A Lightweight UAV-Based SAR System for Human Detection and Monocular Geolocation

Hamza Ghitri[1,2]
[1]Silesian University of Technology, Gliwice, Poland
[2]University of Ain Temouchent, Ain Temouchent, Algeria

***Abstract*—Timely localization of missing persons remains a critical challenge in search-and-rescue (SAR) operations, where manual analysis of UAV imagery limits response speed and operational efficiency. This paper presents a lightweight end-to-end UAV-based SAR pipeline integrating real-time human detection, monocular GPS geolocation, and automated alerting into a deployable embedded system. Human detection is performed using a YOLOv8n model trained through a two-stage transfer learning strategy using the VisDrone and HERIDAL datasets. Detected persons are geolocated using a lightweight geometric algorithm that estimates ground coordinates directly from standard UAV telemetry without requiring additional sensors.**

The proposed system is evaluated on a custom annotated dataset of 300 UAV frames collected under four controlled flight conditions varying in altitude and camera angle. The final model achieves 0.921 precision, 0.926 recall, and 0.965 mAP@0.5, outperforming single-stage training baselines. The geolocation module achieves a mean localization error of 1.01 m across 60 independent measurements under controlled conditions. For real-time onboard deployment, the model is optimized using W8A16 post-training quantization and benchmarked on a Qualcomm RB3 Gen 2 NPU, achieving a median inference latency of 37.3 ms at 960×960 resolution while preserving detection accuracy.

The complete implementation, trained models, and evaluation dataset are publicly released to support reproducible research and future development in deployable SAR systems.

***Index Terms*—Search and Rescue, UAV, Aerial Human Detection, Edge AI, Monocular Geolocation, Embedded Vision, YOLOv8, Edge Deployment, NPU Quantization**

I. Introduction

Search-and-rescue (SAR) operations require rapid localization of missing persons across large and often hazardous environments, where delays can significantly impact survival outcomes [1]. Unmanned aerial vehicles (UAVs) have become increasingly valuable in SAR missions due to their ability to rapidly survey difficult terrain and provide real-time aerial imagery. However, current workflows still rely heavily on human operators manually monitoring video streams, a process that is time-consuming and prone to error [1].

Recent advances in deep learning have significantly improved aerial human detection, particularly through lightweight real-time architectures such as YOLO [3]. Nevertheless, detection alone is insufficient for practical SAR deployment. Reporting only the UAV GPS position can introduce substantial localization error, especially at higher altitudes or under oblique viewing angles, where the detected target may be several meters away from the drone's ground projection.

While previous studies have explored aerial detection and UAV-based geolocation independently, comparatively few works address the complete end-to-end pipeline under real deployment constraints. Lightweight systems integrating real-time detection, monocular target geolocation, and automated operational response using only standard UAV telemetry remain limited.

This paper presents a lightweight end-to-end UAV-based SAR system integrating aerial human detection, monocular GPS geolocation, and automated alert generation. A YOLOv8n detector is trained using a two-stage VisDrone→HERIDAL transfer learning strategy, while target coordinates are estimated through a lightweight geometric projection algorithm requiring no additional sensors beyond standard UAV telemetry. The system further incorporates an automated alert module that transmits estimated target coordinates and visual evidence directly to rescue operators.

The proposed framework is evaluated on a custom UAV dataset collected under four controlled flight conditions. The final model achieves 0.921 precision, 0.926 recall, and 0.965 mAP@0.5, while the geolocation module achieves a mean localization error of 1.01 m. For embedded deployment, the model is optimized using W8A16 quantization and benchmarked on a Qualcomm RB3 Gen 2 NPU, achieving 37.3 ms median inference latency at 960×960 resolution.

The main contributions of this paper are summarized as follows:

1. End-to-End SAR Pipeline:A unified framework integrating real-time aerial human detection, monocular GPS geolocation, and automated alert generation into a single deployable UAV-SAR system.

2. Two-Stage Transfer Learning Strategy: A curriculum-style training approach (VisDrone → HERIDAL) combining large-scale aerial pretraining with SAR-specific adaptation.

3. Lightweight Monocular Geolocation: A geometric localization algorithm based solely on standard UAV telemetry, achieving approximately meter-level localization accuracy without additional sensing hardware.

4. Embedded Edge Deployment: A hardware-aware quantization pipeline enabling real-time W8A16 inference on a resource-constrained Qualcomm RB3 Gen 2 NPU platform.

II. Related Work

A. Aerial Human Detection

Human detection from UAV imagery remains challenging due to small object size, varying viewpoints, cluttered backgrounds, shadows, and pose variability. These difficulties are amplified in SAR scenarios, where targets may appear partially occluded, camouflaged, or lying on the ground while occupying only a small number of pixels.

Recent advances in deep learning, particularly the YOLO family of single-stage detectors, have significantly improved real-time aerial detection performance. Kucukayan and Karacan [8] proposed a YOLOv8-based detector for drone imagery, while Zhong et al. [9] introduced a lightweight architecture optimized for efficient inference on edge devices. Zhang et al. [1] identify small-object scale and background complexity as major challenges in UAV-based SAR perception.

The VisDrone dataset [5] has become a standard benchmark for aerial object detection due to its scale and diversity. However, its predominantly urban scenes differ substantially from wilderness SAR environments. To address this limitation, the HERIDAL dataset [6], [7] was introduced specifically for aerial SAR applications, containing wilderness imagery with standing, sitting, and lying subjects under challenging outdoor conditions. This motivates curriculum-style transfer learning, where large-scale aerial pretraining is followed by SAR-specific adaptation.

B. UAV-Based Target Geolocation

Estimating target GPS coordinates from monocular UAV imagery remains challenging without additional sensing hardware. Several studies have explored target geolocation from UAV imagery using monocular vision and UAV telemetry. Pan et al. [10] proposed a monocular-vision-based moving-target geolocation method using UAV imagery, Zhao et al. [11] combined monocular detection with GPS and IMU data for vehicle geolocation, while Sanyal et al. [12] demonstrated geometric target projection using YOLO-based detection and UAV telemetry. Zheng et al. [13] further improved localization accuracy through multi-stage coordinate transformations, though at increased calibration complexity.

In contrast, this work adopts a lightweight closed-form geometric formulation based solely on standard UAV telemetry, enabling approximate real-time target localization without additional sensing hardware.

C. End-to-End SAR Systems

Most UAV-based SAR research focuses on isolated components such as aerial detection or localization, while comparatively few works address the complete operational pipeline under embedded deployment constraints.

Calabro and Marchetti [14] proposed a drone-assisted framework integrating aerial detection with communication infrastructure for locating distressed individuals. However, the system depends on external infrastructure and does not provide a lightweight onboard perception and localization pipeline.

Recent edge-AI platforms equipped with dedicated NPUs have enabled efficient onboard inference for UAV applications [18]. The Qualcomm RB3 Gen 2 platform, providing up to 12 TOPS through its Hexagon NPU, represents one such deployment target for real-time embedded AI systems [19].

In contrast to prior work, this paper presents a lightweight end-to-end UAV-SAR pipeline integrating real-time aerial human detection, monocular GPS geolocation, and automated alert generation within a unified deployable framework using only monocular imagery and standard UAV telemetry.

