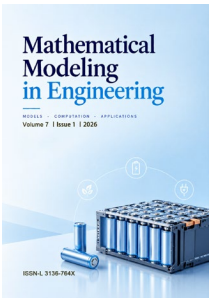
UNIVERSITY OF WEST ATTICA
eISSN: 3136-764X

Mathematical Modeling in Engineering

Vol. 7, No. 2, April, 2026



Bayesian Temporal Fusion Transformer Model for Thermal Target Following in Infrared Panoramic Sequences

Tasos Chatzopoulos[a,*], Xenofon Gavalas[a]

[a] *Department of Computer Engineering, University of Peloponnese, Tripoli, 22100, Greece*

Received 28 January 2026; Revised 25 April 2026; Accepted 30 April 2026; Published 09 May 2026

Abstract

Thermal target following in infrared panoramic inspection sequences is challenging because small angular targets, intermittent heat distractors, panorama-boundary crossings, and robot rotation produce ambiguous observations. This study proposes a Bayesian Temporal Fusion Transformer (B-TFT) that combines wrapped panoramic descriptors, robot-motion context, gated temporal fusion, and posterior uncertainty within a shared representation for tracking and control. Candidate thermal regions are encoded with angular position, scale, contrast, edge, and background statistics; these observations are fused with velocity, yaw rate, route context, and previous commands. A Bayesian projection head estimates the target-state distribution, while predictive variance reduces command aggressiveness and widens the next-frame search range when evidence is weak. Evaluation used 42.6 h of long-wave infrared data comprising 128,400 annotated panoramic frames and 15 target categories collected in corridor, substation, tunnel, and outdoor night routes. Against DeepSORT-Thermal, ConvLSTM, Kalman-Transformer, and deterministic TFT baselines, B-TFT achieved a 94.7% following success rate and a median center error of 3.8 pixels. Expected calibration error decreased from 0.087 for deterministic TFT to 0.031, and average recovery after target reappearance required 0.76 s. End-to-end onboard inference required 31.4 ms per frame. These results indicate that motion-aware Bayesian temporal fusion can preserve identity under thermal ambiguity and provide a calibrated signal for conservative inspection-robot control.

Keywords: Infrared Panoramic Sequence, Thermal Target Following, Bayesian Temporal Fusion Transformer, Inspection Robot, Uncertainty Calibration

1. Introduction

Infrared panoramic imaging has been applied to perception problems of inspection robots in substations, tunnels, factory corridors and other areas, and can effectively overcome challenges such as poor or absent visible light [1]. Therefore, the system needs to keep a heat-emitting object or area within a certain distance of the moving robot, avoid obstacles, and maintain coverage of the inspection area [2]. It is not a typical visual tracking problem because of the wrapped spatial coordinates in a panoramic image, the change in the apparent target width with robot yaw, and many thermal objects with weak texture, even when their temperatures differ considerably from that of the background [3]. In the maintenance inspection, the objects of monitoring are not limited to people and vehicles; heated bearings, cable joints, transformer bushings, pipe leakage zones, mobile workers, and temporary thermal reflections may also be included [4]. Robot platforms are saving sequences of continuous thermal panoramas more frequently than narrow-field clips, and now temporal reasoning is required [5]. A model that has achieved the goal but fails to output the confidence

* Corresponding author: tasos.cha@ce.uop.gr

https://doi.org/10.68406/mme.2026.vol7iss2nm4:41-53

➢ This is an open access article under the CC BY-NC-SA 4.0 license (https://creativecommons.org/licenses/by/4.0/).

© Authors retain all copyrights.

of the decision may still issue an unsafe motion command when the thermal signature is partially occluded, merged with background equipment, or distorted by robot vibration [6].

Traditional infrared tracking commonly combines segmentation, hand-crafted thermal descriptors, and recursive state estimation, which remains attractive for lightweight deployment [7]. Its assumptions become fragile during panoramic patrols when thermal thresholds drift, the robot turns sharply, or several heat sources occupy adjacent angular regions [8]. Deep detectors and Siamese trackers strengthen feature representation, but a confidence score attached to the most likely box does not describe a posterior distribution over future target states [9]. The proposed work therefore focuses on an engineering gap between temporal perception and control: heterogeneous thermal and kinematic variables must be fused across a wrapped sequence, and uncertainty must affect the following command rather than remain a passive output. This requirement motivates a Bayesian temporal-fusion architecture with scene-dependent variable selection and calibrated state prediction.

Temporal fusion provides a suitable basis for panoramic sequence modelling because static route attributes and time-varying observations can be combined while retaining interpretable variable and temporal weights [10]. Probabilistic tracking additionally requires an explicit representation of uncertainty when occlusion, thermal clutter, or motion blur weakens the observation [11]. The proposed B-TFT addresses these requirements through four connected elements. First, target descriptors are expressed in wrapped angular coordinates so that a panorama seam does not create a false physical discontinuity. Second, thermal descriptors, robot velocity, yaw rate, route context, and previous commands enter one gated temporal representation. Third, a Bayesian projection head returns a posterior distribution over angular center, vertical center, scale, and short-horizon offset. Fourth, posterior variance is passed to the controller to reduce aggressive yaw correction and enlarge the search range during ambiguous intervals. The model is designed for onboard operation and is evaluated under partial occlusion, heat distractors, temperature drift, seam crossing, and real-time constraints. Its novelty therefore lies in the deployment-oriented integration of panoramic geometry, heterogeneous temporal fusion, calibrated uncertainty, and risk-aware following control.

The remainder of the paper is organized as follows. Section 2 reviews infrared target tracking and probabilistic temporal learning, with emphasis on the limitations that arise in panoramic robot inspection. Section 3 presents the wrapped observation representation, gated temporal fusion, Bayesian prediction head, training objective, and following-control interface. Section 4 describes the dataset and route-level protocol before evaluating tracking accuracy, uncertainty calibration, robustness, ablation behavior, cross-environment transfer, latency, power, and closed-loop command quality. Section 5 summarizes the contribution, identifies the present limitations, and outlines future development toward multimodal and adaptive inspection systems.

2. Related Work

2.1. Infrared Target Tracking for Inspection Robots

For a long time, infrared target tracking has been based on the idea that thermal contrast can separate the target from the environment more easily than RGB images are affected by visible texture [12]. In actual robot inspection, this assumption is only partially valid because equipment housings, cable terminals and heat-retaining walls may produce thermal patterns similar to those of the target during the short window of the patrol [13]. A typical approach is to first perform background subtraction, local contrast enhancement and blob association, and then use a Kalman or particle filter to maintain the smoothness of the motion of the target hypothesis [14]. The above methods have a relatively small computational cost, but they often lose the trajectory when the target moves out of the panorama or when the robot turns too rapidly for the update rate of the observation filter [15].

Learning-based infrared trackers replace fixed thermal thresholds with representations learned from annotated sequences. Siamese matching, correlation-filter networks, and detector-tracker pipelines can improve localization when the target shape is stable, yet many still treat a frame pair as the primary evidence source. An inspection robot also provides longer-term cues, including repeated heat signatures, platform yaw, recent commands, and predictable background motion. These variables are necessary because identical image displacement can arise from physical target motion or rotation of the robot. Panoramic geometry creates an additional constraint: the left and right image borders correspond to neighboring azimuths rather than a physical break. A tracker operating only in Euclidean image coordinates may therefore interpret a valid seam crossing as disappearance and re-entry. Representing horizontal position as a wrapped angle and conditioning displacement on robot yaw provides a more consistent basis for identity preservation across the panorama boundary.

