

Article

Automated Victim Detection from UAV Thermal Infrared Imagery for Nighttime Search and Rescue Using Multi-Pose Ground Camera Data

Shiori Kubo ^{1,*} , Koudai Yamada ² and Hidenori Yoshida ³¹ Institute of Education, Research and Regional Cooperation for Crisis Management Shikoku, Kagawa University, Takamatsu 761-0396, Japan² Graduate School of Science for Creative Emergence, Kagawa University, Takamatsu 761-0396, Japan³ Faculty of Engineering and Design, Kagawa University, Takamatsu 761-0396, Japan

* Correspondence: kubo.shiori.a6@kagawa-u.ac.jp; Tel.: +81-87-864-2621

Highlights

What are the main findings?

- We developed an automated person detection method using Grounding DINO, in which thermal infrared imagery from fixed-point ground cameras is adapted to UAV viewpoints through spatial resizing, padding, and position-based augmentation.
- The proposed scheme improved Recall over the baseline in complex, low-contrast environments, with padding and position-based augmentation as the dominant contributors, and it generalized to a completely unseen public dataset through altitude-aware input rescaling without any retraining.

What are the implications of the main findings?

- The training-data construction strategy addresses the scarcity of UAV-based thermal infrared datasets by leveraging easily collectible ground-based imagery, and the approach generalizes to nighttime acquisition.
- The Grounding DINO-based system supports real-time inference at approximately 24 FPS on a desktop-class GPU, providing a practical technical foundation for nighttime UAV-based search and rescue operations.

Abstract

The survival probability of persons requiring rescue after large-scale earthquakes or landslides decreases rapidly with time, and nighttime search operations are constrained by limited visibility. Satellite remote sensing enables wide-area observation at night, but its limited spatial resolution restricts the identification of individual persons. This study develops an automated method for detecting persons in thermal infrared imagery captured by Unmanned Aerial Vehicles (UAVs) using the deep learning model Grounding DINO. To address the limited availability of Unmanned Aerial Vehicle (UAV)-based training data, thermal infrared imagery captured by fixed-point ground cameras was used for training. Spatial resizing and padding were applied to emulate UAV viewpoints and mitigate the ground-aerial domain gap. The proposed model outperformed the baseline on real-world datasets, with substantial Recall improvements in environments with low thermal contrast. An ablation study confirmed that spatial resizing, padding, and position-based augmentation each contribute progressively to detection performance. A comparison with the lightweight YOLOv8n and DETR-based RT-DETR detectors indicated that the Transformer



Academic Editor: Susana Lagüela López

Received: 8 May 2026

Revised: 17 June 2026

Accepted: 1 July 2026

Published: 8 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

architecture alone does not account for reliable detection in scenes with complex thermal noise. This robustness instead derives from the language-grounded semantic priors of Grounding DINO. An extended evaluation on nighttime imagery indicated that the proposed approach generalizes to nighttime acquisition.

Keywords: disaster response; deep learning; unmanned aerial vehicle; human detection; spatial resizing

1. Introduction

In post-disaster search and rescue operations following large-scale earthquakes or landslides, the survival probability of entrapped victims is strongly governed by the time elapsed since the hazardous event. This decline is particularly pronounced for victims trapped beneath collapsed structures or sediment debris. Statistics from the 1995 Great Hanshin-Awaji Earthquake indicate that the survival probability of individuals requiring rescue dropped rapidly from 74.9 % on the day of the disaster to 24.2 % the following day, and further to 5.4 % after 72 h [1,2]. This 72-h threshold is commonly referred to as the “golden hours”, indicating that search and rescue operations within this timeframe are essential for maintaining survival probability.

However, in actual disaster scenarios, search and rescue operations conducted at night are subject to substantial physical and institutional constraints, even within these golden hours. In particular, both the INSARAG Guidelines [3], issued by the International Search and Rescue Advisory Group (INSARAG), and the Safety Management Manual for Firefighting and Emergency Operations [4] prioritize the safety of rescue personnel. Consequently, when adequate safety cannot be ensured because of poor visibility, such operations are suspended or scaled down [5,6], resulting in the loss of numerous life-saving opportunities.

