Engineering and Technology Journal e-ISSN: 2456-3358

Volume 11 Issue 07 July-2026, Page No.-10940-10944

DOI: 10.47191/etj/v11i07.17, I.F. – 8.595

Characterizing Confidence Compression in Cross-Domain Thermal Human Detection for Search and Rescue

Nura Mansour Ahmed[1], Ruya Yilmaz[2]

[1]Department of Software Engineering/Computer Engineering, Istanbul Aydin University, Istanbul 34295, Turkey

[2]Faculty of Engineering and Natural Sciences, Computer Engineering Department, Atlas University, Istanbul, Turkey

ABSTRACT: This paper addresses the "thermal crossover" phenomenon in autonomous Search and Rescue (SAR) drones, in which solar-heated objects mimic human body heat and drive high false-alarm rates in single-stage detectors. We propose a two-stage cascade pairing a permissive YOLOv8-Nano scanner with a fine-tuned EfficientNet-B2 refiner adapted to the HIT-UAV aerial thermal benchmark. On a held-out evaluation, the proposed model attains a balanced classification accuracy of about 95% at its operating threshold, substantially reducing false alarms over a single-stage baseline while preserving human recall. Separately, we characterize a Confidence Compression phenomenon: using a ground-level thermal dataset as a reference, we find that under domain mismatch the verifier's output confidence scores are pinned below roughly 0.64 on the 0–1 probability scale. This is a calibration effect—it concerns the confidence values the verifier assigns to each detection, not how often it is correct—and domain-matched adaptation relieves it. An analytical model further projects real-time feasibility on the NVIDIA Jetson Orin Nano.

KEYWORDS: Thermal Clutter Rejection, YOLOv8, EfficientNet-B2, Domain Adaptation, Hard Negative Mining, Jetson Orin Nano, Edge Inference.

I. INTRODUCTION

Search and Rescue (SAR) operations increasingly rely on drone platforms operating in communication-constrained environments [1]. Thermal imaging enables detection in darkness, fog, and smoke, but the "thermal crossover" phenomenon—wherein solar-heated surfaces temporarily match human skin temperature (~33 °C)—causes high false-alarm rates in single-stage detectors, wasting battery resources and degrading mission reliability [2].

Autonomous aerial SAR increasingly depends on thermal infrared sensing because long-wave infrared (LWIR) cameras detect the heat signature of a human body independently of ambient illumination, enabling operation at night and through visual obscurants that defeat RGB cameras [2]. As uncooled microbolometer sensors have become smaller, cheaper, and light enough for small multirotor platforms, thermal-equipped drones have moved from a specialist capability to a practical tool for wide-area search over terrain that is slow or dangerous for ground teams. This shift places the burden of reliability on the on-board detection software rather than the sensor itself.

The move toward autonomy raises the stakes further. When a human pilot monitors a live thermal feed, an occasional false alarm is a minor nuisance the operator filters out instinctively. When the platform plans its own flight path and allocates its own energy, however, every detection becomes an input to an automated decision, and spurious alerts can propagate into wasted maneuvers or premature termination of a promising search leg. Reliable clutter rejection is thus not merely a perception nicety but a precondition for trustworthy autonomous behaviour in the field.

In this setting, the dominant failure mode is not missed detection but false alarm. Thermal clutter rejection—the task of separating genuine human heat signatures from thermally similar background objects—is therefore central to operational SAR. Each false alarm consumes operator attention, diverts flight time, and, in autonomous mission planning, can trigger unnecessary loiter or descent maneuvers that drain the limited energy budget of a battery-powered platform. Because a search may span thousands of frames, even a modest per-frame false-positive rate accumulates into an operationally unacceptable number of spurious alerts over a single sortie.

Existing approaches bifurcate along a speed–accuracy trade-off. Single-stage detectors such as YOLOv8n [3] are embedded-feasible but cannot resolve thermal ambiguity through threshold tuning alone. Two-stage detectors achieve the required precision but exceed drone real-time budgets (~180 GFLOPS, >250 ms latency). Recent models such as LRI-YOLO [5] improve efficiency but retain single-stage logic and do not address thermal crossover by design.

