9 | 8.9 | 8.0 | 96.2 |
| Elevated Tracking | 14.5 | 27.5 | 10.3 | 22.8 | 7.7 | 28.8 | 9.4 | 8.8 | 8.0 | 93.1 |

consistency (98.6%) while maintaining an average separation of 8.2 m from the target, with a minimum of 6.2 m, just above the $d_{\text{safe}} = 6.0$-m threshold. The high peak acceleration of 32.4 m/s$^2$ and maximum pitch of 15.0° reflect sharp control inputs required to navigate tight geometry while preserving continuous visual lock.

The forest trail scenario emphasizes sustained high-speed tracking over unstructured terrain. Here, the target exceeds 53 km/h, and the UAV maintains a mean relative distance of 10.0 m, dipping to 8.4 m at its closest—comfortably within the $d_{\text{safe}} = 8.0$-m constraint. Control demands remain moderate, with limited acceleration spikes and a maximum pitch of 19.6°, indicating smooth flight despite image blur, shadows, and environmental noise. NOVA's object detector continues to operate reliably under these fast and noisy conditions, with 94.5% detection consistency.

In the building transition trial, the system faces full GPS dropout and sharp illumination shifts when passing through

a metallic hangar. NOVA adapts with increased control activity, reaching 22.6-m/s$^2$ peak acceleration and sustaining an average pitch of 5.3°. The UAV maintains an average target distance of 9.9 m and a minimum of 8.9 m, both inside the $d_{\text{safe}} = 8.0$-m threshold, while detection remains robust at 96.2%.

The elevated tracking experiment introduces an artificial vertical offset, forcing the UAV to maintain a higher flight altitude. At these distances, stereo depth estimation becomes less reliable and the target appears smaller in the image. NOVA compensates with stronger dynamic control, producing a maximum pitch of 28.8° and acceleration up to 22.8 m/s$^2$. The UAV sustains a mean distance of 9.4 m and a minimum of 8.8 m, again within the $d_{\text{safe}} = 8.0$-m constraint, and achieves 93.1% detection consistency despite reduced image resolution and more cluttered backgrounds.

Across all environments, NOVA consistently avoids collisions, preserves target visibility, and respects user-defined separation thresholds. These results underscore the system's robustness and adaptability across challenging real-world conditions.

**TABLE 2. Tracking performance consistency across multiple indoor–outdoor trials. Metrics include mean and minimum relative distance, user-defined safety threshold, and relative velocity statistics. NOVA maintains stable performance across all trials without tuning or adaptation.**

| Trial | Direction | Relative Distance (m) | | | Rel. Velocity (m/s) | |
|---|---|---|---|---|---|---|
| | | Mean | Min | $d_{\text{safe}}$ | Mean | Max |
| 1 | Forward | 9.9 | 8.9 | 8.0 | 0.1 | 0.5 |
| 2 | Forward | 9.5 | 8.7 | 8.0 | 0.2 | 0.4 |
| 3 | Reverse | 9.1 | 8.9 | 8.0 | 0.1 | 0.5 |
| 4 | Reverse | 9.3 | 8.6 | 8.0 | 0.2 | 0.5 |

### F. REPEATABILITY AND ROBUSTNESS

Robust operation in the real world requires more than handling isolated challenges. A tracking system must perform consistently across repeated trials, despite minor variations in conditions. To evaluate this, we repeat the full indoor–outdoor transition experiment four times under nominally identical setups. Two trials proceed in the forward direction, while two others are conducted in reverse, introducing mirrored geometry, altered lighting transitions, and different entry angles.

Despite these differences, NOVA maintains stable and consistent behavior. Fig. 13 presents sample onboard images from each run, illustrating reliable visual tracking under changes in viewpoint, shadowing, and background. No manual intervention is required between trials.

Quantitative results in Table 2 confirm this consistency. The repeatability study offers a straightforward statistical analysis across multiple independent trials, reporting mean, minimum, and velocity ranges to characterize performance. In every run, the UAV maintains safe separation well below the user-defined threshold of $d_{\text{safe}} = 8.0$ m, while keeping relative velocity and average spacing tightly bounded. Notably, the reversed-direction flights match the forward cases, showing that NOVA's behavior is robust and not reliant on environment-specific priors.

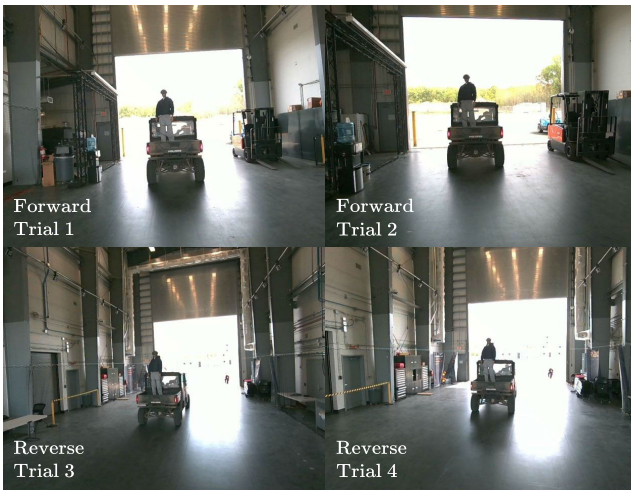



**FIGURE 13. Robust target tracking across repeated trials. Sample RGB frames from four separate indoor–outdoor transition experiments, showing variations in lighting, entry angle, and background. Table 2 reports quantitative relative distance and velocity statistics across these same trials, confirming repeatability and statistical consistency. Together, the images and table illustrate that NOVA maintains target lock and stable coupling at building entry/exit.**

### G. GPS SIGNAL QUALITY AND ITS LIMITATIONS

Although NOVA operates without GPS during flight, we use fused GPS+IMU estimates post hoc to visualize global trajectories and assess how environmental geometry affects localization quality. This provides a comparative baseline for understanding where GPS-based methods remain reliable and where they degrade.