To address this issue, lighting vehicles were deployed to support nighttime search and rescue operations during the 2018 Hokkaido Eastern Iburi Earthquake and the July 2018 Heavy Rain Disaster, although their effective operational range remained limited [7,8]. In the 2016 Kumamoto Earthquake, the main shock occurred at 01:25, which delayed initial damage assessment activities, such as the identification of collapsed houses, until the following morning [9]. In the 2024 Noto Peninsula Earthquake, which occurred at approximately 16:00, complex disasters including widespread building collapses, large-scale fires, and road network disruptions significantly constrained both the nighttime deployment of search and rescue teams and helicopter-based information gathering [10,11].

Against this background, remote sensing technologies have been increasingly employed to acquire wide-area information even at night, without requiring rescue teams or investigators to physically enter hazardous disaster zones. Synthetic Aperture Radar (SAR) is effective for large-scale damage assessment because of its all-weather and day-and-night observation capabilities. As an active remote sensing system that transmits and receives microwave signals, SAR visualizes surface conditions independently of daylight and weather. This capability enables rapid situational assessment immediately after a disaster. For example, Voigt et al. discussed the importance of Satellite-based Emergency Mapping (SEM) and reported that mapping is completed within approximately 6 to 8 h after satellite image acquisition, which allows SEM to be extended to various situations such as extreme weather events and pandemics [12]. Tuele et al. developed a method that automatically and immediately maps flood areas using Sentinel-1 data even in poor weather [13]. The use of satellite imagery such as SAR has therefore enabled relatively

rapid acquisition of information on large-scale disaster conditions. However, because of limitations in observation frequency and resolution, it is unsuitable for local and immediate searches that identify individual victims [12]. Wang et al. proposed a few-shot change-detection method that jointly exploits optical and SAR imagery for rapid disaster response [14]. Arai developed a Generative Adversarial Network (GAN)-based method that converts SAR images into optical images to facilitate landslide-area detection under cloudy or nighttime conditions [15].

In contrast, Unmanned Aerial Vehicles (UAVs), which have recently been adopted across a wide range of application domains, are expected to be effective for victim search operations. They can also acquire high-resolution disaster information from lower altitudes and support the rapid mapping of damage. For example, Restas systematized the use of drones in disaster response. Compared with conventional aircraft and satellites, drones were reported to be highly effective for situational awareness in forest fires and nuclear accidents, and for identifying the locations of victims [16]. Turner et al. further demonstrated that UAVs are effective for high-precision landslide monitoring [17]. VanExel et al. combined UAV and satellite imagery to detect climate-related disasters such as flooding and desertification using deep convolutional networks [18]. Shukla and Shukla integrated UAV-based disaster-zone classification with crowd-density estimation to improve situational awareness in affected areas [19].

Beyond search and rescue, UAV-based deep-learning object detection has been actively investigated in adjacent application domains. For example, UAV-based monitoring has been applied to the inspection of power transmission lines [20], agricultural field surveying [21], and infrastructure damage assessment [22]. These studies share the technical challenge of detecting small or partially occluded targets in aerial imagery, and the methodological insights from these adjacent domains have informed the design of detection systems for disaster response.

Highly maneuverable UAVs are therefore considered suitable for high-frequency data acquisition at close range, and they are expected to play a crucial role in the victim search operations that constitute the focus of this study. For example, Lygouras et al. proposed an automated victim detection framework for search and rescue operations, in which a deep learning-based human detection model was deployed onboard a UAV [23]. Sambolek et al. compared and evaluated deep learning-based person detection methods using UAV aerial imagery and demonstrated the effectiveness of person detection in search and rescue scenarios [24]. Bejiga et al. demonstrated that victims buried beneath snow or rubble can be detected using a pre-trained model as a feature extractor, applied to images captured by UAVs in an environment simulating an avalanche disaster [25]. More recently, Qu et al. addressed the scarcity of post-disaster imagery by using generative adversarial networks to synthesize diverse disaster scenes, which improved human detection in UAV-assisted search and rescue [26].