This paper reports three scientific contributions. (i) Confidence Compression characterization: we identify and name a cross-domain phenomenon in which the verifier's output confidence scores (not its accuracy) for thermal human targets compress toward a ceiling of roughly 0.64 on the 0–1 probability scale at aerial altitude, motivating a dedicated verification stage. (ii) Cascade architecture within embedded latency constraints: a coarse-to-fine two-stage pipeline pairing a permissive single-stage scanner with an EfficientNet-B2 verification refiner, substantially reducing per-frame false alarms over single-stage detection. (iii) A domain-adaptation strategy that establishes a replicable protocol for adapting generic image classifiers to the top-down aerial thermal geometry that induces Confidence Compression.

A. Organization of Paper

The remainder of this paper is organized as follows: Section II reviews related work, Section III details the two-stage cascade methodology and training strategy, Section IV presents the results and discussion, and Section V concludes with limitations and future work.

II. RELATED WORK

Thermal imaging enables SAR detection under conditions where RGB cameras fail, including night operations, heavy smoke, and dense fog [2]. Uncooled LWIR microbolometers provide sufficient thermal sensitivity for human detection at the operational altitudes covered by aerial benchmarks such as HIT-UAV, which extends up to 130 m [1]. The critical limitation is the thermal crossover phenomenon: inanimate surfaces including solar-heated rocks, metallic debris, and warm asphalt temporarily reach temperatures that intersect the ~33 °C human skin surface, causing standard threshold-based detection to produce false alarms.

A. Thermal Sensing for Aerial Search and Rescue

Thermal infrared has a long history in SAR because it images the quantity of direct interest—emitted body heat—rather than reflected light. Early airborne systems relied on cooled sensors that were heavy, power-hungry, and costly, restricting thermal search to crewed aircraft. The maturation of uncooled LWIR microbolometer arrays changed this economics: modern sensors operate without cryogenic cooling, weigh tens of grams, and draw little power, which is what makes them viable payloads for small unmanned platforms [2]. Public aerial thermal benchmarks now capture pedestrians from altitudes and oblique angles representative of real drone missions, providing the data needed to train and evaluate detection models under realistic geometry [1].

However, sensing capability alone does not guarantee reliable detection. At altitude, a human occupies few pixels, the viewing angle is close to nadir, and the scene contains many warm objects. These conditions differ sharply from the ground-level, front-facing imagery on which many pedestrian models were originally developed, and the resulting domain gap is a recurring source of degraded performance in aerial thermal detection.

B. The Thermal Crossover Problem

The core difficulty in thermal clutter rejection is thermal crossover: over the course of a day, inanimate surfaces such as sun-warmed rock, asphalt, metal, and machinery pass through the same apparent-temperature range as exposed human skin. During these intervals a purely intensity-based detector cannot separate a person from the background on brightness alone, because the two are, momentarily, radiometrically indistinguishable. Traditional threshold- and histogram-based methods, which assume the target is consistently hotter than its surroundings, break down precisely when a search is most likely to be conducted—during and after daylight hours.

The underlying mechanism is thermal inertia. Surfaces absorb solar radiation through the day and re-emit it slowly after sunset, so the apparent temperature of the background is not fixed but follows a diurnal cycle that repeatedly sweeps through the human-skin range. Emissivity differences compound the ambiguity: two surfaces at the same physical temperature can present different apparent temperatures to the sensor, and a low-emissivity object can register as human-warm even when it is not. Any detector that hopes to be dependable across a real mission window must therefore be robust to a background whose thermal appearance is continually changing.

This motivates a shift from intensity cues to learned spatial and textural features. Deep detectors can, in principle, exploit shape, gait posture, and local context that survive thermal crossover, but doing so at the compute and latency budgets available on an airborne edge processor remains an open engineering problem.