Fig. 14 summarizes GPS performance across all four representative scenarios. The top rows show the global position trace, color-coded by estimated horizontal error (eph). The bottom plot reports the number of visible satellites over time. Together, these metrics reflect how surrounding structures influence satellite visibility and positioning accuracy.

In the urban maze and forest trail scenarios, satellite visibility remains high, typically above 20 satellites, and the estimated horizontal error stays below 1.0 m. These environments, although partially obstructed, allow for relatively consistent signal reception. The GPS data in these trials remains smooth and usable throughout the mission.

In contrast, the building transition and elevated offset experiments exhibit significant signal degradation. During the building transition mission, satellite count drops sharply as the UAV enters the hangar, with corresponding increases in estimated error, often exceeding 3.5 m. These effects are due to occlusion, multipath reflections, and direct signal loss.

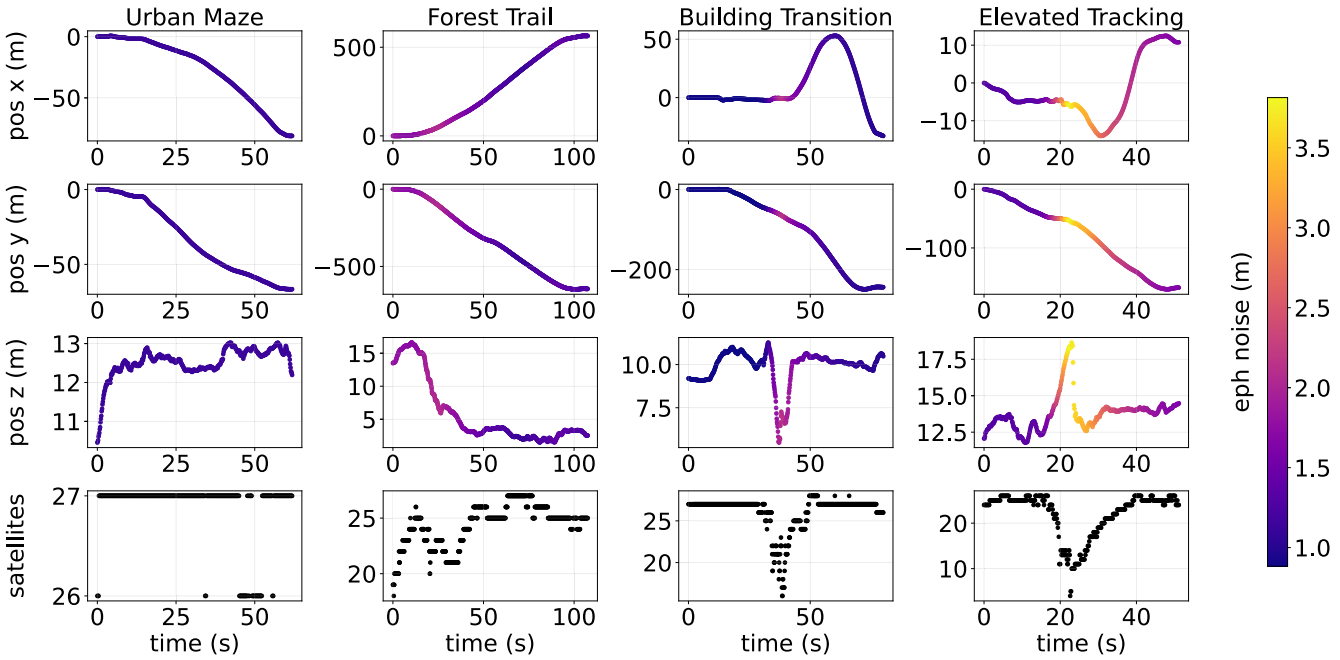



**FIGURE 14. GPS signal degradation across outdoor tracking scenarios. The top rows show fused GPS+IMU global position traces for each of the four outdoor missions, color-coded by estimated horizontal positional error (eph). Lighter colors indicate higher uncertainty. The bottom plot shows the number of GPS satellites tracked over time. In the urban maze and forest trail scenarios, signal quality remains high, with low positional error and stable satellite lock. In contrast, the building transition and elevated tracking missions exhibit significant degradation, including extended satellite dropout and increased eph. This effect is especially pronounced during indoor segments and when the UAV flies near structural ceilings, highlighting the limitations of GPS-based localization in partially enclosed or cluttered environments.**

The elevated tracking scenario presents even more severe GPS degradation. Although the trajectory includes both indoor and outdoor segments, the required vertical offset places the UAV closer to the ceiling during the hangar phase. This reduces the already limited visibility to the sky and exacerbates signal dropout. In several segments, the number of visible satellites falls to near zero, and position estimates become highly uncertain.