2.2. Temporal Fusion and Bayesian Sequence Learning

Temporal deep learning methods provide a route for combining image-derived variables with platform motion, and recurrent models were among the earliest choices for thermal sequence prediction [16]. ConvLSTM-based trackers can learn short-range spatiotemporal continuity, but they often compress all variables into a hidden state whose contribution

is difficult to interpret. Attention models relax this bottleneck by assigning weights to relevant time steps and variables, which is valuable in panoramic inspection because the useful cue may be a target contrast pulse several frames before occlusion rather than the current low-contrast frame [17]. The Temporal Fusion Transformer is particularly relevant because it was designed to integrate static covariates, known future inputs, observed historical inputs, and interpretable temporal attention in a single framework [18].

Probabilistic tracking adds another requirement: the model should not only estimate where the target is, but also express how uncertain the estimate is when visual evidence becomes weak. Bayesian neural networks, Monte Carlo dropout, deep ensembles, and evidential prediction heads have been used to represent epistemic or aleatoric uncertainty in perception systems [19]. In mobile robot following, this information can affect speed limitation, re-detection behavior, and handover between perception and control. The research gap addressed here is the lack of a panoramic thermal following model that combines TFT-style variable fusion with Bayesian posterior calibration and evaluates the resulting uncertainty under real inspection-like thermal ambiguity. A second gap is methodological: uncertainty is often reported as an isolated calibration index, whereas the following task requires uncertainty to influence the downstream action. If uncertainty rises before a visible tracking failure, the robot can reduce speed, enlarge the search gate, or delay aggressive yaw correction. If uncertainty remains low during a true target transition, the controller can continue following without unnecessary oscillation. This paper therefore treats calibration and closed-loop behavior as connected engineering measures.

3. Bayesian TFT Model for Thermal Target Following

3.1. Panoramic Thermal Observation Representation

The proposed model treats each panoramic infrared frame as a wrapped thermal observation coupled with robot motion variables. A frame-level encoder first extracts target-centered descriptors from candidate thermal regions, including center angle, vertical position, angular width, mean target temperature, local contrast, edge sharpness, and a compact background heat histogram. These descriptors are not used as isolated detections; they are aligned into a temporal window so that the model can distinguish genuine target displacement from apparent displacement caused by robot yaw. The input vector at time step t is denoted by x_t, and the target state y_t contains horizontal angular center, vertical center, scale, and short-horizon following offset. Robot velocity, yaw rate, and recent control commands are included as observed context variables rather than external corrections. Their joint representation is specified by the observation equation below.

$$x_t = [z_t, v_t, \omega_t, u_{t-1}, c_t] \tag{1}$$

In this equation, z_t denotes the image-derived thermal descriptor, v_t is the linear velocity, ω_t is the yaw rate, u_{t-1} is the previous control command, and c_t is a compact scene-condition vector. The observation window length L is selected to cover 2.4 s of robot motion, which is long enough to include common occlusion intervals without forcing the attention block to model full-route history. Figure 1 presents the complete system workflow from panoramic acquisition and candidate extraction to Bayesian sequence inference and following-control output.

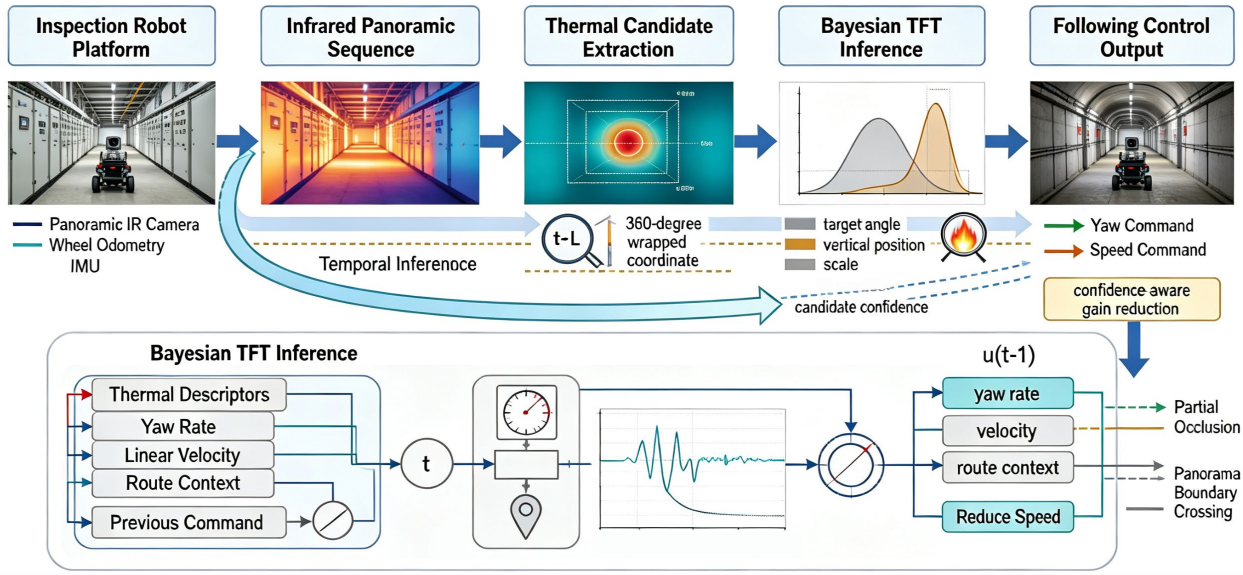


Figure 1. Workflow from infrared panoramic acquisition and candidate extraction to Bayesian TFT inference and robot following control

The candidate extraction stage deliberately keeps multiple thermal regions when their scores are close, because early pruning can remove the correct target in scenes containing hot pipes or reflective metal surfaces. The subsequent sequence model receives candidate-level features together with robot motion, allowing it to reject a distractor if its displacement is inconsistent with the platform trajectory. This arrangement shifts part of the ambiguity resolution from frame-level thresholding to temporal inference, which is more appropriate for inspection routes where the robot observes the same object from several successive viewpoints.

3.2. Gated Temporal Fusion and Bayesian Prediction

After variable selection and static enrichment, the model compresses the wrapped observation history into a temporally fused hidden state. The following temporal-fusion equation makes this mapping explicit while retaining the static route descriptor as a separate input.

$$h_t = TFT_\theta(x_{t-L:t}, s) \tag{2}$$

The temporal-fusion equation produces h_t from the selected history and the static route descriptor. Its TFT block contains gated residual networks, variable selection layers, static enrichment, and interpretable multi-head temporal attention. In this implementation, static enrichment does not refer to a fixed environment map; it refers to slowly varying route attributes, including scene type, expected thermal clutter level, and robot camera mounting height. The variable selection layer suppresses descriptors that are unreliable in a particular scene, such as target edge sharpness during motion blur or raw temperature during outdoor wind cooling. The resulting hidden representation h_t is passed to a Bayesian prediction head so that the model returns a distribution instead of a single point estimate. The gating mechanism also prevents a single noisy variable from dominating the hidden state over the whole temporal window. For example, the temperature descriptor may be highly informative during a cold night patrol, but less reliable inside a substation where energized equipment creates several stable high-temperature regions. Conversely, yaw rate becomes more informative when a target crosses the panorama boundary because the expected wrapped displacement can be inferred from the robot's rotation. The attention block therefore acts on a representation already filtered by scene-dependent relevance.