C. Single-Stage Detection at the Edge

Single-stage detectors dominate embedded deployment because they resolve detection in one forward pass, trading some accuracy for speed. The YOLO family is the canonical example; its latest iterations use anchor-free, decoupled detection heads and efficient backbones to achieve real-time throughput at modest parameter counts [3]. Lightweight infrared-specific variants further reduce cost through partial and group convolutions tailored to UAV deployment [5]. These models are well suited to the coarse scanning role—finding every candidate quickly—but their single decision boundary makes them prone to the false alarms that thermal crossover induces when operated at the permissive thresholds needed to preserve recall.

YOLOv8n represents the current state-of-the-art for edge-constrained single-stage detection, achieving 8.9 GFLOPS with 3.2M parameters through its anchor-free, decoupled detection head and C2f backbone [3]. Despite this efficiency, false positive rates as high as 28% were observed for the single-stage baseline in this study's thermally cluttered evaluation. Two-stage models achieve higher discrimination precision but are infeasible at drone scale: Faster R-CNN (ResNet-50) runs at ~180 GFLOPS and >250 ms latency on embedded hardware. Our work introduces a dedicated hard-negative verification stage to bypass this trade-off.

D. Two-Stage and Cascade Approaches

Two-stage detectors such as Faster R-CNN separate region proposal from classification and generally achieve higher precision, but their computational cost is an order of magnitude larger than single-stage models and typically exceeds the real-time budget of an embedded flight computer. Cascade architectures offer a middle path: a cheap, high-recall front end proposes candidates, and a more discriminative back end verifies only those candidates. Because the verifier runs on a small number of crops rather than the full frame, its cost scales with the number of proposals rather than image resolution, which keeps the average per-frame budget bounded. This work follows that cascade philosophy, pairing a permissive scanner with a dedicated verification stage to reject thermal clutter without abandoning real-time operation.

The efficiency of a cascade depends on the statistics of the scene. When most frames contain few or no candidate regions—as in sparse wilderness search—the verifier is seldom invoked and the amortized cost approaches that of the scanner alone. In dense or highly cluttered scenes the verifier runs more often, but this is also where its discrimination is most valuable. A well-designed cascade thus spends compute adaptively, concentrating effort where ambiguity is greatest rather than paying a fixed high cost on every frame as a monolithic two-stage detector would.

E. Domain Gap and Robust Evaluation

A further theme in recent steganalysis-adjacent and detection literature is that models tuned on one data distribution often degrade when deployed on another. In aerial thermal detection this domain gap arises from differences in altitude, viewing angle, sensor response, and scene composition between training and deployment. Techniques such as transfer learning, targeted data augmentation, and regularization schemes that discourage over-confident decision boundaries [6] are commonly used to narrow this gap. The recurring lesson is that a classifier's reported accuracy on an in-domain test split can substantially overstate its behaviour once the operating domain shifts.

This makes evaluation methodology as important as architecture. Class-balanced held-out sets, threshold selection driven by the receiver operating characteristic, and confidence-aware reporting all help expose failure modes that headline accuracy figures conceal. The present study adopts this stance: rather than reporting a single optimistic accuracy, it examines how the classifier's confidence distribution itself shifts under domain mismatch, which is what surfaces the Confidence Compression effect.

III. MATERIALS AND METHODS

A. Two-Stage Cascade Pipeline

The inference pipeline operates in three steps: (1) Stage 1 scans the full thermal frame using a YOLOv8-Nano detector at a permissive confidence threshold, forwarding all proposals above it to maximize recall; (2) detected bounding boxes are cropped and padded by 10% to capture contextual border pixels, then resized to 224 × 224 px; (3) Stage 2 classifies each crop as human or clutter using EfficientNet-B2, outputting a binary accept/reject decision per ROI via a ROC-Youden optimal threshold.

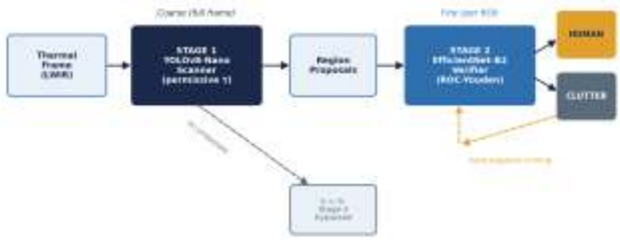

Fig. 1. Two-stage cascade overview. A permissive Stage 1 scanner proposes candidate regions from the full thermal frame; Stage 2 verifies each region as human or clutter. Stage 2 is bypassed when no proposals are generated (k = 0), and Stage 1 false positives are re-labeled as clutter through hard-negative mining.