Robust detection under degraded imaging conditions has also been pursued in UAV-related tasks, where YOLO-RAW enhances the detection of UAVs under rain, sensor noise, and motion blur [27]. Open-world recognition frameworks have likewise been developed to identify previously unseen target categories in remote-sensing imagery, including an incremental ship-recognition framework that extracts unknown classes while mitigating catastrophic forgetting [28]. These two directions, namely robustness to degraded imagery and open-set extensibility, are both pertinent to victim detection in thermal infrared imagery and motivate the detector adopted in this study.

However, most UAVs are equipped with visible-light cameras. This makes them effective in daylight but limits their applicability at nighttime disaster sites. The use of thermal infrared cameras, which operate independently of visible light, is therefore effective

for nighttime search operations. This study aims to develop a method for automatically detecting victims from the air using UAVs equipped with thermal infrared cameras under nighttime search conditions.

The high-precision detection of victims using deep learning models requires a large and diverse dataset acquired under conditions similar to those encountered during operation. However, images captured at disaster sites are extremely scarce [23]. Thermal infrared images are even rarer, and few datasets include poses specific to disaster sites, such as victims lying on the ground or crouching.

In this study, thermal infrared data captured by fixed-point ground cameras serve as a substitute for the rare and costly UAV imagery used in training, because such data can be collected safely and easily. To mitigate the domain gap between ground-view and UAV-acquired aerial imagery, which arises from differences in perspective and distance, we apply spatial resizing and padding to the ground-based imagery to emulate UAV altitudes. We further adopt Grounding DINO, a Transformer-based object detection model, and evaluate its performance under real-world conditions that include diverse poses commonly observed at disaster sites, such as lying down and crouching.

The contributions of this study are threefold. First, we propose a training-data construction strategy that adapts thermal infrared imagery captured by fixed-point ground cameras to the UAV viewpoint through spatial resizing, padding, and position-based augmentation. This strategy addresses the scarcity of UAV-acquired thermal training data for victim detection. Second, we provide a systematic analysis of the ground-to-aerial domain gap, including an ablation study that isolates the contribution of each preprocessing component. Third, we conduct a comprehensive evaluation covering partial occlusion, nighttime acquisition, an external public dataset, and comparisons with Convolutional Neural Network (CNN)-based and DETR-based detectors. The novelty of this study therefore lies in the data-centric adaptation strategy rather than in the detection architecture itself.

2. Methodology

2.1. Study Area and Thermal Infrared Data Acquisition

We assumed a nighttime disaster search and rescue scenario, and captured thermal infrared data using UAVs equipped with thermal infrared cameras, as well as fixed-point ground cameras. All data were captured at the Hayashi-cho Campus of Kagawa University in Japan.

We captured training and validation data using fixed-point cameras at two locations, specifically the courtyard and the sofa area, which are indicated by the pink frames in Figure 1. These environments represent open outdoor and indoor spaces with distinct background temperature distributions. Furthermore, the inclusion of the sofa area enabled the capture of images containing diverse poses, such as crouching, sitting, and lying down, in addition to pedestrians.

Test data were captured using UAVs equipped with thermal infrared cameras at parking lots and sports grounds, which are indicated by the light blue frames in Figure 1. These sites were configured to represent victim search at actual disaster sites. During data capture, the scenes included not only pedestrians but also individuals lying on the ground or partially occluded by objects, representing delayed evacuees or incapacitated persons.

The thermal infrared camera mounted on the UAV operates in the Long-Wave Infrared (LWIR) band. The captured thermal images have a resolution of 1280×960 px and a thermal sensitivity of less than 50 mK. The flight altitude was set to approximately 20 m. This altitude enables smooth screening in residential areas without the need to avoid obstacles, as typical utility poles are 10–15 m high.

The thermal infrared images used for training and validation were captured using fixed-point ground cameras operating in the same LWIR band as the UAV-mounted camera. To reproduce the viewing geometry and height of UAV-based imaging as closely as possible, image capture was conducted at approximately 4.4 m above ground level in the sofa area and approximately 9.8 m in the courtyard. This setup enabled the capture of thermal infrared images containing diverse poses expected at disaster sites, including walking, crouching, sitting, and lying down. These ground-view images were subsequently utilized as training and validation data by applying the preprocessing described later to emulate the person scale observed in UAV-based thermal infrared imagery.