Uncertainty enters the prediction stage through Bayesian marginalization rather than through a confidence value appended after inference. The predictive distribution is written as the next equation, which integrates over plausible projection parameters conditioned on the training data.

$$p(y_{t+\tau} \mid x_{1:t}) = \int p(y_{t+\tau} \mid h_t, \theta) p(\theta \mid D) d\theta \tag{3}$$

Rather than serving as a deterministic mapping, this predictive equation is implemented through a variational posterior over the final projection parameters and a heteroscedastic likelihood for the target state. During training, the objective combines negative log likelihood, a Kullback-Leibler regularization term, and a smoothness penalty on the expected target displacement. During inference, multiple posterior samples generate an empirical predictive distribution. The control command uses the predictive mean as the tracking target and uses predictive variance as a speed-reduction term when the model is uncertain. Figure 2 is the detailed B-TFT architecture, including descriptor streams, gated fusion, temporal attention, Bayesian posterior sampling, and control gating. The posterior variance is not treated as a cosmetic confidence display. It directly modifies the command by reducing the yaw correction gain when uncertainty is high, which helps avoid overreaction to a temporary heat distractor. The same uncertainty also enlarges the next-frame candidate search range, giving the perception module a better chance to recover after a short disappearance. This coupling between probabilistic perception and conservative control is central to the proposed design. The encoder and TFT blocks remain deterministic, whereas Bayesian treatment is confined to the final projection head; heteroscedastic likelihood variance represents observation-dependent noise, and posterior projection variability represents model uncertainty. The sampling configuration is fixed before test evaluation so that calibration and runtime are assessed under the same inference procedure.

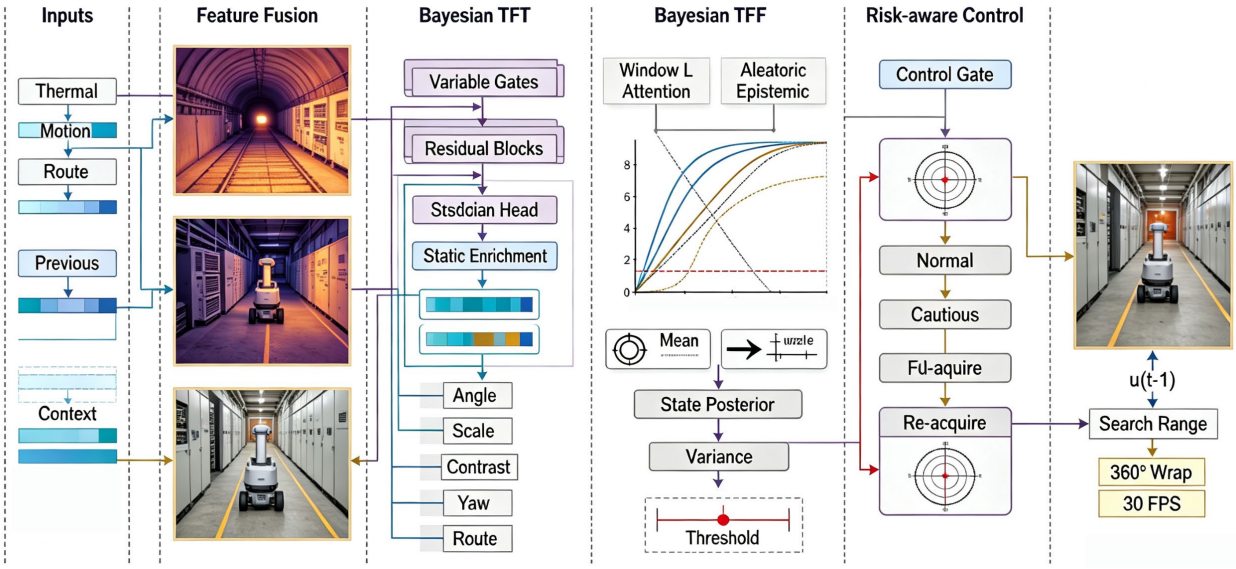


Figure 2. Bayesian Temporal Fusion Transformer (B-TFT) architecture for thermal target following

3.3. Training Objective and Following Interface

At the perception-control boundary, the posterior moments are converted into a conservative following command. The control equation below uses expected target offset for steering and subtracts an uncertainty-dependent term before the command reaches the robot.

$$u_t = K\mathbb{E}[y_{t+1}] - \lambda Var(y_{t+1}) \tag{4}$$

Within the control equation, K maps expected target offset to yaw correction, while λ determines how strongly posterior variance suppresses an aggressive response. Training is performed with sequence mini-batches sampled from complete patrol runs rather than independently shuffled frames. This strategy preserves the correlation between robot motion and panoramic target displacement. Candidate labels are generated from manually corrected thermal bounding boxes converted into angular coordinates, and short disappearance intervals are retained instead of removed, because the model must learn to recover from missing evidence. The loss is weighted by target size and thermal contrast so that small targets and low-contrast targets are not dominated by visually easy samples. No bibliographic claims are introduced in this chapter because the section describes the model and implementation proposed in this paper. The expected offset is expressed in normalized angular coordinates, and the uncertainty-adjusted yaw and speed commands are clipped to the admissible robot limits before actuation. This conservative gating is an empirical safety mechanism rather than a formal closed-loop stability guarantee.

The optimizer uses warm-up followed by cosine decay, and training windows are sampled with higher probability near boundary crossings, occlusion starts, and target reappearance points. These sampling rules increase exposure to the events that most often cause following failure while leaving the original temporal order intact. Validation is performed at the route level rather than by random frame split, which prevents nearly identical neighboring frames from appearing in both training and validation partitions. The following interface receives the posterior mean, posterior variance, and a target-visibility flag derived from the candidate stream. When the variance is low, the controller uses a normal yaw correction gain and keeps the target near the center of the thermal panorama. When the variance exceeds the caution threshold, the controller reduces yaw acceleration and allows a wider angular tolerance, preventing the robot from chasing a transient hot distractor. This design keeps perception and control numerically simple while allowing uncertainty to affect real motion.

4. Experimental Data and Analysis

4.1. Dataset and Evaluation Protocol

The experimental dataset was collected from four inspection environments: indoor corridors, a power substation mock route, a concrete service tunnel, and an outdoor night patrol path. It contained 42.6 h of long-wave infrared panorama, 128,400 annotated frames, and 15 target categories including personnel, hot joints, overheated bearings, small service vehicles, valves, cable clusters, and temporary heat sources. Thermal-robot evaluation protocols generally need both stable targets and controlled background heat disturbances to reveal whether a tracker preserves identity rather than merely detecting the warmest region [20]. The present protocol therefore separated normal sequences from stress-test sequences with blur, partial occlusion, target disappearance, and heat distractors. Table 1 lists the main protocol settings

and representative numerical values used for the following experiments. The camera produced a wrapped panoramic strip at 30 frames per second, and each sequence was synchronized with wheel odometry and inertial yaw measurements. The annotation format stored target center angle, vertical center, target scale, visibility state, and distractor proximity. To avoid inflated performance from adjacent frames, training, validation, and test partitions were assigned by complete patrol run rather than by random frame, a separation strategy used to limit temporal leakage in sequence evaluation [21]. Normal-route evaluation measured ordinary following behavior, while stress-test evaluation measured the ability to maintain identity when the target was small, partly hidden, or visually similar to nearby heat sources. The evaluation code treated a following attempt as successful only when the target remained within the allowed angular band and the identity was not exchanged with a distractor during the same interval. A short re-detection grace period was permitted after full occlusion, but the event was counted as a failure if the controller issued a yaw command toward an incorrect thermal region. This rule made the metric stricter than frame accuracy alone and better aligned with inspection safety.