III. Proposed System

This work proposes a lightweight end-to-end UAV-based search-and-rescue (SAR) framework integrating real-time aerial human detection, monocular GPS geolocation, and automated alert generation within a unified deployable pipeline. The system is designed for both offline analysis of recorded UAV footage and real-time onboard deployment on embedded edge hardware. An overview of the complete architecture is shown in Figure 1.

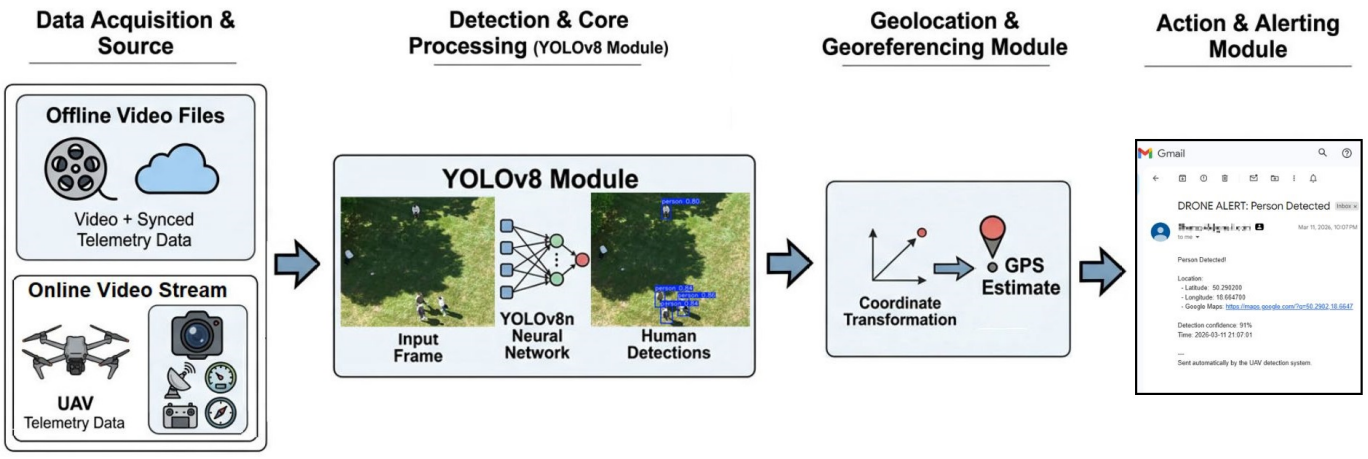

Fig. 1. System architecture of the proposed end-to-end UAV-based SAR System.

The proposed framework operates sequentially through four processing stages. First, aerial video frames and synchronized telemetry data are acquired either from live UAV streams or recorded flight sessions. The telemetry includes GPS position, altitude, compass heading, and camera field-of-view parameters required for geolocation estimation.

Second, each input frame is processed by a fine-tuned YOLOv8n model optimized for aerial SAR conditions. The detector produces bounding boxes and confidence scores corresponding to detected persons within the scene. The model is trained using a two-stage transfer learning strategy combining large-scale aerial pretraining on VisDrone with SAR-specific adaptation on HERIDAL.

Third, the detected bounding box coordinates are transformed into approximate ground GPS coordinates using a lightweight geometric projection algorithm based solely on standard UAV telemetry. Unlike approaches requiring LiDAR, stereo vision, or depth estimation hardware, the proposed method estimates target location using monocular imagery together with camera geometry and UAV orientation data.

Finally, an automated alert module transmits actionable detection information to rescue operators. The generated alert includes the estimated target GPS coordinates, a navigation link, detection confidence, timestamp, and annotated visual evidence frame, enabling rapid operational response without continuous manual monitoring of the UAV feed.

The modular architecture enables compatibility with resource-constrained embedded platforms while maintaining real-time inference capability. The complete pipeline is designed to minimize hardware requirements and support practical deployment in lightweight UAV-assisted SAR operations.

A. Model Selection and Training

Selecting an appropriate detection architecture is critical for UAV-based SAR applications, where accurate human detection must be achieved under strict real-time and resource-constrained deployment requirements [1]. In this work, YOLOv8n [4] is selected as the detection backbone due to its favorable balance between detection accuracy, computational efficiency, and embedded deployment suitability.

Compared to two-stage detectors such as Faster R-CNN [17], single-stage YOLO-based architectures provide significantly lower inference latency, making them more appropriate for real-time UAV operation. Among the YOLOv8 variants, YOLOv8n was specifically chosen because its reduced parameter count and lightweight architecture enable efficient inference on resource-constrained edge platforms while maintaining competitive detection performance.

YOLOv8 further introduces anchor-free detection, improved feature aggregation, and optimized training strategies that improve small-object detection performance [3]. This is particularly important in aerial SAR scenarios, where human targets frequently occupy only a small portion of the image, especially at higher altitudes.

Direct fine-tuning from generic ImageNet pretraining is insufficient for reliable deployment in wilderness SAR environments due to the substantial domain gap between natural-image datasets and aerial SAR imagery. To address this limitation, a two-stage transfer learning strategy is adopted. In the first stage, YOLOv8n is fine-tuned on a filtered subset of the VisDrone dataset [5], providing large-scale aerial imagery across diverse urban environments, scales, and illumination conditions for generalized aerial feature learning. In the second stage, the model is further adapted on the HERIDAL dataset [6], which captures SAR-specific conditions including wilderness terrain, camouflaged subjects, partial occlusions, and challenging poses such as lying or sitting individuals.

This curriculum-style progression from general aerial perception to SAR-specific adaptation improves robustness under real operational conditions and is consistent with recent findings reported by Ciccone and Ceruti [15]. Detailed dataset descriptions, training configuration, and deployment optimization are provided in Sections IV, V-A, and III-D, respectively.

B. Geolocation Algorithm

The proposed geolocation module estimates the ground position of detected persons using only standard onboard UAV telemetry, including GPS position, altitude, compass heading, and camera field of view. Unlike naive approaches that directly report the UAV GPS position upon detection, the proposed method explicitly models the spatial displacement between the drone and the target. The algorithm assumes a near-horizontal UAV attitude and a locally planar ground surface, under which image-space coordinates can be projected into approximate ground coordinates through camera geometry.

The geolocation pipeline operates through three sequential stages illustrated in Figures 2–4.

1) Pixel-to-Metric Projection: Let (c_x, c_y) denote the pixel coordinates of the detected bounding box center, and let W and H represent the image width and height in pixels. The displacement relative to the image center is defined as:

$$\delta_x = c_x - \frac{W}{2}, \qquad \delta_y = -\left(c_y - \frac{H}{2}\right) \tag{1}$$

Using the UAV altitude h and the horizontal and vertical camera fields of view α and β, the ground sampling distance per pixel is estimated as:

$$s_x = \frac{h \cdot \tan\left(\frac{\alpha}{2}\right)}{W/2}, \qquad s_y = \frac{h \cdot \tan\left(\frac{\beta}{2}\right)}{H/2} \tag{2}$$

The corresponding metric displacement relative to the UAV ground projection becomes:

$$d_x = \delta_x \cdot s_x, \qquad d_y = \delta_y \cdot s_y \tag{3}$$

This formulation converts image-space offsets into approximate ground-plane distances measured in meters.