These observations highlight the limits of GPS-based localization in built environments. Even in outdoor settings, partial occlusion or reflective surfaces can compromise satellite visibility and introduce substantial pose uncertainty. NOVA's design avoids these failure modes by relying exclusively on onboard visual and inertial sensing, which remains operational across all tested scenarios regardless of external infrastructure availability.

### *H. ABLATION STUDIES*

To evaluate the contribution of individual components within the NOVA framework, we conduct a series of ablation studies. Each study isolates one module and examines its effect on overall system performance, while keeping the remainder of the stack fixed. The goal is to assess how specific design elements contribute to robustness, accuracy, and safety during target tracking in unstructured environments.

#### 1) ADAPTIVE ZOOM STRATEGY

The adaptive zoom module improves detection robustness by dynamically cropping the image around the expected target location before passing it to the detector. This strategy is designed to address two key challenges: i) suppressing visually similar distractors that can trigger identity switches and ii) preserving the detector's ability to recognize small, distant targets by maintaining spatial resolution over the region of interest.

##### a) Multitarget Robustness

We first assess the role of the adaptive zoom strategy in scenes with multiple visually similar objects. The setup includes three mannequins placed at different positions, all resembling each other in size and appearance. The UAV is instructed to track one specific mannequin while the others remain in view.

Fig. 15 shows representative trials. Without zoom, the detector occasionally confuses the target with a nearby distractor, leading to identity switches or temporary tracking failure. With adaptive zooming enabled, the system consistently maintains the correct target across all runs. This supports the hypothesis that reducing visual clutter at the detector input helps identify the target in ambiguous scenes.

While our results show that adaptive zooming already mitigates most identity switches in cluttered views by isolating and resolving the designated target, conventional

**FIGURE 15.** Impact of adaptive zoomed-in view multitarget scenes. In cluttered scenes with multiple visually similar targets (white), tracking can become unstable when relying on full-frame detection, often resulting in identity switches or drift. The adaptive zoom module addresses this by cropping tightly around the prompted target, suppressing distractors, and maintaining detection focus. Shown here are drone trajectories (blue) for three separate trials, each prompted to track a different mannequin. NOVA successfully adheres to the correct target in all cases.

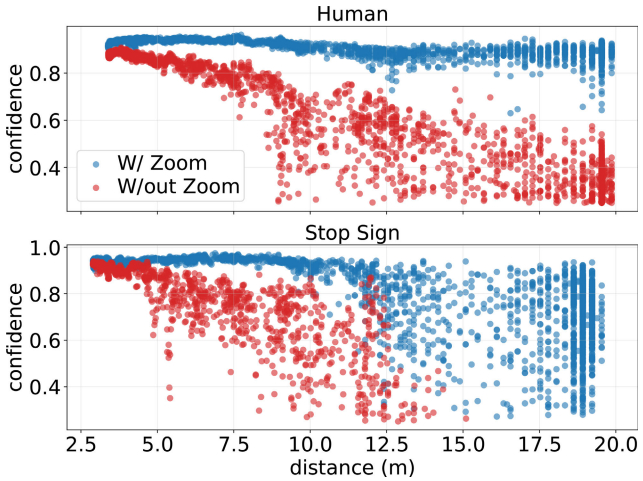



**FIGURE 16.** Effect of adaptive zooming on detection confidence versus distance. Detection confidence for a human (top) and a stop sign (bottom) is compared with and without the adaptive zoom module. The robot was handheld toward and away from fixed targets in an open indoor space, and the perception pipeline was run in both modes. Without zooming, detection confidence degrades sharply beyond 10–12 m due to visual clutter and scale compression. The zooming strategy, by focusing the input on the target and cropping out irrelevant context, maintains high confidence and extends detection range to the camera's depth limit (>20 m).

reidentification approaches such as ByteTrack or SORT [41], [42] remain complementary. These modules would become necessary in the case where another object of the same category physically overlaps within the zoom crop.

### b) Long-Range Detection

We further evaluate the zoom module's impact on detection reliability at extended distances. Two static targets, a human and a stop sign, are placed in an open indoor space. The robot is handheld and moved toward and away from each target, while the onboard perception stack is run twice: once with zooming enabled and once with full-resolution images passed directly to the detector. Detection confidence is recorded for each case.

Fig. 16 summarizes the results. Without zoom, the detector's confidence drops rapidly beyond 10–12 m, often failing to register the target altogether. In contrast, the zoom module enables consistent detections up to the camera's maximum effective range of 20 m. By cropping out irrelevant regions and rescaling the image around the target, zooming mitigates scale compression and distractor interference, boosting confidence and range. While higher resolution full-frame inference (e.g., 640 $\times$ 480) could also improve accuracy, it is not feasible for real-time operation on embedded hardware, as it requires processing four times more pixels than our 320 $\times$ 240 zoom crops. Adaptive zoom, therefore, provides a practical balance: retaining high-frequency detections and extended range within the computational limits of onboard hardware.

Together, these findings highlight that adaptive zoom contributes to robust tracking in two key regimes: it improves resilience to distractors in multiobject environments and extends detection range under scale-constrained conditions.

### 2) DEPTH COMPLETION

The raw stereo depth output from the robot's onboard sensor is often sparse and unreliable, particularly in low-texture regions, near thin structures, or under degraded lighting. These limitations reduce the effectiveness of downstream obstacle avoidance and planning components, especially in fast or cluttered environments.