Table 1. Evaluation protocol and representative quantitative settings

Protocol Block	Panoramic Thermal Setting	Measured Variable	Representative Value
Indoor corridor patrol	360-degree long-wave infrared sequence with reflective wall sections	Frame count / mean speed	34,800 / 0.62 m/s
Power substation patrol	Static hot equipment background with moving robot platform	Target classes / ambient range	5 / 18.4-31.7 C
Tunnel inspection	Low-contrast cylindrical panorama and intermittent occlusion	Occlusion ratio / sequence length	0.28 / 19.6 min
Outdoor night route	Wide thermal field with wind-driven background drift	Mean target size / route distance	41 px / 2.8 km
Stress-test subset	Artificial blur, heat distractors, and target disappearance windows	Distractor count / blur level	9 per min / sigma 1.6

The route-level partition assigned 70% of patrol runs to training, 10% to validation, and 20% to testing while preserving the relative frequency of the four environments. No test route shared a continuous recording session with a training route. Thermal frames were normalized with per-sensor calibration coefficients, but no per-route temperature equalization was applied, because removing ambient drift would make the stress condition artificially simple. Training samples used 24-frame windows with a 12-frame stride; evaluation advanced one frame at a time so that delay and recovery measurements reflected the online operating cycle. Each baseline received the same candidate regions, robot-motion signals, initialization boxes, and disappearance grace period. Hyperparameters were selected on complete validation routes, after which the chosen configuration was frozen before test evaluation. Five runs with different initialization seeds were used for neural models, and reported rates are averages over those runs. Route-level bootstrap resampling with 1,000 replicates was used to examine the stability of method rankings rather than to treat adjacent frames as independent observations. The primary metric, following success, combined angular containment and identity continuity over an interval. Center error measured image-plane localization, recovery time measured the delay from target reappearance to stable reacquisition, and warning lead time measured the interval between an uncertainty threshold crossing and the first subsequent loss event. Calibration was quantified over ten confidence bins, with bins populated from complete sequences to preserve difficult temporal segments. Runtime was measured on the onboard processor after a two-minute warm-up, and latency included candidate encoding, temporal inference, posterior sampling, and command generation. Power values were averaged over the same route windows rather than read from isolated kernels. These controls made the comparisons internally consistent while retaining the temperature drift, motion transients, and candidate ambiguity encountered during actual patrol. Because adjacent frames are temporally correlated, they are not treated as independent samples for significance testing; complete patrol runs are the independence unit for partitioning and bootstrap resampling. The bootstrap analysis therefore supports ranking stability rather than a frame-level claim of statistical significance.

4.2. Tracking Accuracy Calibration and Robustness

The first comparison evaluated following success rate, center error, scale error, and identity continuity against four baselines: DeepSORT-Thermal, ConvLSTM tracker, Kalman-Transformer, and deterministic TFT. Figure 3(a) shows that B-TFT reached 94.7% following success on normal routes and 89.2% on stress-test routes, whereas deterministic TFT reached 91.5% and 83.6%. The 5.5-point normal-to-stress decrease for B-TFT was smaller than the 7.9-point decrease of deterministic TFT and the 10.6-point decrease of DeepSORT-Thermal. Because all methods were evaluated on identical complete routes, this narrower loss cannot be attributed to a favorable frame subset; it reflects better identity retention when weak observations persisted across several consecutive frames. Figure 3(b) separately reports the center-error distribution. The B-TFT median was 3.8 pixels, compared with 4.7 pixels for deterministic TFT, 5.6 pixels for Kalman-Transformer, and 7.4 pixels for ConvLSTM, while its upper whisker ended at 9 pixels rather than 12-18 pixels for the three learned or hybrid competitors. Tracking evaluation commonly separates localization precision from retention because a method can have a reasonable median error and still suffer control-critical tail events [22]. Here, the compressed upper tail coincided with 18 fewer target-loss events than the strongest baseline in the tunnel subset. Frame

review showed that the gain was concentrated at target reappearance and panorama-boundary crossing: B-TFT preferred candidates whose wrapped displacement agreed with recent yaw motion, whereas the baselines more often selected the strongest instantaneous thermal response. The result links the numerical tail reduction to a specific operating mechanism rather than to a uniform offset in otherwise similar predictions.

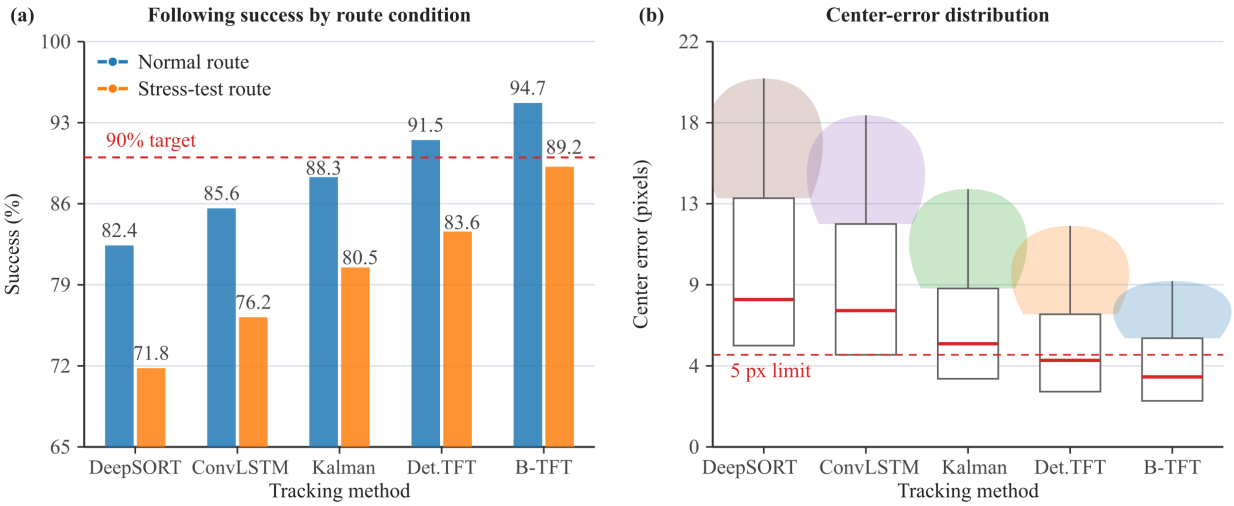


Figure 3. Tracking accuracy comparison: (a) following success under normal and stress-test routes; (b) center-error distribution across all test sequences

Calibration was examined because a following robot must slow down or re-acquire the target when the prediction is unreliable. Figure 4(a) compares reliability curves over ten confidence bins, and B-TFT reduced expected calibration error to 0.031 compared with 0.087 for deterministic TFT and 0.114 for DeepSORT-Thermal. Figure 4(b) shows a different but complementary response: posterior uncertainty increased from 0.42 to 1.86 normalized units as occlusion ratio rose from 0.05 to 0.55, whereas the deterministic spread changed only from 0.38 to 0.52. Li et al. [23] showed that uncertainty-aware adaptation can expose unreliable predictions when scene statistics shift. As shown in the near-diagonal B-TFT curve, most of the high-confidence decisions were correct, and thus the controller did not cause a slowdown in the normal route section. A monotonic rise in uncertainty provided the missing time-to-event information; at about 35% occlusion, uncertainty entered the caution region, and then many substation clips showed an increase in the center error. Probabilistic alarms are only practical if the threshold-crossing event occurs before the damage to be avoided happens [24]. The first crossing had several 30-Hz control updates, and a reduced yaw gain and a wider candidate gate were used in this experiment. High-variance intervals also overlapped with the clips that human reviewers had marked as ambiguous, but the variance was low at clear boundary crossings; therefore, the model reacted to evidence conflict rather than to panoramic wrapping alone. The caution threshold was selected on validation routes and frozen before the test sequences were evaluated, preventing retrospective threshold adjustment to the displayed calibration results.