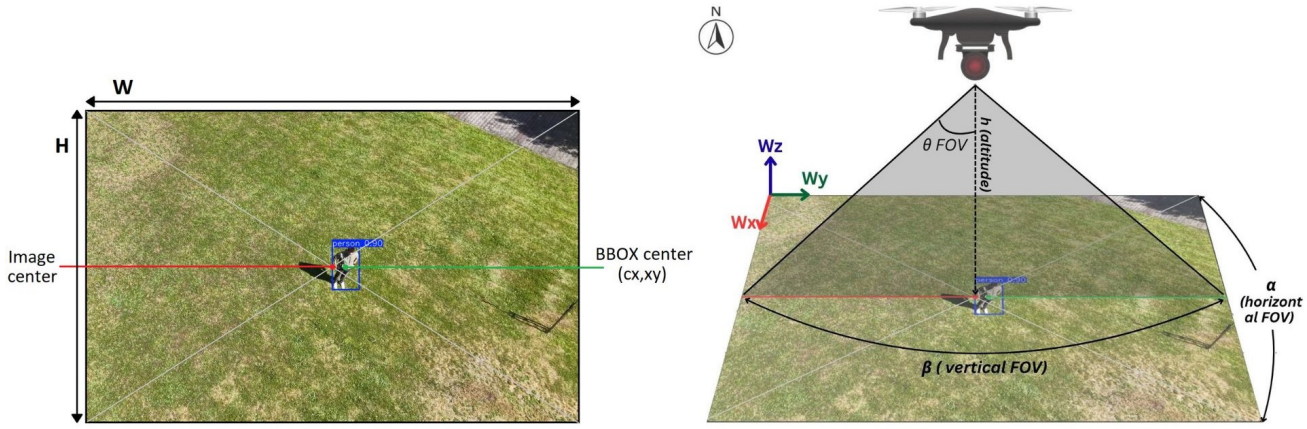

Fig. 2. Geometric interpretation of pixel offset and ground sampling distance.

2) Heading Compensation: The displacement vector (d_x, d_y) is initially expressed in the image coordinate frame. To align it with geographic coordinates, a rotation using the UAV compass heading ψ is applied:

$$\begin{pmatrix} d'_x \\ d'_y \end{pmatrix} = \begin{pmatrix} \cos\psi & -\sin\psi \\ \sin\psi & \cos\psi \end{pmatrix} \begin{pmatrix} d_x \\ d_y \end{pmatrix} \tag{4}$$

where $\psi = 0°$ corresponds to geographic North and increases clockwise. This transformation compensates for UAV orientation, ensuring that the estimated displacement is correctly aligned with the Earth reference frame.

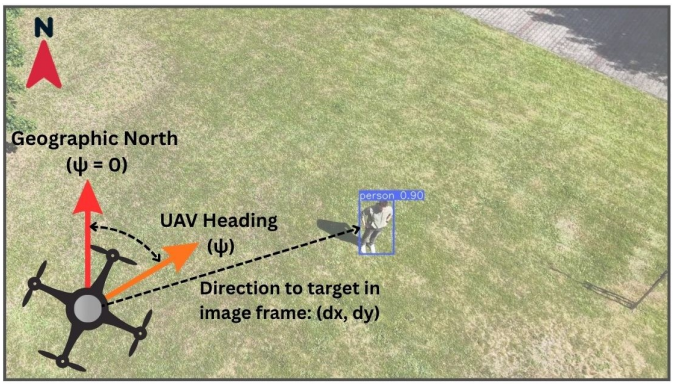

Fig. 3. Rotation of image-frame displacement into geographic coordinates.

3) GPS Coordinate Estimation: The rotated displacement (d'_x, d'_y) is converted into latitude and longitude offsets using the Earth's circumference $C = 40{,}075{,}000$ m. The latitude offset is computed as:

$$\Delta\phi = \frac{d'_y}{C} \times 360 \tag{5}$$

Longitude estimation additionally requires a cosine correction to compensate for meridian convergence:

$$\Delta\lambda = \frac{d'_x}{C \cdot \cos\left(\frac{\pi\phi_{\text{drone}}}{180}\right)} \times 360 \tag{6}$$

where ϕ_{drone} denotes the UAV latitude. The final estimated target coordinates are:

$$\phi_{\text{target}} = \phi_{\text{drone}} + \Delta\phi, \qquad \lambda_{\text{target}} = \lambda_{\text{drone}} + \Delta\lambda \tag{7}$$

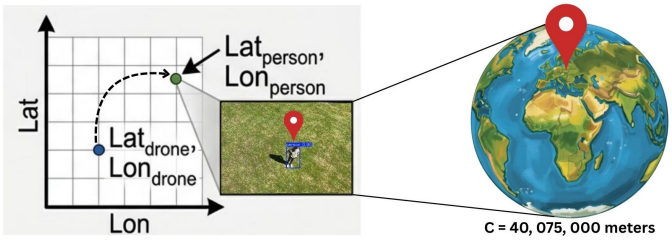

Fig. 4. Conversion of metric displacement into GPS coordinate offsets.

The proposed formulation is intentionally lightweight and designed for operationally useful approximate target localization rather than high-precision geodetic surveying. The complete closed-form solution executes in under 1 ms per detection, making it suitable for real-time onboard deployment on resource-constrained edge platforms.

Localization accuracy depends primarily on GPS precision, altitude reliability, camera calibration accuracy, and the validity of the planar-ground assumption. As evaluated in Section V-C, localization error increases progressively with altitude and oblique viewing angles due to amplified projection uncertainty and perspective distortion.

Although evaluated using a DJI Air 3S platform, the method is hardware-agnostic and can be adapted to any UAV platform providing camera field-of-view parameters and synchronized telemetry data. The required algorithm inputs are summarized in Table I.

C. Automated Alert System

To complete the operational SAR workflow, the proposed system incorporates a lightweight automated alert module that transmits detection results directly to rescue operators without requiring continuous manual monitoring of the UAV

TABLE I
SUMMARY OF GEOLOCATION ALGORITHM INPUTS AND THEIR SOURCES.

Input	Symbol	Description	Source
Bounding box center	(c_x, c_y)	Pixel coordinates of detected person	YOLOv8n output
Image dimensions	W, H	Frame width and height in pixels	Camera configuration
Horizontal FOV	α	Camera horizontal field of view	DJI Air 3S (76°)
Vertical FOV	β	Camera vertical field of view	DJI Air 3S (49°)
Altitude	h	Drone height above ground (m)	DJI SRT file
Compass heading	ψ	Drone yaw angle (° from North)	DJI SRT file
Drone position	ϕ, λ	Drone latitude and longitude	DJI SRT file

video feed. Once a person is detected and geolocated, the system generates a single alert per detection session, preventing repeated notifications and reducing alert flooding during continuous operation.

Each alert contains the estimated GPS coordinates of the detected target, a navigation link for rapid field deployment, the detection timestamp, and the nearest resolved location obtained through reverse geocoding. An annotated image frame is attached as visual confirmation, allowing rescue operators to verify the detection before dispatching ground teams.

Figure 5 shows a representative alert generated during field evaluation.