To address this challenge, NOVA incorporates a disparity-based depth completion module. It leverages monocular priors and disparity cues to fill in missing or noisy depth regions, producing denser and smoother maps. Fig. 17 presents qualitative comparisons between the raw stereo maps and the completed depth output across several representative scenarios. The completed maps better preserve obstacle geometry, enable earlier obstacle detection, and improve motion planning and control safety.

### 3) HISTOGRAM-BASED MODE FILTERING

Accurate target localization requires reliable estimation of depth within the predicted bounding box. However, raw depth values are often noisy and may include background

FIGURE 17. **Qualitative results of depth completion. (Top) RGB inputs. (Middle) Raw stereo depth from the onboard sensor, which suffers from missing or noisy regions, especially around thin structures, glass, and foliage. (Bottom) Completed depth maps produced by our fusion module. The system recovers fine details, such as railings and branches, that are critical for collision avoidance but often missed by stereo matching alone. This enables safe navigation in visually complex environments.**

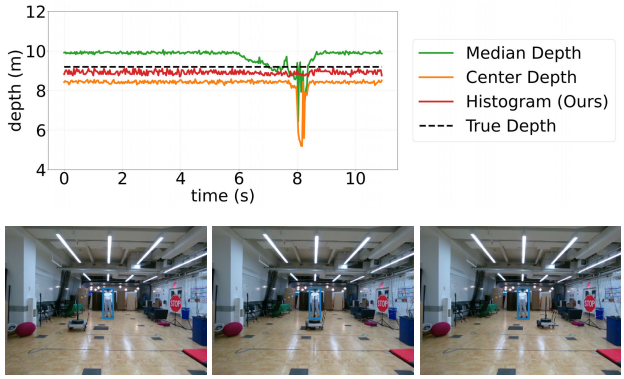



FIGURE 18. **Robustness of depth estimation methods during occlusion. (Bottom) A cart with an obstacle occludes the target mid-experiment, inducing background intrusion in the bounding box. (Top) The plot shows depth estimates over time using three strategies: center-pixel, mean-pixel, and our histogram-based mode filtering. Ground truth is manually measured. The mean estimator exhibits large spikes during occlusion, and the center-pixel fails intermittently. The histogram-based method remains stable and close to the ground truth throughout.**

clutter, particularly in dynamic scenes with partial occlusion. To address this, we introduce a histogram-based mode filtering approach that selects the most frequent depth bin within the bounding box, offering a robust estimate that is resilient to outliers and background interference.

To evaluate this method, we conduct an experiment where the robot hovers in front of a stationary target. A cart carrying a vertical pole is then moved horizontally between the robot and the target, creating a temporary occlusion. During the sequence, we compare three depth estimation strategies: reading the depth value at the center pixel of the bounding box, averaging all valid depth pixels within the box, and applying our proposed histogram-mode filtering.

The ground truth distance to the target is measured manually for reference. As shown in Fig. 18, the center-pixel approach produces unstable estimates during occlusion, while the mean-pixel method exhibits a consistent bias toward the background. In contrast, the histogram-based strategy maintains a stable and accurate estimate throughout the occlusion event, closely matching the ground truth.

This improved robustness in depth estimation directly enhances control performance, enabling more stable tracking under visual uncertainty and in cluttered environments.

## VI. LIMITATIONS AND FUTURE WORKS

NOVA operates under two key assumptions that currently define its functional scope and deployment regime.

First, the system assumes that the target is initially visible in the onboard camera's field of view and belongs to a known object category supported by the detector. While this assumption is reasonable in prompted missions or preconfigured tracking scenarios, it limits applicability in open-world settings where targets may appear later, enter from arbitrary directions, or lack a predefined semantic label. A natural extension is to integrate open-set detection and continuous reidentification, where the detector is not restricted to a fixed set of classes and targets can instead be specified through visual prompts, user feedback, or learned embeddings [77], [78]. Recent work in category-agnostic tracking and self-supervised object discovery further suggests promising pathways to relax this assumption and enable flexible engagement with previously unseen targets [47].

Second, the system relies on the ability to estimate the target's depth using stereo-based depth completion. While this approach is sufficient at close to mid-range (typically up to 30–40 m), it breaks down when the target is detected at longer distances and stereo cues become unreliable. In these conditions, detection may still succeed due to the zoom module, but the depth estimates become noisy or flat, making the x-axis velocity (change in relative depth) unobservable. As a result, the UAV may accelerate rapidly toward the target with limited feedback. In flight experiments, the robot eventually stabilizes once the depth becomes consistent, but the initial approach is often jerky and fast, reflecting the lack of velocity feedback in the unobservable regime.

To relax the reliance on stereo depth, several alternative strategies could be explored. One approach is to augment the sensing stack with additional depth-aware sensors, such as radar or lightweight range finders, which are less sensitive to texture and lighting and can extend the operational range [79], [80], [81]. Another is to infer short-term relative motion using optical flow, either from dense image features or from bounding box displacements of detected objects, including in open-set configurations [82]. A third direction involves registering successive point clouds generated from

depth completion, even if sparse or noisy, to recover frame-to-frame translation [83]. These approaches would allow for estimation of instantaneous robot velocity without relying on world-frame position or persistent maps, and could remain robust under partial feature disruption due to their minimal or stateless memory requirements.