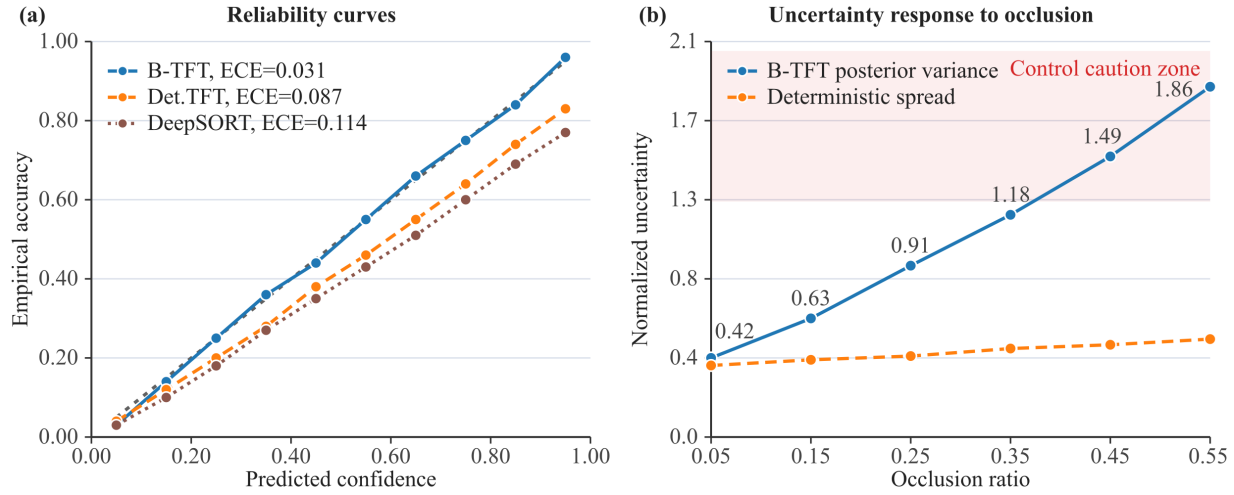


Figure 4. Calibration behavior: (a) reliability curves over confidence bins; (b) posterior variance response under increasing occlusion ratio

Robustness tests were organized around target size, heat contrast, occlusion length, and robot yaw rate. Figure 5(a) presents a heat map of success rate over target size and thermal contrast; B-TFT reached 88% at the 28-pixel, 4.5 C operating point and exceeded 90% through most cells above that diagonal. Below 20 pixels or below 3.0 C, success fell to 42-75%, defining a region in which candidate evidence was too weak for temporal fusion to remove ambiguity consistently. Figure 5(b) resolves the stress response into five normalized axes: B-TFT obtained 0.93 for occlusion tolerance, 0.91 for boundary crossing, 0.88 for distractor rejection, 0.90 for blur resistance, and 0.95 for steady-route continuity. Multidimensional stress tests are used in field robotics because aggregate accuracy can conceal sharply different failure modes [25]. The heat map and radar plot expose two such modes. Small low-contrast targets failed because several candidate regions had similar local statistics, whereas lower-seam targets failed because floor reflection compressed the vertical gradient and destabilized scale. The largest radar separation occurred in distractor rejection and boundary crossing, where motion context and wrapped geometry provide information unavailable to a frame-only score. Blur resistance improved by a smaller margin, suggesting that the temporal model could bridge brief degradation but could not reconstruct spatial detail absent from every frame in the window. These patterns define a usable sensor envelope: targets above 28 pixels with at least 4.5 C contrast support normal following, while observations below that boundary require slower motion and more frequent re-detection.

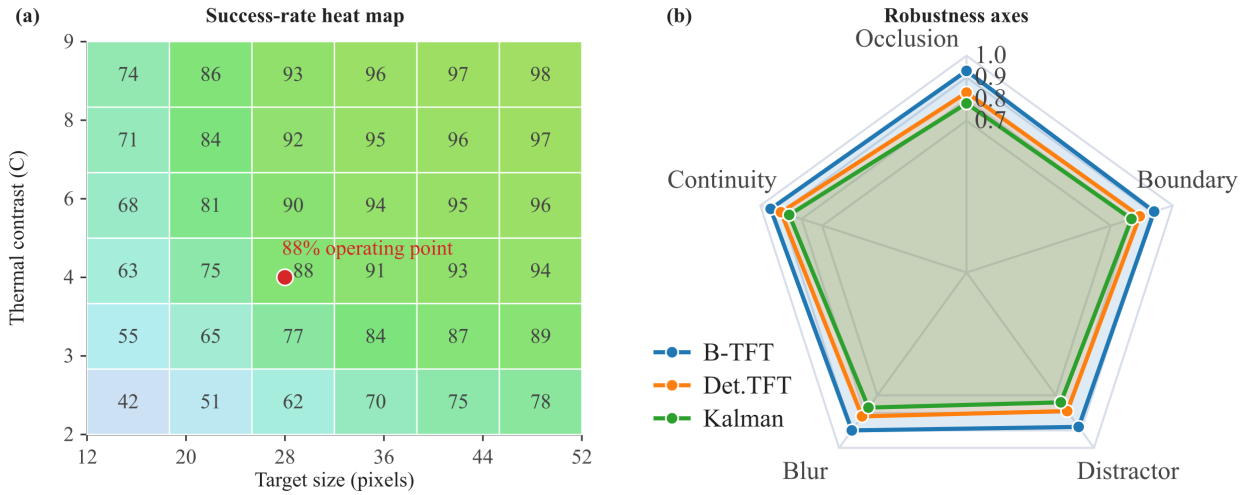


Figure 5. Robustness analysis: (a) success-rate heat map over target size and thermal contrast; (b) radar comparison of five robustness axes

4.3. Ablation Transfer and Closed Loop Deployment

Ablation experiments measured how much each component contributed to tracking and following control. Figure 6(a) shows that removing robot-motion context increased mean center error from 3.8 to 5.1 pixels, while removing the Bayesian head raised recovery time from 0.76 to 1.18 s. The divergence identifies separate roles: motion context mainly corrected panoramic displacement during fast turns, whereas posterior sampling mainly controlled how quickly the system committed to a candidate after occlusion. Figure 6(b) shows that disabling gated variable selection produced the largest success-rate decrease, from 94.7% to 88.5%, even though attention remained active. This result indicates that attention alone could not prevent unstable temperature or edge descriptors from entering the temporal representation. Figure 6(c) maps latency for five window lengths and four hidden widths. The selected 24-frame, 96-unit setting required 31.4 ms per frame, leaving about 1.9 ms within a 30-fps cycle; longer windows or 128-160-unit widths approached or exceeded the available budget. Embedded perception should therefore be judged jointly by accuracy and deadline compliance [26]. Figure 6(d) relates posterior uncertainty to absolute error over 70 observations and yields a Spearman coefficient of 0.72. Selective Bayesian treatment of a prediction head can retain useful uncertainty while avoiding the cost of sampling the complete feature encoder [27]. The scatter widens above the 1.25 caution threshold, but most large errors remain in the high-uncertainty region, which is the asymmetry needed for conservative control. When gating was removed, gross occlusion still raised variance, yet conflicting motion descriptors generated weaker warnings; uncertainty quality therefore depended on the representation passed to the Bayesian head as well as on posterior sampling.