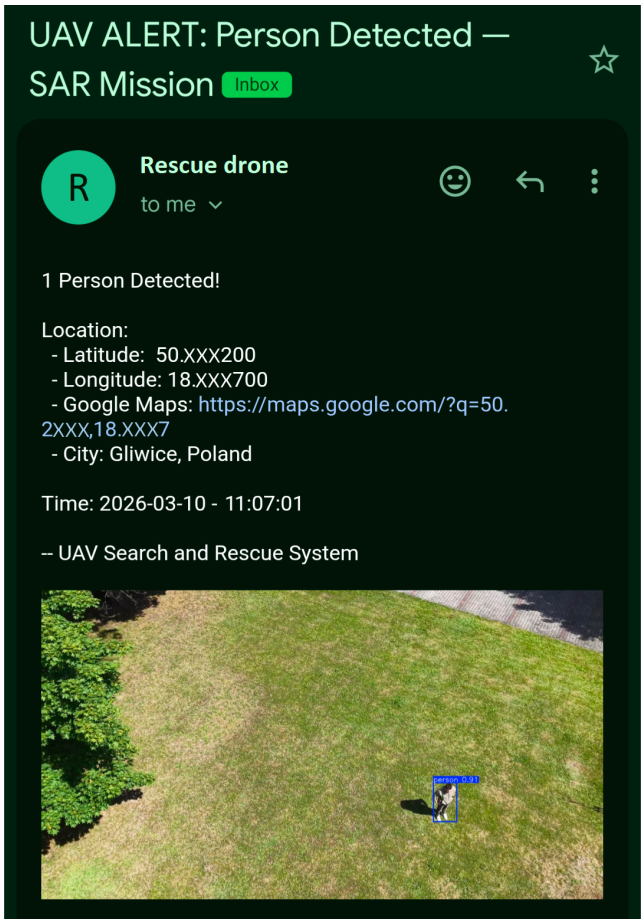

Fig. 5. Example alert transmitted to rescue operators following target detection and geolocation.

The notification pipeline is designed to operate with minimal infrastructure requirements and requires only a network connection, enabling compatibility with both onboard 4G-enabled UAV deployment and ground-station-based operation. By integrating automated detection, localization, and notification within a unified workflow, the system reduces operator workload and shortens the response time between aerial detection and field intervention.

D. Model Optimization

To enable efficient onboard inference on the Qualcomm RB3 Gen 2 platform, the trained YOLOv8n model is optimized for integer-based NPU execution using a post-training quantization pipeline. The original PyTorch model is first exported to ONNX format, which serves as an intermediate representation compatible with the Qualcomm AI Hub deployment workflow.

Post-training quantization (PTQ) is subsequently applied using a representative calibration set of SAR imagery. A hybrid W8A16 quantization scheme is adopted, where model weights are quantized to INT8 while activations remain in INT16 precision. This configuration was selected empirically after observing that fully INT8 quantization, although faster, introduced noticeable degradation in the detection of small and low-contrast human targets. Retaining 16-bit activations preserves fine-grained spatial information critical for reliable aerial SAR detection while still significantly reducing computational cost and memory usage relative to full floating-point inference.

Performance evaluation was conducted using Qualcomm AI Hub on physical QCS6490 hardware [19]. The final W8A16 model achieves a median inference latency of 37.3 ms at 960×960 input resolution with a peak memory footprint of 13–20 MB. These results confirm the suitability of the proposed pipeline for real-time embedded deployment on resource-constrained UAV edge platforms.

IV. DATASETS

Three datasets serve complementary roles within the experimental pipeline: VisDrone for large-scale aerial pretraining, HERIDAL for SAR-specific domain adaptation, and a custom UAV evaluation dataset for controlled real-world performance assessment.

A. VisDrone Dataset

The VisDrone dataset [5] is one of the largest publicly available benchmarks for UAV-based object detection, containing 10,209 aerial images annotated with more than 540,000 object instances across 10 categories. Due to its diversity in altitude, scale, illumination, and urban scene composition, it provides a strong foundation for generalized aerial feature learning.

For this work, only the two human-related classes (*pedestrian* and *people*) were retained and merged into a single *human* category to align with the binary detection objective of SAR operations. The resulting filtered dataset composition is summarized in Table II.

Fig. 6. Example aerial images from the VisDrone dataset.

TABLE II
VISDRONE DATASET COMPOSITION AFTER CLASS FILTERING.

Split	Images	Human Instances
Training	6,471	∼39,000
Validation	548	∼3,300
Test	1,610	∼9,700
Total	8,629	∼52,000

Although VisDrone provides large-scale aerial diversity, its predominantly urban environments differ substantially from the wilderness conditions encountered in SAR missions. This domain mismatch motivates the second-stage SAR-specific adaptation on HERIDAL.

B. HERIDAL Dataset

The HERIDAL (Human hERI-view aerial Dataset for seArch and rescuE) dataset [7] was specifically designed for aerial SAR research in wilderness environments. The dataset contains approximately 1,684 high-resolution images captured over mountainous terrain, forests, rocky regions, and grassland environments representative of real SAR conditions. Unlike urban aerial benchmarks, HERIDAL focuses exclusively on human detection under challenging outdoor conditions.

The dataset includes subjects in standing, sitting, and lying poses, frequently partially occluded by vegetation, shadows, or terrain structures. In many images, targets occupy only a small number of pixels due to UAV altitude and image resolution, making the dataset particularly relevant for aerial SAR perception tasks. The dataset composition is summarized in Table III.

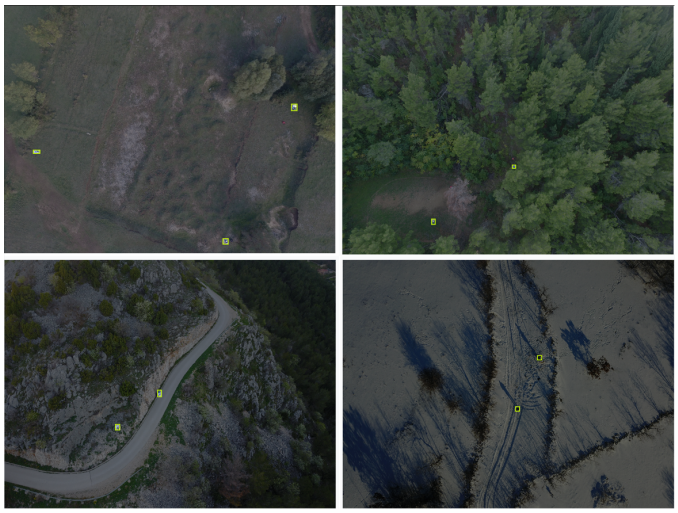

Fig. 7. Example SAR-oriented aerial images from the HERIDAL dataset.

TABLE III
HERIDAL DATASET COMPOSITION.

Split	Images	Human Instances
Training	1,583	2,996
Test	101	314
Total	1,684	3,310

Compared to VisDrone, HERIDAL more accurately reflects the visual conditions encountered during real SAR operations, including irregular terrain, camouflaged targets, and difficult viewing conditions. However, its relatively limited scale motivates initialization from larger aerial datasets prior to SAR-specific fine-tuning.

C. Custom Evaluation Dataset

To evaluate real-world system performance under controlled operational conditions, a dedicated UAV evaluation dataset was collected using a DJI Air 3S at 4K resolution and 60 FPS. The dataset was designed to assess detection robustness and geolocation performance across varying flight altitudes, viewing angles, human poses, and potential false-positive conditions.