These assumptions are not intrinsic limitations of the overall framework, but they define the boundaries of the current implementation. Future work targeting open-set object understanding and depth-agnostic motion estimation could expand NOVA's applicability to more uncertain, long-range, or semantically unconstrained tracking scenarios.

## VII. CONCLUSION

We presented NOVA, an onboard visual-inertial framework for agile target tracking in unstructured and GPS-denied environments. The system avoids reliance on external localization, global maps, or precomputed scene priors. Instead, it formulates perception, estimation, and control directly in the target's reference frame, using only stereo vision and inertial sensing. The approach combines a lightweight detector with adaptive zooming, depth completion, and visual-inertial state estimation, followed by an NMPC that operates under collision-aware constraints derived from onboard sensing.

We validated NOVA across a series of real-world trials that stress the system across different terrain types, motion regimes, and sensory conditions. The system demonstrated consistent performance in forested, urban, and mixed indoor–outdoor environments, maintaining visual lock, respecting safety distances, and operating without manual tuning. Additional experiments evaluated generalization across viewpoint offsets, repeatability over multiple trials, and robustness under degraded GPS and perceptual ambiguity.

Ablation studies confirmed the contribution of individual components, such as adaptive zooming for long-range detection and identity preservation. The system's limitations were also analyzed, including assumptions about initial visibility, semantic category constraints, and the reliance on stereo-based depth at range.

Together, these results support the conclusion that robust, real-time target tracking can be achieved using only onboard sensing, even in the absence of external infrastructure or structured environments. Future works will focus on relaxing current assumptions, extending to open-set target representations, and enabling perception-driven control under long-range uncertainty.

## REFERENCES

[1] S. Wu, R. Li, Y. Shi, and Q. Liu, "Vision-based target detection and tracking system for a quadcopter," *IEEE Access*, vol. 9, pp. 62043–62054, 2021.

[2] M. Xu, A. Hu, and H. Wang, "Visual-impedance-based human–robot cotransportation with a tethered aerial vehicle," *IEEE Trans. Ind. Informat.*, vol. 19, no. 10, pp. 10356–10365, Oct. 2023.

[3] A. Hu, M. Xu, H. Wang, and H. Castañeda, "Vision-based impedance control of an aerial manipulator using a nonlinear observer," *IEEE Trans. Autom. Sci. Eng.*, vol. 20, no. 2, pp. 1441–1451, Apr. 2023.

[4] J. Gu, T. Su, Q. Wang, X. Du, and M. Guizani, "Multiple moving targets surveillance based on a cooperative network for multi-UAV," *IEEE Commun. Mag.*, vol. 56, no. 4, pp. 82–89, Apr. 2018.

[5] N. Bashir, S. Boudjit, and S. Zeadally, "A closed-loop control architecture of UAV and WSN for traffic surveillance on highways," *Comput. Commun.*, vol. 190, pp. 78–86, Jun. 2022.

[6] H. Huang, A. V. Savkin, and W. Ni, "Online UAV trajectory planning for covert video surveillance of mobile targets," *IEEE Trans. Autom. Sci. Eng.*, vol. 19, no. 2, pp. 735–746, Apr. 2022.

[7] L. Quan et al., "Formation flight in dense environments," *CoRR*, 2022.

[8] G. A. Di Caro and A. W. Z. Yousaf, "Multi-robot informative path planning using a leader-follower architecture," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 10045–10051.

[9] T. Miki, P. Khrapchenkov, and K. Hori, "UAV/UGV autonomous cooperation: UAV assists UGV to climb a cliff by attaching a tether," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 8041–8047.

[10] G. Niu, Q. Yang, Y. Gao, and M.-O. Pun, "Vision-based autonomous landing for unmanned aerial and ground vehicles cooperative systems," *IEEE Robot. Autom. Lett.*, vol. 7, no. 3, pp. 6234–6241, Jul. 2022.

[11] M. Demirhan and C. Premachandra, "Development of an automated camera-based drone landing system," *IEEE Access*, vol. 8, pp. 202111–202121, 2020.

[12] P. Vlantis, P. Marantos, C. P. Bechlioulis, and K. J. Kyriakopoulos, "Quadrotor landing on an inclined platform of a moving ground vehicle," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2015, pp. 2202–2207.

[13] F. Chaumette and S. Hutchinson, "Visual servo control. II. Advanced approaches," *IEEE Robot. Autom. Mag.*, vol. 14, no. 1, pp. 109–118, Mar. 2007.

[14] S. Hutchinson, G. D. Hager, and P. I. Corke, "A tutorial on visual servo control," *IEEE Trans. Robot. Autom.*, vol. 12, no. 5, pp. 651–670, May 1996.

[15] S. Cho and D. H. Shim, "Sampling-based visual path planning framework for a multirotor UAV," *Int. J. Aeronaut. Space Sci.*, vol. 20, no. 3, pp. 732–760, Sep. 2019.

[16] A. A. Oliva, E. Aertbeliën, J. De Schutter, P. R. Giordano, and F. Chaumette, "Towards dynamic visual servoing for interaction control and moving targets," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2022, pp. 150–156.

[17] S. Raj, P. R. Giordano, and F. Chaumette, "Appearance-based indoor navigation by IBVS using mutual information," in *Proc. 14th Int. Conf. Control, Autom., Robot. Vis. (ICARCV)*, Nov. 2016, pp. 1–6.