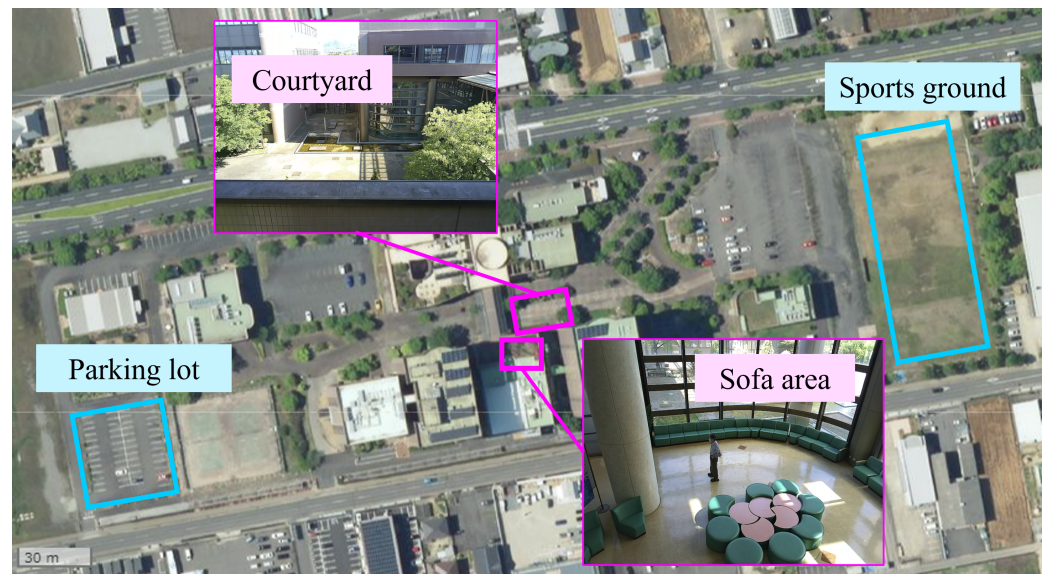


Figure 1. Study sites and experimental data collection areas. The pink frames indicate the locations where training and validation data were captured using stationary thermal infrared cameras, corresponding to the courtyard and the sofa area. The light blue frames represent the areas where test data were captured using a UAV-mounted thermal infrared camera, corresponding to the parking lot and the sports ground. The background aerial photograph was obtained from GSI Maps [29].

2.2. AI-Based Victim Detection Method

We adopted Grounding DINO [30], a Transformer-based object detection model, for victim detection in thermal infrared images. Grounding DINO is an open-set object detection model that enables the detection of arbitrary objects by processing image and text information multimodally.

In addition to its architectural advantages over CNN-based detectors, Grounding DINO was selected over alternative detection paradigms for two practical reasons specific to disaster-response scenarios. First, the open-set detection capability of Grounding DINO enables future extension to disaster-specific concepts such as “injured person” or “crawling victim” through prompt modification, without the need for additional model retraining. By comparison, closed-set detectors such as YOLO [31], Faster R-CNN [32], and recent DETR variants [33] require retraining for any new target class. Second, the integration of the BERT-based text encoder with the visual backbone allows the model to incorporate linguistic priors during detection. This characteristic is particularly beneficial when targets exhibit unusual appearances due to occlusions, pose variations, or low thermal contrast. We further validate this choice empirically through a comparative experiment with YOLOv8n in Section 3.3, where Grounding DINO consistently outperforms the lightweight CNN-based detector in challenging environments such as sports grounds and debris-covered scenes.

Architecturally, Grounding DINO integrates an image backbone based on the Swin Transformer with a text backbone based on BERT, and visual and linguistic features are fused through cross-attention layers within the feature enhancement module. This architecture incorporates global contextual information through self-attention and enables advanced contextual understanding grounded in the concepts specified by text prompts [30]. This capability is expected to provide robust detection for open-set instances that are rare in conventional training datasets, such as persons partially buried under debris or in crouching poses commonly observed at disaster sites, as defined linguistically [34].