Data collection was conducted under four flight configurations obtained by combining two altitudes (15 m and 30 m) with two camera orientations (45° oblique and 90° nadir), as summarized in Table IV. Video sequences were recorded continuously for 75 seconds per condition. Frames were subsequently extracted at 1 FPS and resized to 1920×1080 prior

Fig. 8. Representative frames from the four flight conditions.

to annotation. Given the continuous UAV and subject motion at 60 FPS acquisition, the selected extraction rate provides sufficient scene diversity while avoiding excessive temporal redundancy.

The dataset includes human poses and scenarios relevant to SAR operations, including standing, sitting, lying, grouped, and moving subjects. The lying pose is particularly important because it represents injured or unconscious victims, a condition largely underrepresented in existing aerial detection benchmarks such as VisDrone.

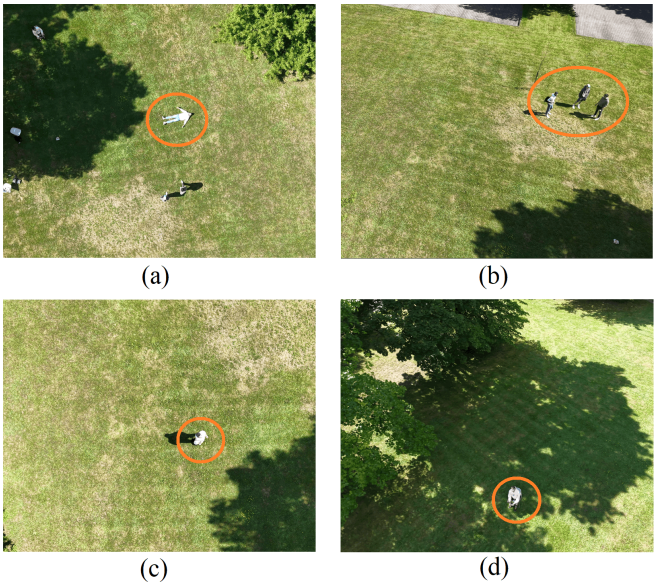

Fig. 9. Representative SAR scenarios including lying, standing, grouped, and seated subjects. (a) person lying on the ground. (b) group of people standing. (c) single person standing. (d) single person in a chair.

To evaluate robustness against false positives, visually similar distractor objects such as chairs and bags were intentionally introduced into the scene while remaining unannotated. These objects were selected because they exhibit visual characteristics that may resemble human targets from aerial viewpoints, particularly at higher altitudes. Cast shadows were also preserved within the dataset to evaluate the model's ability to distinguish human subjects from shadow projections under varying illumination conditions.

Fig. 10. Examples of intentionally introduced false-positive distractors and shadow conditions. (a) a collection of scattered chairs and bags. (b) an example of people's shadows being reflected on the ground.

All 300 frames were manually annotated using Roboflow with tight bounding boxes. The annotation protocol includes partially occluded individuals while excluding instances smaller than 10 pixels in height due to insufficient visual information for reliable labeling. No data augmentation was applied in order to preserve the geometric consistency between image-space coordinates and ground positions required for geolocation evaluation. The complete evaluation dataset is publicly available [22].

TABLE IV
CUSTOM EVALUATION DATASET: FLIGHT CONDITIONS, FRAME COUNTS, AND HUMAN INSTANCES.

Condition	Altitude	Camera Angle	Duration	Frames Extracted	Human Instances
Video 1	15m	90°	75s	75	111
Video 2	15m	45°	75s	75	140
Video 3	30m	45°	75s	75	206
Video 4	30m	90°	75s	75	206
Total			**300s**	**300**	**663**

V. Results and Evaluation

A. Experimental Setup

All experiments were conducted using Python, PyTorch, and the Ultralytics YOLOv8 framework [4]. To ensure a fair and reproducible comparison, three training strategies were evaluated under a unified configuration: VisDrone-only training, HERIDAL-only training, and the proposed two-stage VisDrone→HERIDAL transfer learning approach. The complete training configuration is summarized in Table V.

All models were trained using a fixed random seed, and the custom evaluation dataset was fully excluded from all training and hyperparameter selection stages. This separation ensures that all reported results reflect generalization performance under unseen operational conditions.

TABLE V
Unified training configuration for all model variants.

Parameter	Value
Base architecture	YOLOv8n
Epochs	50
Input resolution	960×960 px
Batch size	16
Optimizer	AdamW
Learning rate	0.001
Early stopping	patience = 20
Mosaic / Mixup	1.0 / 0.1
Training hardware	NVIDIA T4 GPU

For embedded deployment evaluation, the final model was optimized using the W8A16 quantization pipeline described in Section III-D and benchmarked on the Qualcomm RB3 Gen 2 (QCS6490) through Qualcomm AI Hub [19] at 960×960 input resolution.

Geolocation accuracy was evaluated using a stationary reference subject whose ground-truth position was approximated from georeferenced satellite imagery in Google Maps. The estimated annotation uncertainty of this procedure is approximately 1–2 m. Per-frame UAV telemetry, including GPS position, altitude, and heading, was extracted from the DJI SRT metadata stream.

Detection performance is reported using Precision, Recall, mAP@0.5, and mAP@0.5:0.95 following standard object detection evaluation protocols [3]. Geolocation accuracy is measured using the Haversine distance between predicted and reference GPS coordinates.

B. Detection Performance

Three model variants were evaluated to isolate the contribution of each training stage and quantify the domain gap between general aerial imagery and SAR-specific operational conditions.

1) Quantitative Results

Table VI summarizes the detection performance of all three training strategies. The VisDrone-only model achieves an mAP@0.5 of 0.597, indicating limited generalization from urban aerial imagery to wilderness SAR environments. In contrast, the HERIDAL-only model achieves 0.941 mAP@0.5, confirming that SAR-specific training data is essential for reliable performance under operational conditions.

TABLE VI
Detection performance of all three model variants on the custom SAR evaluation dataset.

Training Strategy	Precision	Recall	mAP@0.5	mAP@0.5:0.95
VisDrone only	0.580	0.567	0.597	0.367
HERIDAL only	0.845	0.911	0.941	0.757
Proposed	**0.921**	**0.926**	**0.965**	**0.777**

The proposed two-stage VisDrone→HERIDAL strategy further improves all metrics, reaching 0.965 mAP@0.5 and 0.777 mAP@0.5:0.95. This suggests that large-scale aerial pretraining on VisDrone provides a generalized feature foundation that is subsequently specialized for SAR conditions through HERIDAL fine-tuning. These findings are consistent with Ciccone and Ceruti [15], who reported the use of a two-stage transfer learning strategy with VisDrone for general aerial object detection and HERIDAL for SAR-specific fine-tuning.

The gap between mAP@0.5 and mAP@0.5:0.95 mainly reflects bounding-box localization imprecision rather than missed detections. Two recurring effects contribute to this behavior: closely grouped individuals are occasionally merged into a single predicted box, and cast shadows are sometimes partially included within detections, increasing box area and reducing IoU under stricter thresholds.

2) Per-Condition Analysis

Table VII reports performance across the four flight conditions. Detection recall exceeds 0.955 for both 15 m configurations but decreases at 30 m altitude, where human targets occupy fewer pixels and contain less discriminative visual information.

The 30 m / 90° condition produces the lowest recall (0.874) despite maintaining a high mAP@0.5 of 0.980. Two factors contribute to this behavior. First, nadir imagery captures primarily the head and shoulders, causing light-colored clothing to blend with the ground and increasing false negatives. Second, the top-down perspective reduces apparent separation between nearby individuals, occasionally producing merged detections over groups.