[18] P. Geneva, K. Eckenhoff, W. Lee, Y. Yang, and G. Huang, "OpenVINS: A research platform for visual-inertial estimation," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 4666–4672.

[19] T. Qin, P. Li, and S. Shen, "VINS-mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Trans. Robot.*, vol. 34, no. 4, pp. 1004–1020, Aug. 2018.

[20] D. Scaramuzza and Z. Zhang, "Visual-inertial odometry of aerial robots," 2019, *arXiv:1906.03289*.

[21] M. Labbé and F. Michaud, "RTAB-map as an open-source LiDAR and visual simultaneous localization and mapping library for large-scale and long-term online operation," *J. Field Robot.*, vol. 36, no. 2, pp. 416–446, Mar. 2019.

[22] W. G. Aguilar, G. A. Rodríguez, L. Álvarez, S. Sandoval, F. Quisaguano, and A. Limaico, "Visual SLAM with an RGB-D camera on a quadrotor UAV using on-board processing," in *Proc. 14th Int. Work-Conf. Artif. Neural Netw. (IWANN)*, 2017, pp. 596–606.

[23] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147–1163, Oct. 2015.

[24] Y. Wang and A. Zell, "Improving feature-based visual SLAM by semantics," in *Proc. IEEE Int. Conf. Image Process., Appl. Syst. (IPAS)*, Dec. 2018, pp. 7–12.

[25] T. Tzanetos et al., "Ingenuity Mars helicopter: From technology demonstration to extraterrestrial scout," in *Proc. IEEE Aerosp. Conf. (AERO)*, Mar. 2022, pp. 1–19.

[26] J. Balaram, M. Aung, and M. P. Golombek, "The ingenuity helicopter on the perseverance rover," *Space Sci. Rev.*, vol. 217, no. 4, p. 56, Jun. 2021.

[27] S. Withrow, W. Johnson, L. A. Young, H. Cummings, J. Balaram, and T. Tzanetos, "An advanced Mars helicopter design," in *Proc. ASCEND*, Nov. 2020, p. 4028.

[28] H. F. Grip et al., "Flying a helicopter on mars: How ingenuity's flights were planned, executed, and analyzed," in *Proc. IEEE Aerosp. Conf. (AERO)*, Mar. 2022, pp. 1–17.

[29] K. Zhang, Y. Shi, and H. Sheng, "Robust nonlinear model predictive control based visual servoing of quadrotor UAVs," *IEEE/ASME Trans. Mechatronics*, vol. 26, no. 2, pp. 700–708, Apr. 2021.

[30] D. Guo and K. K. Leang, "Image-based estimation, planning, and control for high-speed flying through multiple openings," *Int. J. Robot. Res.*, vol. 39, no. 9, pp. 1122–1137, Aug. 2020.

[31] P. Serra, R. Cunha, T. Hamel, D. Cabecinhas, and C. Silvestre, "Landing of a quadrotor on a moving target using dynamic image-based visual servo control," *IEEE Trans. Robot.*, vol. 32, no. 6, pp. 1524–1535, Dec. 2016.

[32] D. Zheng, H. Wang, J. Wang, S. Chen, W. Chen, and X. Liang, "Image-based visual servoing of a quadrotor using virtual camera approach," *IEEE/ASME Trans. Mechatronics*, vol. 22, no. 2, pp. 972–982, Apr. 2017.

[33] J. Thomas, G. Loianno, K. Daniilidis, and V. Kumar, "Visual servoing of quadrotors for perching by hanging from cylindrical objects," *IEEE Robot. Autom. Lett.*, vol. 1, no. 1, pp. 57–64, Jan. 2016.

[34] B. Ma et al., "Target tracking control of UAV through deep reinforcement learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 6, pp. 5983–6000, Jul. 2023.

[35] A. Dionigi, M. Leomanni, A. Saviolo, G. Loianno, and G. Costante, "Exploring deep reinforcement learning for robust target tracking using micro aerial vehicles," in *Proc. 21st Int. Conf. Adv. Robot. (ICAR)*, Dec. 2023, pp. 506–513.

[36] Y. Mao, F. Gao, Q. Zhang, and Z. Yang, "An AUV target-tracking method combining imitation learning and deep reinforcement learning," *J. Mar. Sci. Eng.*, vol. 10, no. 3, p. 383, Mar. 2022.

[37] W. Zhang, K. Song, X. Rong, and Y. Li, "Coarse-to-fine UAV target tracking with deep reinforcement learning," *IEEE Trans. Autom. Sci. Eng.*, vol. 16, no. 4, pp. 1522–1530, Oct. 2019.

[38] C. Min et al., "Autonomous driving in unstructured environments: How far have we come?" 2024, *arXiv:2410.07701*.

[39] R. Khanam and M. Hussain, "YOLOv11: An overview of the key architectural enhancements," 2024, *arXiv:2410.17725*.

[40] Y. Zhao et al., "DETRs beat YOLOs on real-time object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 16965–16974.

[41] Y. Zhang et al., "ByteTrack: Multi-object tracking by associating every detection box," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 1–21.

[42] N. Aharon, R. Orfaig, and B.-Z. Bobrovsky, "BoT-SORT: Robust associations multi-pedestrian tracking," 2022, *arXiv:2206.14651*.

[43] T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, and Y. Shan, "YOLO-world: Real-time open-vocabulary object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 16901–16911.