B. Loss Function and Domain Adaptation

The refiner is trained with a Weighted Binary Cross-Entropy objective that penalizes missed detections more heavily than false alarms:

$$L = -(1/N) \sum [\, w \cdot y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i) \,]$$

where w = 3.0 is the positive-class weight, the loss is averaged over N samples, and $\hat{y}_i$ is the predicted probability for target y_i. This objective is combined with label smoothing and Mixup regularization [6]. The refiner is adapted from ImageNet initialization to the aerial thermal domain, with an iterative hard negative mining loop that re-labels Stage 1 false positives as clutter, focusing the verifier on the scanner's characteristic failure modes.

IV. RESULTS AND DISCUSSION

A. Decision Threshold Sensitivity Analysis

Evaluation of the baseline cross-domain configuration exposed a pronounced Confidence Compression artifact: the verifier's confidence scores for true human targets saturate near a ceiling of approximately 0.64 (on the 0–1 probability scale), so raising the decision threshold to suppress clutter collapses human recall. This ceiling reflects poor confidence calibration under domain shift rather than a low classification accuracy. Domain-matched adaptation restores the posterior scale, enabling operation at a ROC-Youden optimal threshold. At this operating point the refiner sharply reduces false positives relative to the single-stage baseline while preserving human recall, confirming that the compression is a property of the training-domain mismatch rather than an inherent limit of the classifier.

Under the baseline cross-domain configuration, the verifier's confidence scores for genuine human targets fail to exceed a ceiling of approximately 0.64 on the probability scale. Because the decision threshold must sit below this confidence ceiling to retain recall, the operating point is forced into a regime where clutter and humans are poorly separated, and no single threshold yields both high recall and low false-alarm rate. Domain-matched adaptation restores the posterior scale, decompressing the confidence distribution so that a well-separated operating threshold becomes available. At this threshold the verifier reaches a balanced classification accuracy in the mid-90% range, with high human recall and a low false-alarm rate—a decisive improvement over the single-stage baseline.

B. Computational Complexity and Latency Feasibility

The total pipeline computational cost per frame is analytically modeled as follows:

$$C_{total} = C_{detector} + (k \times C_{refiner})$$

where C_detector ≈ 8.9 GFLOPS and C_refiner ≈ 1.0 GFLOPS per crop [4]. At a typical operating point (k ≈ 8 proposals) with a Stage 2 cost of ~2.0 ms per crop, the projected total execution time is ~39 ms (~26 FPS) on the NVIDIA Jetson Orin Nano, within the 66 ms (15 FPS) real-time budget. As no on-device measurement was performed, this figure is an analytical projection rather than a measured result. When Stage 1 yields no proposals (k = 0), Stage 2 is bypassed entirely and the pipeline cost falls to C_detector ≈ 8.9 GFLOPS. Even under the heavier proposal loads observed on the test distribution, the projected cost remains within the real-time budget, preserving an engineering margin.

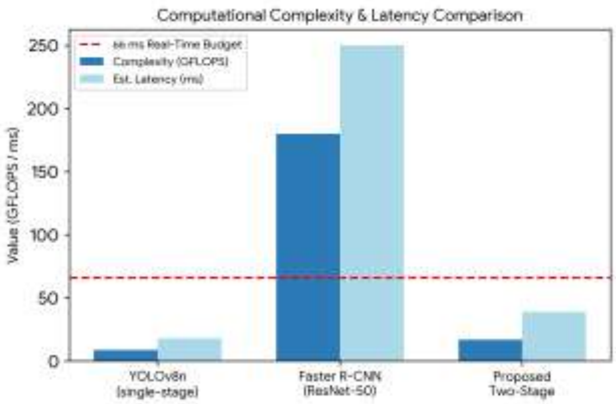

Fig. 2. Computational complexity and latency comparison.