Despite these challenges, precision remains consistently high across all conditions (0.918–0.970), indicating strong robustness against spurious detections under varying altitudes and camera orientations.

3) Qualitative Results

Figure 11 presents representative detections across all evaluation conditions. The model successfully identifies standing, sitting, lying, grouped, and moving subjects while remaining robust against intentionally introduced distractor objects such as chairs and bags.

Figure 12 illustrates the primary failure cases corresponding to the FP and FN counts reported in Table VII. False positives mainly originate from cast shadows and visually similar

TABLE VII
PER-CONDITION DETECTION PERFORMANCE OF THE PROPOSED VISDRONE → HERIDAL MODEL.

Condition	Instances	TP	FP	FN	Precision	Recall	mAP@0.5
15 m / 45°	140	136	10	4	0.955	0.957	0.976
15 m / 90°	111	108	12	3	0.970	0.955	0.982
30 m / 45°	206	190	17	16	0.918	0.922	0.922
30 m / 90°	206	180	14	26	0.928	0.874	0.980
Overall	663	614	53	49	**0.921**	**0.926**	**0.965**

Fig. 11. Detection examples across all flight conditions.

ground objects, while false negatives occur when light-colored clothing blends with the terrain or when nearby individuals are merged into a single detection. These errors are consistent with the limitations of monocular aerial detection under high-altitude and nadir-view conditions.

4) Edge Deployment Results

Table VIII compares inference latency across deployment configurations. The FP32 model achieves 12.6 ms latency on an NVIDIA T4 GPU, suitable for offline processing. For embedded deployment, two quantization strategies were evaluated on the Qualcomm RB3 Gen 2 NPU.

Pure INT8 quantization achieves the lowest latency (11.4 ms) and smallest memory footprint but introduces noticeable degradation on small and low-contrast targets, particularly for lying persons. This behavior is consistent with prior observations regarding the sensitivity of small-object detection to aggressive activation quantization [20].

The proposed W8A16 configuration preserves 16-bit activations while quantizing weights to INT8, achieving 37.3 ms median latency with a 13–20 MB memory footprint at 960×960 resolution. This corresponds to approximately 26.8 FPS while preserving detection accuracy, confirming suitability for real-time UAV deployment.

TABLE VIII
INFERENCE LATENCY COMPARISON ACROSS HARDWARE PLATFORMS.

Hardware Platform	Precision	Median Latency
NVIDIA T4 GPU	FP32	12.6 ms
Qualcomm RB3 Gen 2 NPU	W8A16	37.3 ms
Qualcomm RB3 Gen 2 NPU	INT8	11.4 ms

C. Geolocation Accuracy

A single stationary volunteer subject was positioned at a fixed location throughout all four flight conditions as illustrated in Figure 13. A total of 60 independent measurements were collected by selecting 15 frames from each video sequence at uniform temporal intervals. Ground-truth coordinates were approximated using georeferenced satellite imagery in Google Maps, with an estimated annotation uncertainty of 1–2 m.

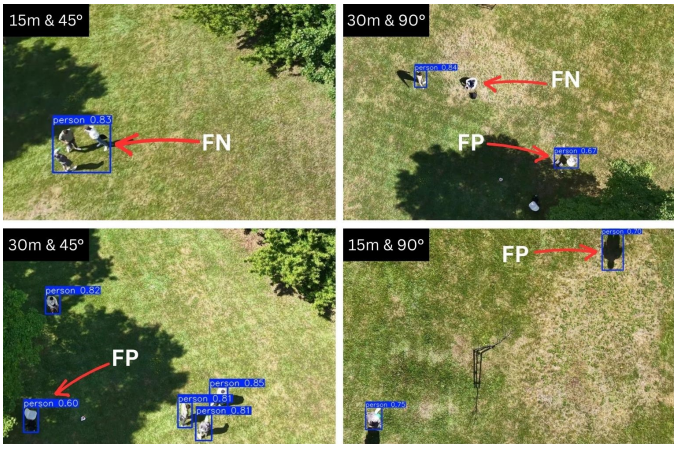

Fig. 12. Representative false-positive and false-negative cases.

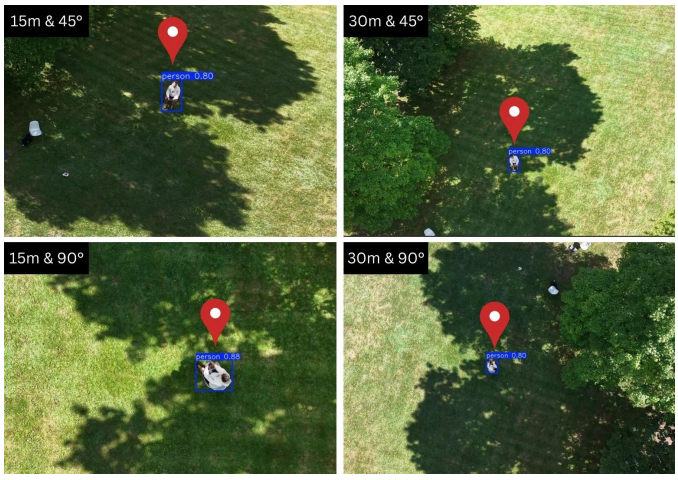

Fig. 13. Aerial views of the stationary target across the four flight conditions.