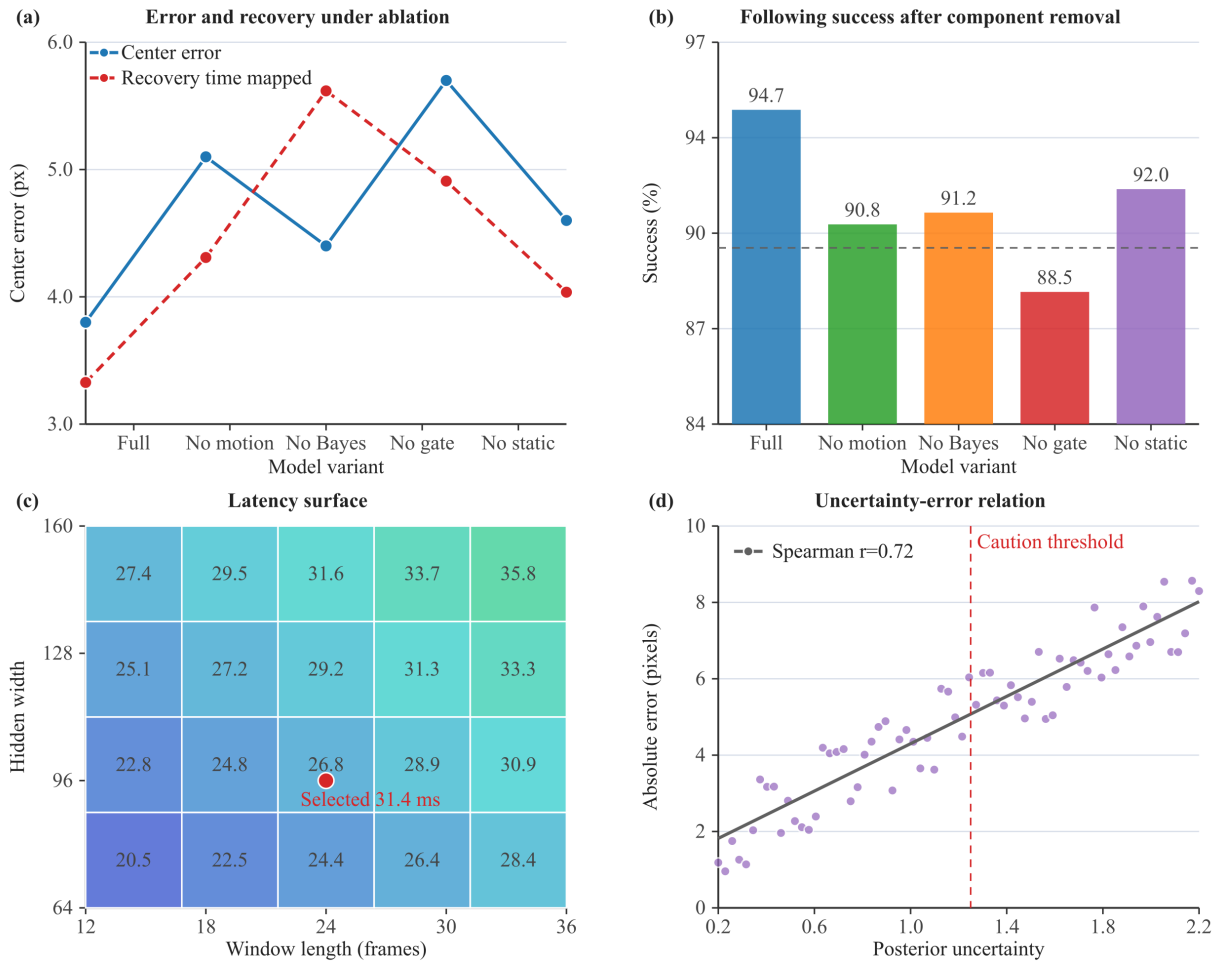


Figure 6. Ablation and efficiency analysis: (a) center error and recovery time under component removal; (b) following success after ablation; (c) latency surface over window length and feature width; (d) uncertainty-error correlation

Generalization was evaluated by training on three environments and testing on the withheld environment. Figure 7(a) shows that substation-to-tunnel transfer produced the largest loss: tunnel success decreased from 92.9% with mixed training to 86.9% when tunnel data were withheld, a 6.0-point gap. The same cross-environment profile also shows smaller but persistent losses of 5.3, 5.4, and 6.4 points for corridor, substation, and outdoor routes, respectively. Infrared domain shift often changes background temperature statistics more strongly than the shape of the tracked target [28]. In the tunnel, elongated wall reflections altered candidate geometry; outdoors, wind introduced intermittent edge blur around small targets. Figure 7(b) measures whether the model exposed those shifts early enough for control action. Uncertainty warnings preceded loss by 0.47-0.62 s across the four withheld environments, whereas visible center-error warnings provided only 0.18-0.24 s. At 30 frames per second, the 0.54-s outdoor uncertainty lead corresponds to roughly 16 control cycles, compared with about six cycles after the center error became visible. The Bayesian component did not remove the transfer loss, but it converted part of an abrupt tracking failure into a detectable pre-failure interval. The remaining gap between environments also shows the limit of route context: a categorical scene descriptor cannot encode reflection patterns or thermal dynamics never represented during training, so deployment in a new facility still benefits from route calibration.

The final experiment examined closed-loop behavior on the inspection robot. Figure 8(a) gives the cumulative distribution of angular following error; 90% of B-TFT commands stayed within 5.8 degrees, compared with 8.9 degrees for deterministic TFT. Figure 8(b) reports the corresponding command dynamics: B-TFT reduced yaw-rate jerk by 21.4% relative to Kalman-Transformer, and the reduction was sustained over the 60-s route rather than confined to isolated turns. Closed-loop tracking assessment must include command smoothness because comparable image-space error can produce different platform vibration and motion blur [29]. Figure 8(c) shows that target-loss events remained at or below 2.1 per 10 min up to 11 thermal distractors per minute, then increased to 3.4 and 4.8 at densities of 13 and 15. The change in slope marks a candidate ambiguity regime in which several heat sources had both similar contrast and similar angular velocity. Figure 8(d) separates the 31.4-ms pipeline into 12.4 ms for encoding, 14.7 ms for TFT inference, 3.1 ms for the Bayesian head, and 1.2 ms for control; the associated component powers sum to 13.6 W. Energy-aware timing is a deployment constraint for autonomous inspection platforms operating on long routes without workstation-class power [30].

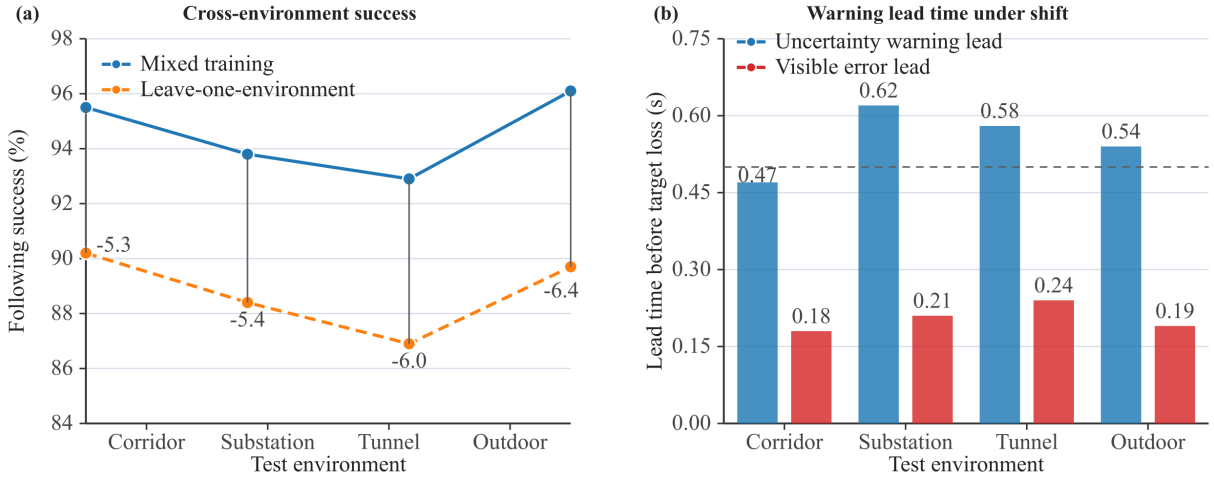


Figure 7. Cross-environment generalization: (a) leave-one-environment following success; (b) lead time of uncertainty warning before target-loss events