C. Operational Implications

The practical value of a verification cascade in SAR is measured less by peak accuracy than by the behaviour of the system across a full sortie. A permissive front end ensures that few genuine victims are discarded early, while the verifier suppresses the clutter-driven alerts that would otherwise erode operator trust and waste flight energy. Because the verifier is invoked only on proposed regions, its cost is naturally low over the many empty frames typical of wide-area search, and rises only when the scene is busy—exactly when additional discrimination is most useful.

This asymmetry between recall at the scanner and precision at the verifier is the central design principle of the architecture. It reframes thermal clutter rejection not as a single classification threshold to be tuned, but as a division of labour between a fast, forgiving detector and a slower, stricter judge. The Confidence Compression finding reinforces this view: when the verifier is trained on mismatched-domain data, its confidence scale collapses and the division of labour fails, whereas domain-matched adaptation restores a usable decision margin.

V. CONCLUSION AND FUTURE WORK

This paper presents the design and evaluation of a two-stage Edge-AI cascade pipeline for thermal clutter rejection in autonomous SAR drone flights. By coupling a permissive single-stage YOLOv8-Nano scanner with a fine-tuned EfficientNet-B2 hard-negative verification refiner, the proposed architecture reaches a balanced classification accuracy of approximately 95% at its operating threshold and substantially reduces false alarms driven by the thermal crossover phenomenon while preserving human recall. An analytical complexity model indicates that the added verification stage remains compatible with real-time operation on resource-constrained airborne edge nodes.

A. Limitation and Future Work

The reported latency remains an analytical projection; on-device benchmarking on the Jetson Orin Nano, including memory-transfer overhead and thermal throttling effects, is left to future work. In addition, the present evaluation is confined to a single dataset and a limited held-out sample, yielding wide confidence intervals. Future research will expand evaluation across additional altitudes, sensors, and thermal datasets to tighten these bounds, investigate altitude-adaptive threshold calibration, and assess the end-to-end cascade coupled to the Stage 1 scanner in flight conditions. More broadly, robust thermal clutter rejection for autonomous SAR will require evaluation beyond any single benchmark: across seasons and times of day when thermal crossover is most severe, across sensor models with differing noise and resolution characteristics, and against moving or partially occluded subjects. Establishing standardized, mission-representative evaluation protocols for this task remains an open community need, and is a prerequisite for fielding such systems with confidence.

Looking further ahead, thermal clutter rejection is likely to benefit from multimodal fusion, combining infrared with visible or depth cues where payload permits, and from temporal reasoning that exploits the fact that a person moves and cools differently from inert clutter. Transformer-based and distillation-compressed detectors may also narrow the gap between the accuracy of large models and the latency ceiling of edge hardware. These directions share a common goal: making autonomous SAR detection dependable not under laboratory conditions but across the full range of environments in which lives may actually depend on it.

ACKNOWLEDGMENT

The authors express gratitude to all who have supported this academic computing research program and provided the benchmarking datasets.

REFERENCES

1. J. Suo, T. Wang, X. Zhang, H. Chen, W. Zhou, and W. Shi, “HIT-UAV: A high-altitude infrared thermal dataset for Unmanned Aerial Vehicle-based object detection,” Sci. Data, vol. 10, p. 227, Apr. 2023, doi: 10.1038/s41597-023-02066-6.
2. S. Yeom, “Thermal image tracking for search and rescue missions with a drone,” Drones, vol. 8, no. 2, p. 53, Feb. 2024, doi: 10.3390/drones8020053.
3. G. Jocher, A. Chaurasia, and J. Qiu, “Ultralytics YOLOv8,” Ultralytics, Los Angeles, CA, USA, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics
4. M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in Proc. 36th Int. Conf. Mach. Learn. (ICML), 2019, pp. 6105–6114.
5. B. Ding, Y. Zhang, and S. Ma, “A lightweight real-time infrared object detection model based on YOLOv8 for unmanned aerial vehicles,” Drones, vol. 8, no. 9, p. 479, Sep. 2024, doi: 10.3390/drones8090479.
6. H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond empirical risk minimization,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2018.