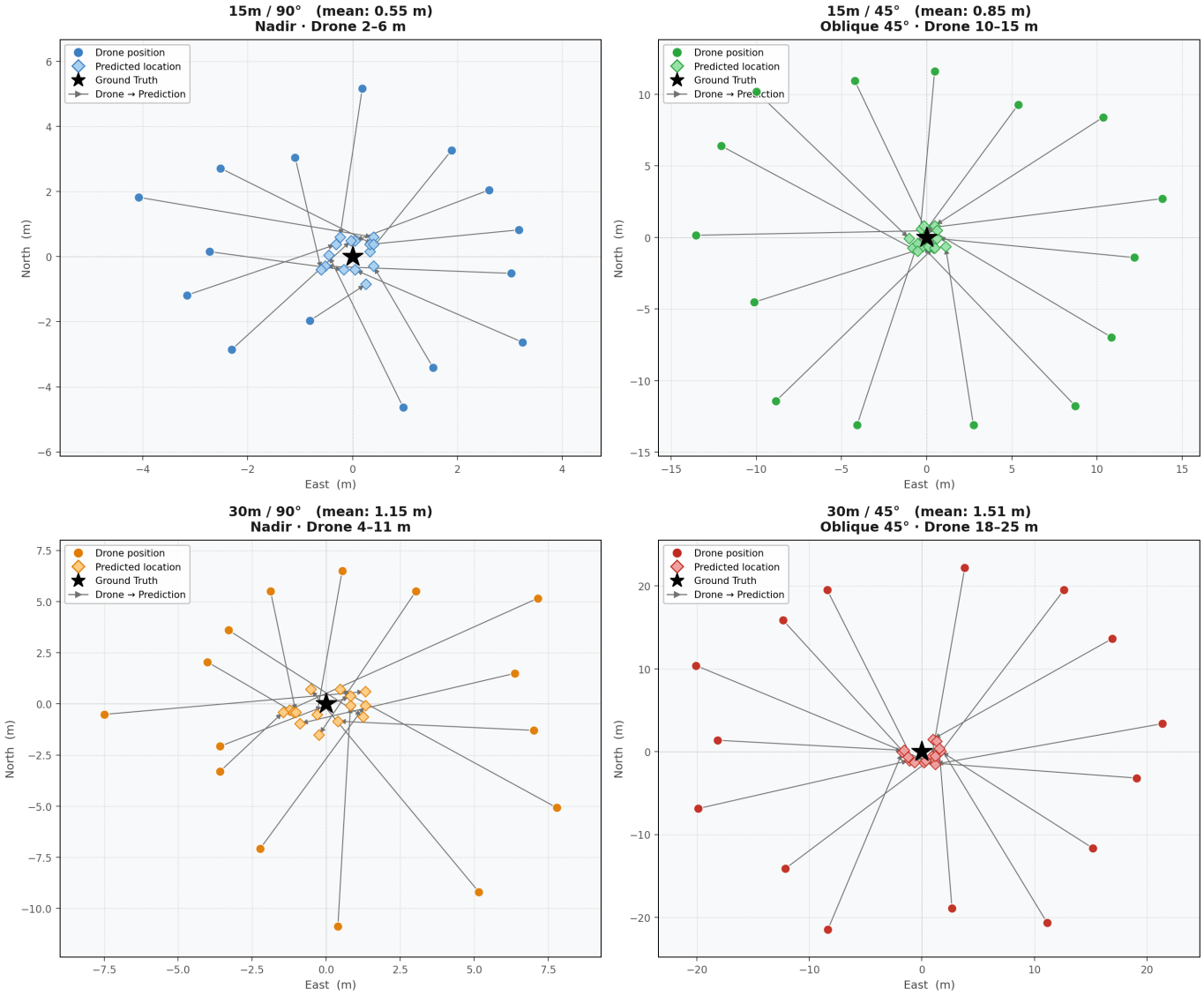

Fig. 14. Estimated GPS positions relative to the fixed ground-truth target location.

Figure 14 illustrates the spatial distribution of estimated target coordinates relative to ground truth. Position estimates remain tightly clustered at 15 m altitude, while spatial dispersion increases at 30 m, particularly under oblique viewing conditions. Localization error was computed using the Haversine distance between estimated and reference GPS coordinates:

$$E = 2R \arcsin\left[\sin^2\frac{\Delta\phi}{2} + \cos\phi_1 \cos\phi_2 \sin^2\frac{\Delta\lambda}{2}\right]^{1/2} \quad (8)$$

where ϕ and λ denote latitude and longitude coordinates.

Table IX summarizes the localization results. The proposed method achieves an overall mean error of 1.01 m across all evaluated conditions. The lowest error is obtained at 15 m nadir view, where the flat-ground assumption is most valid and the target occupies the largest image area. Localization error increases progressively with altitude due to reduced ground sampling resolution and amplified projection sensitivity to GPS and heading uncertainty.

Oblique (45°) conditions consistently produce larger errors than nadir configurations because perspective distortion violates the planar-ground assumption used by the geometric projection model. This effect becomes more pronounced at 30 m altitude, where both angular uncertainty and image-space compression increase simultaneously.

Figure 15 presents the error distribution for each condition with individual measurements overlaid. The relatively narrow interquartile ranges indicate stable localization estimates across varying UAV positions within each flight condition, and no measurement exceeds 2 m error across the complete evaluation set.

Figure 16 further shows the monotonic increase in mean localization error with altitude for nadir and oblique camera configurations.

TABLE IX
GEOLOCATION ERROR PER FLIGHT CONDITION. N DENOTES THE NUMBER OF INDEPENDENT FRAME MEASUREMENTS.

Condition	Altitude	Angle	N	Mean Error (m)	Std Dev (m)	Max Error (m)
C1	15 m	90°	15	0.55	0.14	0.89
C2	15 m	45°	15	0.84	0.21	1.29
C3	30 m	90°	15	1.16	0.29	1.56
C4	30 m	45°	15	1.51	0.25	1.86
Overall			60	**1.01**	**0.43**	**1.86**

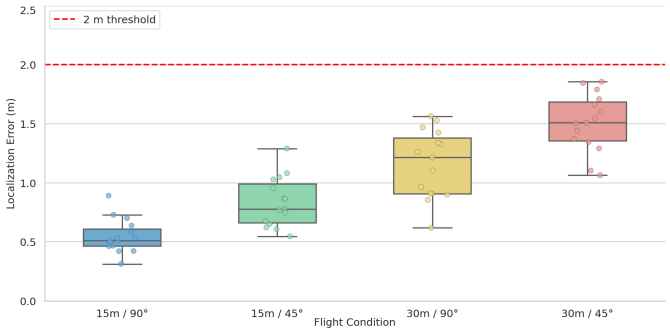

Fig. 15. Distribution of geolocation error across all flight conditions.

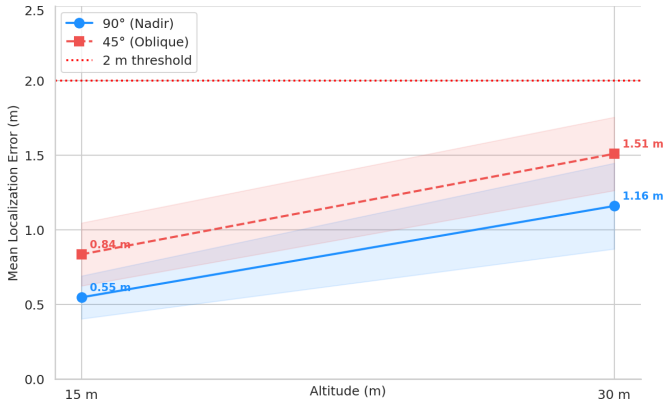

Fig. 16. Mean localization error versus altitude for nadir and oblique camera configurations.

These results demonstrate that standard UAV telemetry is sufficient for operationally useful approximate target geolocation without requiring additional sensing hardware or explicit scene calibration, achieving sub-meter accuracy under favorable conditions while remaining below 2 m error across all evaluated scenarios.

VI. CONCLUSION

This paper presented a lightweight end-to-end UAV-based search-and-rescue pipeline integrating real-time aerial human detection, monocular GPS geolocation, and automated alert generation within a unified deployable framework. The proposed two-stage transfer learning strategy (VisDrone → HERIDAL) achieved 0.921 precision, 0.926 recall, and 0.965 mAP@0.5 on the custom SAR evaluation dataset, outperforming both single-stage baselines. The geometric geolocation module achieved a mean localization error of 1.01 m across 60 independent measurements using only standard UAV telemetry, without requiring additional sensing hardware. For embedded deployment, the W8A16-optimized model achieved 37.3 ms median inference latency on a Qualcomm RB3 Gen 2 NPU, confirming real-time feasibility on resource-constrained edge platforms.

Overall, the results demonstrate that operationally useful human detection and approximate target localization can be achieved using a lightweight monocular UAV system with minimal hardware requirements. By combining perception, localization, and automated notification within a single deployable pipeline, the proposed framework represents a practical step toward real-time AI-assisted SAR operations.

The complete implementation, trained models, and evaluation dataset are publicly available to support reproducible research and future development in deployable UAV-SAR systems [21], [22].