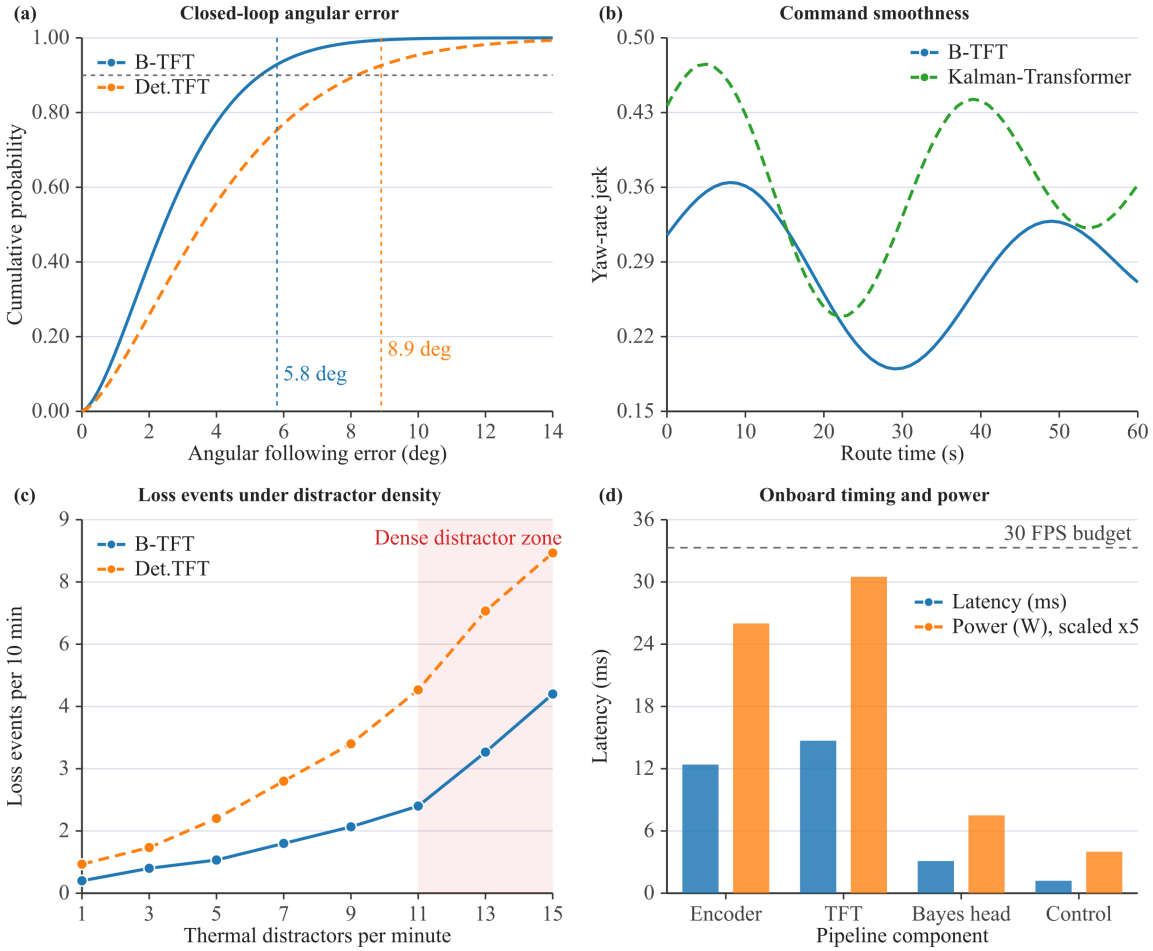


Figure 8. Closed-loop deployment: (a) cumulative angular following error; (b) yaw-rate command smoothness; (c) target-loss events under thermal distractor density; (d) onboard latency and power

The 1.9-ms timing margin was adequate under the reported configuration but leaves little room for a larger candidate set. Frame-by-frame review located residual failures in targets below 20 pixels, disappearances longer than 2.8 s, and same-angular-cell crossings. In the last case, variance rose as intended, yet the posterior mean could still drift toward a hotter distractor when both regions maintained nearly identical motion for more than one second; the controller then reduced yaw gain but could not recover identity until their trajectories separated. Latency and power should therefore be interpreted as end-to-end measurements for the reported onboard configuration, not as hardware-independent complexity guarantees.

5. Conclusion

This paper develops a Bayesian Temporal Fusion Transformer for thermal target tracking and following from infrared panoramic sequences. The proposed pipeline represents target observations in wrapped coordinates, combines thermal and robot-motion variables through gated temporal fusion, and predicts a posterior target state for the following controller. By using predictive variance to moderate commands and adapt candidate search, the method links uncertainty estimation to closed-loop inspection behavior rather than treating confidence as an auxiliary display.The study remains limited by the range of environments and disturbance conditions represented in the dataset. Extreme weather, dense smoke, severe vibration, and long seasonal temperature changes were not included. Deployment on a new robot or facility still requires annotated target boxes, route context, and calibration of the operating threshold. Moreover, Bayesian approximation is restricted to the final prediction head, so uncertainty originating in early feature extraction is represented only indirectly.

Future work will extend the data to more industrial sites and combine infrared panoramas with depth, acoustic, or other complementary measurements. Online adaptation without route-specific labels is also needed to reduce deployment cost. A closer connection between posterior risk and motion planning could coordinate velocity, viewpoint selection, and re-detection behavior, while model compression and uncertainty-preserving distillation may improve suitability for lower-power embedded platforms.

6. Declarations

6.1 Author Contributions

Tasos Chatzopoulos contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. Xenofon Gavalas contributes to validation, analysis, investigation, draft preparation. All authors have read and agreed with the manuscript before its submission and publication.

6.2 Funding and Acknowledgments

This research received no specific financial support from any funding agency.

6.3 Institutional Review Board Statement

Not applicable.

6.4 Informed Consent Statement

Not applicable.

6.5 Data Availability

Not applicable.

6.6 Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

6.7 Declaration of generative AI and AI-assisted technologies

During the preparation of this manuscript, the authors utilized Grammarly for linguistic polishing, logical organization and draft revision. The generative AI tool was not involved in research design, data collection, data analysis, or the equation of core conclusions. All outputs generated by AI were reviewed, revised and verified manually by the authors. The authors take full responsibility for all content of this manuscript.