[44] A. Wang, L. Liu, H. Chen, Z. Lin, J. Han, and G. Ding, "YOLOE: Real-time seeing anything," 2025, *arXiv:2503.07465*.

[45] N. Ravi et al., "SAM 2: Segment anything in images and videos," in *Proc. 13th Int. Conf. Learn. Represent.*, 2025.

[46] S. Liu et al., "Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection," in *Proc. Eur. Conf. Comput. Vis.*, Nov. 2024, pp. 38–55.

[47] A. Saviolo, P. Rao, V. Radhakrishnan, J. Xiao, and G. Loianno, "Unifying foundation models with quadrotor control for visual tracking beyond object categories," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2024, pp. 7389–7396.

[48] S. Hu, Q. Wang, F. Wang, and Y. Li, "Finite-time dynamic visual servo control for quadrotor tracking unknown motion target," *Nonlinear Dyn.*, vol. 113, no. 7, pp. 6959–6977, Apr. 2025.

[49] Y. Kumar and S. B. Roy "Adaptive IBVS based planar non-holonomic target tracking for quadrotors," in *Proc. Int. Conf. Unmanned Aircr. Syst. (ICUAS)*, Jun. 2024, pp. 201–208.

[50] M. Leomanni, F. Ferrante, A. Dionigi, G. Costante, P. Valigi, and M. L. Fravolini, "Quadrotor control system design for robust monocular visual tracking," *IEEE Trans. Control Syst. Technol.*, vol. 32, no. 6, pp. 1995–2008, Nov. 2024.

[51] Y. Jiang, H. Wang, and W. Yu, "Perception-aware model predictive control for target tracking with UAVs," in *Proc. 14th Asian Control Conf.*, 2024, pp. 998–1003.

[52] A. Altan and R. Hacıoğlu, "Model predictive control of three-axis gimbal system mounted on UAV for real-time target tracking under external disturbances," *Mech. Syst. Signal Process.*, vol. 138, Apr. 2020, Art. no. 106548.

[53] A. H. González and D. Odloak, "Robust model predictive controller with output feedback and target tracking," *IET Control Theory Appl.*, vol. 4, no. 8, pp. 1377–1390, Aug. 2010.

[54] N. Sugie, "A model of predictive control in visual target tracking," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-1, no. 1, pp. 2–7, Jan. 2010.

[55] A. Saviolo, N. Picello, J. Mao, R. Verma, and G. Loianno, "Reactive collision avoidance for safe agile navigation," 2024, *arXiv:2409.11962*.

[56] T. Do, L. C. Carrillo-Arce, and S. I. Roumeliotis, "High-speed autonomous quadrotor navigation through visual and inertial paths," *Int. J. Robot. Res.*, vol. 38, no. 4, pp. 486–504, Apr. 2019.

[57] A. Loquercio, A. Saviolo, and D. Scaramuzza, "AutoTune: Controller tuning for high-speed flight," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 4432–4439, Apr. 2022.

[58] M. Kulkarni, B. Moon, K. Alexis, and S. Scherer, "Aerial field robotics," 2024, *arXiv:2401.10837*.

[59] B. Lindqvist, S. S. Mansouri, A.-A. Agha-Mohammadi, and G. Nikolakopoulos, "Nonlinear MPC for collision avoidance and control of UAVs with dynamic obstacles," *IEEE Robot. Autom. Lett.*, vol. 5, no. 4, pp. 6001–6008, Oct. 2020.

[60] Y. Li, G. Lu, D. He, and F. Zhang, "Robocentric model-based visual servoing for quadrotor flights," *IEEE/ASME Transactions on Mechatronics*, vol. 28, no. 4, pp. 2155–2166, Apr. 2023.

[61] B. Zhang et al., "CoNi-MPC: Cooperative non-inertial frame based model predictive control," *IEEE Robot. Autom. Lett.*, vol. 8, no. 12, pp. 8082–8089, Dec. 2023.

[62] B. Zhang, X. Chen, Q. Chen, C. Xu, F. Gao, and Y. Cao, "Global-state-free obstacle avoidance for quadrotor control in air-ground cooperation," *IEEE Robot. Autom. Lett.*, vol. 10, no. 7, pp. 6688–6695, Jul. 2025.

[63] T. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[64] L. Yang et al., "Depth anything V2," 2024, *arXiv:2406.09414*.

[65] L. Oth, P. Furgale, L. Kneip, and R. Siegwart, "Rolling shutter camera calibration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1360–1367.

[66] G. Loianno, M. Watterson, and V. Kumar, "Visual inertial odometry for quadrotors on SE(3)," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2016, pp. 1544–1551.

[67] J. H. E. Cartwright and O. Piro, "The dynamics of Runge–Kutta methods," *Int. J. Bifurcation Chaos*, vol. 2, no. 3, pp. 427–449, 1992.

[68] A. Saviolo, J. Frey, A. Rathod, M. Diehl, and G. Loianno, "Active learning of discrete-time dynamics for uncertainty-aware model predictive control," *IEEE Trans. Robot.*, vol. 40, pp. 1273–1291, 2024.

[69] A. Saviolo, G. Li, and G. Loianno, "Physics-inspired temporal learning of quadrotor dynamics for accurate model predictive trajectory tracking," *IEEE Robot. Autom. Lett.*, vol. 7, no. 4, pp. 10256–10263, Oct. 2022.