Limitations

Several limitations should be acknowledged. First, the custom evaluation dataset is relatively small and was collected in a single outdoor environment, which may limit generalization to diverse terrains and conditions. Second, geolocation evaluation was performed using a single stationary subject, while ground-truth coordinates were approximated through georeferenced satellite imagery, introducing an estimated positional uncertainty of approximately 1–2 m. Third, the proposed geolocation algorithm assumes a locally planar ground surface and near-nadir camera orientation, which may reduce accuracy under complex terrain or extreme viewing angles.

Future Work

Future work will focus on improving robustness, scalability, and deployment realism. Expanding the dataset to include diverse environments, moving subjects, adverse weather conditions, and night-time imagery would improve model generalization under real SAR scenarios. Geolocation accuracy could be improved through RTK GPS integration or multi-sensor fusion. Tiling-based inference strategies such as SAHI would address the recall degradation observed at higher altitudes. Finally, multi-object tracking and night-vision support would further extend the system's applicability in real SAR deployments.

Ethics and Privacy Statement

All field data were collected under controlled conditions with voluntary participation from members of the research team. The dataset was prepared for research purposes only, and no sensitive personal information is intentionally released. To protect privacy, exact GPS coordinates and detailed location information were removed or generalized in the public manuscript and dataset. Public examples are limited to non-sensitive geographic context when needed to describe the experimental setting.

Data and Code Availability

1) The source code, and deployment resources are publicly available on GitHub [21]: https://github.com/7amzaGH/UAV-SAR-Human-Detection-and-Geolocation.
2) The evaluation dataset is publicly available on Roboflow Universe [22]: https://universe.roboflow.com/hamzaghitri/uav-sar-human-detection-dataset.

ACKNOWLEDGMENTS

The author thanks Jakub Gutt, Wojciech Seman, and Krzysztof Połeć for their assistance during UAV field data collection and controlled evaluation activities conducted in this work.

REFERENCES

[1] X. Zhang, Y. Feng, N. Wang, G. Lu, and S. Mei, "Aerial Person Detection for Search and Rescue: Survey and Benchmarks," *Journal of Remote Sensing*, vol. 5, Art. no. 0474, 2025, doi: 10.34133/remotesensing.0474.

[2] T. Manzini and R. Murphy, "Open Problems in Computer Vision for Wilderness SAR and the Search for Patricia Wu-Murad," arXiv:2307.14527, 2023, doi: 10.48550/arXiv.2307.14527.

[3] J. Terven, D.-M. Córdova-Esparza, and J.-A. Romero-González, "A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS," *Machine Learning and Knowledge Extraction*, vol. 5, no. 4, pp. 1680–1716, 2023, doi: 10.3390/make5040083.

[4] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8," GitHub repository, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[5] P. Zhu *et al.*, "Detection and Tracking Meet Drones Challenge," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7380–7399, 2021, doi: 10.1109/TPAMI.2021.3119563.

[6] D. Bozic-Stulic, Z. Marusic, and S. Gotovac, "Deep Learning Approach in Aerial Imagery for Supporting Land Search and Rescue Missions," *International Journal of Computer Vision*, vol. 127, pp. 1256–1278, 2019, doi: 10.1007/s11263-019-01177-1.

[7] P. Pyrrö, "HERIDAL: Human hERI-view aerial Dataset for seArch and resCuE (Keras-RetinaNet PASCAL VOC format)," Zenodo, Nov. 2021, doi: 10.5281/zenodo.5662351. [Online]. Available: https://doi.org/10.5281/zenodo.5662351.

[8] G. Kucukayan and H. Karacan, "YOLO-IHD: Improved Real-Time Human Detection System for Indoor Drones," *Sensors*, vol. 24, no. 3, p. 922, 2024, doi: 10.3390/s24030922.

[9] H. Zhong, Y. Zhang, Z. Shi, Y. Zhang, and L. Zhao, "PS-YOLO: A Lighter and Faster Network for UAV Object Detection," *Remote Sensing*, vol. 17, p. 1641, 2025, doi: 10.3390/rs17091641.

[10] T. Pan, B. Deng, H. Dong, J. Gui, and B. Zhao, "Monocular-Vision-Based Moving Target Geolocation Using Unmanned Aerial Vehicle," *Drones*, vol. 7, no. 2, p. 87, 2023, doi: 10.3390/drones7020087.

[11] X. Zhao, F. Pu, W. Wang, H. Chen, and Z. Xu, "Detection, Tracking, and Geolocation of Moving Vehicle From UAV Using Monocular Camera," *IEEE Access*, vol. 7, pp. 101160–101170, 2019, doi: 10.1109/ACCESS.2019.2929760.

[12] S. Sanyal, S. Bhushan, and K. Sivayazi, "Detection and Location Estimation of Object in Unmanned Aerial Vehicle Using Single Camera and GPS," in *Proceedings of the 2020 First International Conference on Power, Control and Computing Technologies (ICPC2T)*, 2020, pp. 73–78, doi: 10.1109/ICPC2T48082.2020.9071439.

[13] Z. Zheng, H. Liu, Z. Ye, M. Wang, H. Li, X. Lu, and Q. Li, "Target Localization of a Quadrotor UAV with Multi-Level Coordinate System Transformation Based on Monocular Camera Position Compensation," *Electronics*, vol. 14, no. 22, Art. no. 4371, 2025, doi: 10.3390/electronics14224371.

[14] A. Calabro and E. Marchetti, "Transponder: Support for Localizing Distressed People through a Flying Drone Network," *Drones*, vol. 8, no. 9, Art. no. 465, 2024, doi: 10.3390/drones8090465.

[15] F. Ciccone and A. Ceruti, "Real-Time Search and Rescue with Drones: A Deep Learning Approach for Small-Object Detection Based on YOLO," *Drones*, vol. 9, no. 8, Art. no. 514, 2025, doi: 10.3390/drones9080514.

[16] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 779–788, arXiv:1506.02640.

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *Advances in Neural Information Processing Systems*, vol. 28, 2015, arXiv:1506.01497.

[18] R. Jayanth, N. Gupta, and V. Prasanna, "Benchmarking Edge AI Platforms for High-Performance ML Inference," arXiv:2409.14803, 2024, doi: 10.48550/arXiv.2409.14803.

[19] Qualcomm Technologies, "Qualcomm RB3 Gen 2 Development Kit," Qualcomm Developer Network, 2024. [Online]. Available: https://www.qualcomm.com/developer/hardware/rb3-gen-2-development-kit. Accessed: May 10, 2026.

[20] A. Aldubaikhi and S. Patel, "Advancements in Small-Object Detection (2023–2025): Approaches, Datasets, Benchmarks, Applications, and Practical Guidance," *Applied Sciences*, vol. 15, no. 22, Art. no. 11882, 2025, doi: 10.3390/app152211882.

[21] H. Ghitri, "UAV-SAR Human Detection and Geolocation," GitHub repository, 2026. [Online]. Available: https://github.com/7amzaGH/UAV-SAR-Human-Detection-and-Geolocation

[22] H. Ghitri, J. Gutt, W. Seman, and K. Połeć, "UAV-SAR Human Detection and Geolocation Dataset," Roboflow Universe dataset, 2026. [Online]. Available: https://universe.roboflow.com/hamzaghitri/uav-sar-human-detection-dataset