The choice of using a single text prompt (“person”) was determined through preliminary experiments. We compared alternative prompt configurations, including disaster-related terms such as “victim” and “human body”, as well as compound prompts such as “person, victim, body”. The single prompt “person” yielded the most stable detection performance with the fewest false positives. This stability can be attributed to the fact that the BERT-based text encoder of Grounding DINO has been pretrained on a large-scale corpus in which “person” provides the most general and well-grounded semantic representation of the human category. We therefore adopted the single prompt “person” throughout the experiments reported in this paper.

To achieve high-precision detection while reducing significant annotation costs, we fine-tuned a pre-trained Grounding DINO model (groundingdino_swint_ogc) using the ground-based data described in Section 2.1.

2.3. Dataset Preparation and Domain Gap Handling

We constructed the dataset by extracting frames containing persons from the captured thermal infrared videos. A domain gap exists between the training and validation data captured by fixed-point ground cameras and the test data captured by UAVs, which is attributed to differences in shooting distance and angle. The primary factor contributing to this domain gap is the mismatch in the number of pixels occupied by the target, referred to as Pixels on Target (PoT). Persons in UAV images captured at an altitude of 20 m appear as extremely small regions relative to those in ground-based images. Using ground-based data directly for training could therefore reduce detection accuracy because of this scale mismatch.

Annotations were generated using the Computer Vision Annotation Tool (CVAT) [35]. Each person visible in the captured frames was enclosed by a tight axis-aligned bounding box covering the entire visible body, including limbs and head. When body parts were partially occluded by objects or other body parts, only the visible portions were enclosed. To ensure consistency, all annotations were performed by a single annotator who applied consistent criteria across the dataset. Frames containing no person, or in which all persons were too heavily occluded to be reliably identified, were excluded from the dataset.

To formalize this mismatch, PoT can be expressed as the number of pixels representing a target area on the ground, which is inversely related to the ground sampling distance (GSD). The GSD can be written as

$$\text{GSD} = \frac{W_{\text{ground}}}{N_{\text{pixel}}} \quad (1)$$

where W_{ground} is the ground-projected width of the sensor footprint and N_{pixel} is the corresponding image width in pixels. Using this definition, PoT can be approximated as

$$\text{PoT} \approx \frac{A_{\text{target}}}{\text{GSD}^2} \quad (2)$$

where A_{target} denotes the ground-projected area of the person. These relationships indicate that PoT decreases rapidly as GSD increases, which makes persons appear as small thermal sources in UAV-based thermal infrared imagery and amplifies the scale mismatch between ground-view and aerial domains.

To address this, the application of spatial resizing and padding to the ground-based imagery (Figure 2) emulates UAV altitudes. We calculated the reduction ratio for the ground-based imagery based on the pixel height of persons in the UAV images.

To determine the spatial resizing factor in a principled manner, the reduction ratio was defined based on the pixel height of a person observed in the UAV-based imagery and that in the ground-based imagery. Specifically, the resizing factor s was computed as

$$s = \frac{P_{\text{UAV}}}{P_{\text{ground}}} \quad (3)$$

where P_{UAV} denotes the pixel height of a person in the UAV-based image and P_{ground} denotes that in the corresponding ground-based image.