[70] L. Bauersfeld, E. Kaufmann, P. Foehn, S. Sun, and D. Scaramuzza, "NeuroBEM: Hybrid aerodynamic quadrotor model," *Robotics, Sci. Syst. Found.*, Jun. 2021.

[71] F. Crocetti, J. Mao, A. Saviolo, G. Costante, and G. Loianno, "GaPT: Gaussian process toolkit for online regression with application to learning quadrotor dynamics," 2023, *arXiv:2303.08181*.

[72] A. Saviolo and G. Loianno, "Learning quadrotor dynamics for precise, safe, and agile flight control," *Annu. Rev. Control*, vol. 55, pp. 45–60, 2023.

[73] E. Kaufmann, L. Bauersfeld, and D. Scaramuzza, "A benchmark comparison of learned control policies for agile quadrotor flight," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2022, pp. 10504–10510.

[74] G. Jocher, J. Qiu, and A. Chaurasia. (Jan. 2023). *Ultralytics YOLO*. [Online]. Available: https://github.com/ultralytics/ultralytics

[75] A. Grunnet-Jepsen, J. N. Sweetser, and J. Woodfill, "Best-known methods for tuning Intel® RealSens™ D400 depth cameras for best performance," Intel Corp., Satan Clara, CA, USA, Tech. Rep., 2018, vol. 1.

[76] R. Verschueren et al., "Acados—A modular open-source framework for fast embedded optimal control," *Math. Program. Comput.*, vol. 14, no. 1, pp. 147–183, Mar. 2022.

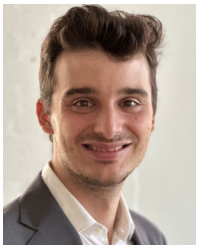
[77] A. Maalouf et al., "Follow anything: Open-set detection, tracking, and following in real-time," *IEEE Robot. Autom. Lett.*, vol. 9, no. 4, pp. 3283–3290, Apr. 2024.

[78] Y. Liu et al., "Opening up open world tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 19023–19033.

[79] M. Nissov, N. Khedekar, and K. Alexis, "Degradation resilient LiDAR-radar-inertial odometry," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2024, pp. 8587–8594.

[80] B. Kim, M. B. Azhari, J. Park, and D. H. Shim, "An autonomous UAV system based on adaptive LiDAR inertial odometry for practical exploration in complex environments," *J. Field Robot.*, vol. 41, no. 3, pp. 669–698, May 2024.

[81] J. Zhang and S. Singh, "Low-drift and real-time LiDAR odometry and mapping," *Auto. Robots*, vol. 41, no. 2, pp. 401–416, Feb. 2017.
[82] A. Alfarano, L. Maiano, L. Papa, and I. Amerini, "Estimating optical flow: A comprehensive review of the state of the art," *Comput. Vis. Image Understand.*, vol. 249, Dec. 2024, Art. no. 104160.
[83] M. Lyu, J. Yang, Z. Qi, R. Xu, and J. Liu, "Rigid pairwise 3D point cloud registration: A survey," *Pattern Recognit.*, vol. 151, Jul. 2024, Art. no. 110408.



**ALESSANDRO SAVIOLO** received the B.Sc. and M.Sc. degrees in computer engineering from the University of Padova, Padua, Italy, in 2018 and 2020, respectively, and the Ph.D. degree in electrical and computer engineering from New York University, Brooklyn, NY, USA, in 2025.

He worked at Flexsight, Padua, from 2020 to 2021, as a Research Engineer on medical and agricultural robotics. He is currently a Senior Autonomy Engineer with Plus, Santa Clara, CA, USA, developing learning-based and model-predictive control methods for large-scale autonomous long-haul trucking. He is a robotics technologist and a researcher. His research focused on adaptive and reactive visual autonomy for safe agile flight in unstructured environments.

**GIUSEPPE LOIANNO** (Member, IEEE) received the Ph.D. degree in robotics from the University of Naples "Federico II", Naples, Italy, in 2014.

He is currently an Associate Professor with the Department of Electrical Engineering and Computer Sciences, Univeristy fo Claifornia Berkeley, and the Director of the Agile Robotics and Perception Laboratory, working on autonomous robots. Prior joining UC Berkeley, he was an Associate Professor with New York University (NYU), New York, NY, USA, and before that a Postdoctoral Researcher, Research Scientist, and team leader with the GRASP Laboratory, University of Pennsylvania, Philadelphia, PA, USA. He has authored or coauthored more than 70 conference papers, journal papers, and book chapters. His research interests include perception, learning, and control for autonomous robots.

Dr. Loianno was the recipient of the NSF CAREER Award in 2022 and DARPA Young Faculty Award in 2022. He was also the recipient of the IROS Toshio Fukuda Young Professional Award in 2022, Conference Editorial Board Best Associate Editor Award at ICRA 2022, Best Reviewer Award at ICRA 2016, and he was selected as Rising Star in AI from KAUST in 2023. He is also currently the Co-Chair of the IEEE RAS Technical Committee on Aerial Robotics and Autonomous Aerial Vehicles. He was the General Chair of the IEEE International Symposium on Safety, Security and Rescue Robotics (SSRR) in 2021 as well as Program Chair in 2019, 2020, and 2022. His work has been featured in a large number of renowned international news and magazines.