7. References

[1] Yuan, D., Zhang, H., Shu, X., Liu, Q., Chang, X., He, Z., & Shi, G. (2024). Thermal infrared target tracking: a comprehensive review. IEEE Transactions on Instrumentation and Measurement, 73, 1–19. https://doi.org/10.1109/tim.2023.3338701

[2] Du, Z., Zhang, G., Zhang, Y., Chen, J., & Zhang, X. (2024). Path planning of substation inspection robot based on high-precision positioning and navigation technology. International Journal of Low-Carbon Technologies, 19, 1754–1765. https://doi.org/10.1093/ijlct/ctae125

[3] Liu, W., Zhang, T., Xu, S., Chang, Q., & Cui, Y. (2024). PanoDetNet: multi-resolution panoramic object detection with adaptive

feature attention. IEEE Access, 12, 104300–104316. https://doi.org/10.1109/access.2024.3435764

[4] Li, X., He, Y., Wang, H., Sun, G., Yu, J., Du, X., Hu, M., & Wang, Y. (2024). Thermal inspection of subsurface defects in wind turbine blade segments under the natural solar condition. IEEE Transactions on Industrial Electronics, 71(9), 11488–11497. https://doi.org/10.1109/tie.2023.3335455

[5] Luo, Y., Li, X., Wang, J., & Chen, S. (2024). Clustering and tracking-guided infrared spatial–temporal small target detection. IEEE Transactions on Geoscience and Remote Sensing, 62, 1–20. https://doi.org/10.1109/tgrs.2024.3384440

[6] Huseljic, D., Herde, M., Hahn, P., Müjde, M., & Sick, B. (2025). Systematic evaluation of uncertainty calibration in pretrained object detectors. International Journal of Computer Vision, 133(3), 1033–1047. https://doi.org/10.1007/s11263-024-02219-z

[7] Zha, C., Luo, S., & Xu, X. (2024). Infrared multi-target detection and tracking in dense urban traffic scenes. IET Image Processing, 18(6), 1613–1628. https://doi.org/10.1049/ipr2.13053

[8] Zhao, K., Zhao, H., Wang, Z., Peng, J., & Hu, Z. (2024). Object-preserving siamese network for single-object tracking on point clouds. IEEE Transactions on Multimedia, 26, 3007–3017. https://doi.org/10.1109/tmm.2023.3306490

[9] Xu, X., Zhan, W., Jiang, Y., Zhu, D., Chen, Y., Guo, J., Li, J., & Liu, Y. (2024). A single-frame infrared small target detection method based on joint feature guidance. Complex & Intelligent Systems, 10(4), 5281–5300. https://doi.org/10.1007/s40747-024-01410-6

[10] Ying, C., & Lu, J. (2024). TFEformer: temporal feature enhanced transformer for multivariate time series forecasting. IEEE Access, 12, 153694–153708. https://doi.org/10.1109/access.2024.3480953

[11] Zhao, R., Wang, K., Xiao, Y., Gao, F., & Gao, Z. (2024). Leveraging Monte Carlo dropout for uncertainty quantification in real-time object detection of autonomous vehicles. IEEE Access, 12, 33384–33399. https://doi.org/10.1109/access.2024.3355199

[12] Fişne, A., Yurtsever, M. M. E., & Eken, S. (2024). Energy-efficient buildings with energy-efficient optimized models: a case study on thermal bridge detection. Cluster Computing, 27(9), 12787–12797. https://doi.org/10.1007/s10586-024-04624-y

[13] Sun, Y., Geng, K., Zhuang, W., Yin, G., Chen, X., Wang, J., & Ding, P. (2024). Moving traffic object detection based on Bayesian theory fusion. Automotive Innovation, 7(3), 418–430. https://doi.org/10.1007/s42154-023-00245-0

[14] Xu, L., Liu, D., & Ma, H. (2024). Robot visual tracking model based on improved GOTURN-LD algorithm. IEEE Access, 12, 179981–179996. https://doi.org/10.1109/access.2024.3503674

[15] Lian, S., & Zhang, F. (2024). TDANet: target-directed attention network for object-goal visual navigation with zero-shot ability. IEEE Robotics and Automation Letters, 9(9), 8075–8082. https://doi.org/10.1109/lra.2024.3440100

[16] Xiao, Y., Liu, Z., Yin, H., Wang, X., & Zhang, Y. (2024). STFormer: a dual-stage transformer model utilizing spatio-temporal graph embedding for multivariate time series forecasting. Journal of Intelligent & Fuzzy Systems, 46(3), 6951–6967. https://doi.org/10.3233/jifs-237250

[17] Han, W., Zhu, T., Chen, L., Ning, H., Luo, Y., & Wan, Y. (2024). MCformer: multivariate time series forecasting with mixed-channels transformer. IEEE Internet of Things Journal, 11(17), 28320–28329. https://doi.org/10.1109/jiot.2024.3401697

[18] Ma, S., Zhao, Y. B., Kang, Y., & Bai, P. (2024). Multivariate time-series modeling and forecasting with parallelized convolution and decomposed sparse-transformer. IEEE Transactions on Artificial Intelligence, 5(10), 5232–5243. https://doi.org/10.1109/tai.2024.3410934

[19] Landgraf, S., Wursthorn, K., Hillemann, M., & Ulrich, M. (2024). DUDES: deep uncertainty distillation using ensembles for semantic segmentation. PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science, 92(2), 101–114. https://doi.org/10.1007/s41064-024-00280-4

[20] Jiao, L., Wei, H., & Pan, Q. (2025). Region and sample level domain adaptation for unsupervised infrared target detection in aerial remote sensing images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 18, 11289–11306. https://doi.org/10.1109/jstars.2025.3561737

[21] Su, Y., Zhang, J., & Li, Q. (2024). SCSformer: cross-variable transformer framework for multivariate long-term time series forecasting via statistical characteristics space. Applied Intelligence, 54(24), 12922–12948. https://doi.org/10.1007/s10489-024-05764-9

[22] Chen, Y., Tang, Y., Xiao, Y., Yuan, Q., Zhang, Y., Liu, F., He, J., & Zhang, L. (2024). Satellite video single object tracking: a systematic review and an oriented object tracking benchmark. ISPRS Journal of Photogrammetry and Remote Sensing, 210, 212–240. https://doi.org/10.1016/j.isprsjprs.2024.03.013

[23] Li, G., Chen, Z., Mao, M., Lin, L., & Fang, C. (2024). Uncertainty-aware active domain adaptive salient object detection. IEEE Transactions on Image Processing, 33, 5510–5524. https://doi.org/10.1109/tip.2024.3413598

[24] Wang, J., Zhao, S., & Liang, S. (2024). Augmenting 3-D object detection through data uncertainty-driven auxiliary framework.

IEEE Transactions on Instrumentation and Measurement, 73, 1–14. https://doi.org/10.1109/tim.2024.3398077

[25] Jafari-Tabrizi, A., Gruber, D. P., & Gams, A. (2024). Exploiting image quality measure for automatic trajectory generation in robot-aided visual quality inspection. The International Journal of Advanced Manufacturing Technology, 132(9-10), 4885–4901. https://doi.org/10.1007/s00170-024-13609-5

[26] Zemčík, T., & Kratochvíla, L. (2024). Mobile robot perception system in ROS 2 on embedded computing platforms. IFAC-PapersOnLine, 58(9), 305–310. https://doi.org/10.1016/j.ifacol.2024.07.414

[27] Wang, H., Ma, J., Zhang, J., & Chen, F. (2025). Graphfusion: robust 3d detection via cross-modal graph and uncertainty-aware Bayesian fusion. IEEE Signal Processing Letters, 32, 3645–3649. https://doi.org/10.1109/lsp.2025.3609243

[28] Luo, D., Xiang, Y., Wang, H., Ji, L., & Ye, M. (2025). Knowledge adaptation for cross-domain moving infrared small target detection. IEEE Transactions on Geoscience and Remote Sensing, 63, 1–14. https://doi.org/10.1109/tgrs.2025.3604069

[29] Li, L., Qiang, J., Xia, Y., & Cao, W. (2025). Adaptive dual closed-loop trajectory tracking control for a wheeled mobile robot on rough ground. Nonlinear Dynamics, 113(3), 2411–2425. https://doi.org/10.1007/s11071-024-10350-2

[30] Singh, S., Molla, A. R., Mondal, A., & Dey, S. (2024). Co-designing perception-based autonomous systems on CPU-GPU platforms. IEEE Embedded Systems Letters, 16(4), 357–360. https://doi.org/10.1109/les.2024.3443135