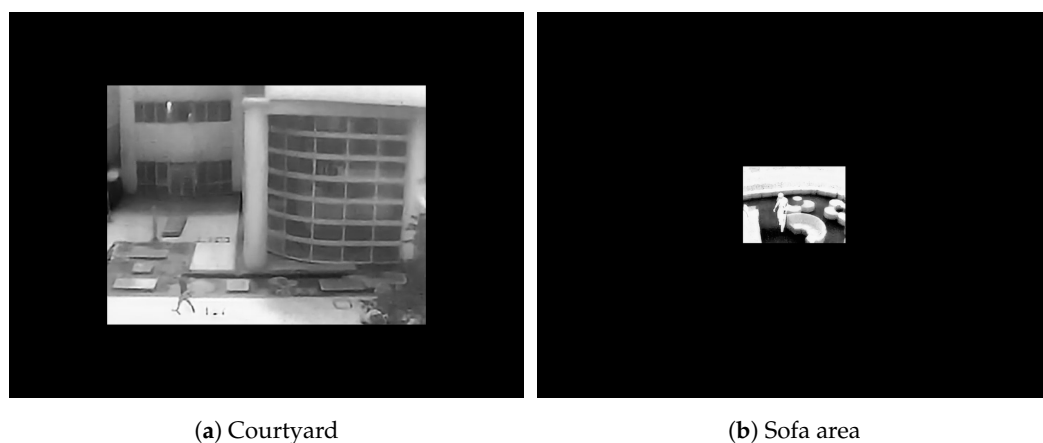


Figure 2. Thermal infrared images after spatial resizing and padding for the courtyard (a) and the sofa area (b). A higher reduction ratio was applied to (b) so that the apparent person size matches the UAV viewpoint at an altitude of 20 m.

We then placed the reduced image in the center of a 480×360 px black background image, corresponding to the model input size, to emulate the appearance of persons from a UAV viewpoint. The proposed spatial resizing process intentionally downsamples person regions in ground-based imagery to emulate a reduction in Pixels on Target (PoT) and optimizes the network to extract meaningful features even from low-resolution thermal sources.

Under our imaging conditions, the ground sampling distance (GSD) at an altitude of 20 m with a spatial resolution of 1280×960 pixels is estimated to be approximately 1.5–2.0 cm per pixel. When the body or upper torso of a person requiring rescue is considered, the target occupies several tens of pixels in the image. Fine details such as limb boundaries and the head are represented as extremely small regions consisting of only a few to several tens of pixels. Such a limited number of pixels constitutes a physical constraint on extracting meaningful features from small targets. Targets smaller than 32×32 pixels, which correspond to the small-object criterion in the MS COCO dataset, are widely recognized to retain few effective spatial features within the network. They are therefore difficult to distinguish from background noise [36]. The proposed spatial resizing and padding strategy intentionally exposes the model to low-resolution targets and aims to optimize the network to achieve high sensitivity to subtle temperature variations.

Thermal infrared imagery can be represented using either of two visualization schemes, namely White Hot, in which high-temperature regions appear bright, and Black Hot,

in which they appear dark. All data were acquired using a unified White Hot scheme. Under this condition, the black regions introduced by the padding process represent backgrounds that emulate extremely low-temperature areas. This setting is expected to encourage the model to learn to distinguish persons, which appear as relatively high-intensity heat sources, as independent local features separated from background noise.

In addition, an important physical constraint in thermal infrared observation is the thermal crossover phenomenon, in which the radiance of the target becomes similar to that of the background [37]. This phenomenon leads to the loss of thermal contrast and consequently makes automatic detection extremely challenging. The proposed domain gap mitigation strategy is also intended to enhance the capability of extracting subtle features under such low-contrast conditions.

For the training data, data augmentation improved robustness against positional changes and expanded the dataset. Five training images were generated from a single image by placing the resized ground-based imagery not only at the center of the background image but also at its four corners, namely the top-right, top-left, bottom-right, and bottom-left (Figure 3). For the validation data, this positional augmentation was excluded, so that only centered resized images were used.

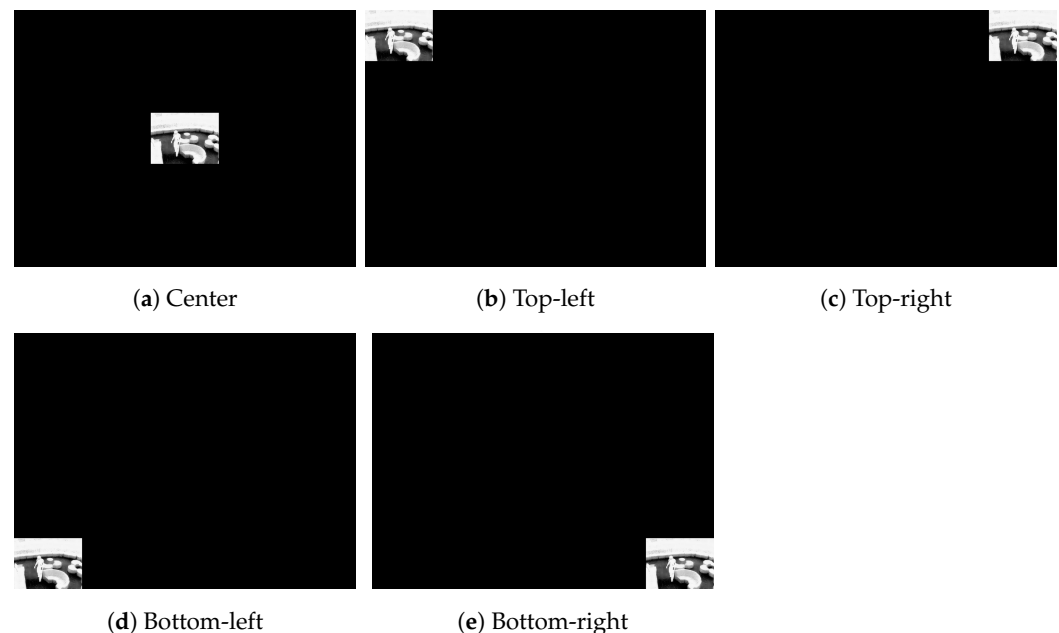


Figure 3. Results of data augmentation after image preprocessing. Five training images were generated from a single original image by placing the resized ground-based imagery at the center and the four corners (top-left, top-right, bottom-left, and bottom-right) of the background image.

Table 1 presents the breakdown of the final dataset. Data augmentation expanded the training data from 1827 to 9135 images, and the validation data consisted of 203 images that were resized only. For the test data, which were designed to represent a search at an actual disaster site, we prepared 20, 10, and 20 images for cases captured by UAVs in a parking lot, a sports ground, and a sports ground with obstacles such as debris, respectively.

2.4. Training Setting and Evaluation Metrics

We used the dataset described in Section 2.3 to train the person-detection model. We adopted AdamW as the optimization algorithm and set the learning rate to 1×10^{-5} , the batch size to 4, and the number of epochs to 50. We set the input image size to 480×360 px and followed the recommended settings of the original model [30] for other hyperparameters. During training, we monitored the loss and detection accuracy on the

validation set. To avoid overfitting, we selected the model weights from epoch 13, when validation performance was optimal, as the final detection model.

Table 1. Breakdown of the constructed dataset used for training, validation, and testing.

Category	Location	Number of Images
Training data	Courtyard	6725
	Sofa area	2410
Validation data	Courtyard	53
	Sofa area	150
Test data	Parking lot	20
	Sports ground	10
	Sports ground (with debris)	20

The baseline model in all experiments denotes the publicly released Grounding DINO (Swin-T backbone) pre-trained on large-scale visible-light image–text data and applied in a zero-shot manner without any additional training on thermal infrared imagery. The identical text prompt (“person”), input resolution, and confidence threshold were used for the baseline and the proposed model. The two models differ only in whether the proposed training-data construction strategy was applied.

During model training, Grounding DINO employs a contrastive learning loss between the image and text modalities [38]. For object localization, an L1 loss is combined with a Generalized Intersection over Union (GIoU) loss that evaluates overlap between predicted and ground-truth bounding boxes, designed to improve regression stability for small targets [38]. In particular, because small coordinate deviations can substantially affect evaluation metrics in small-target detection, this combination of advanced loss functions contributes to improved detection accuracy for small thermal sources that are characteristic of a UAV viewpoint.

The quantitative evaluation of detection performance utilized Precision, Recall, F1-score, Intersection over Union (IoU), and mean Average Precision (mAP), which are standard metrics in object detection tasks. The mAP was reported at an IoU threshold of 0.5 (mAP@0.5) and over the COCO-style range from 0.5 to 0.95 with a step of 0.05 (mAP@0.5:0.95) [36]. For Precision, Recall, F1-score, and IoU, the IoU threshold was set to 0.5 and detections with a confidence score of 0.25 or higher were defined as valid predictions. For mAP, all detections were ranked by confidence score and evaluated following the COCO evaluation protocol.

All training and inference experiments were conducted on a workstation equipped with an NVIDIA GeForce RTX 4080 SUPER GPU with 16 GB of memory, running Ubuntu 24.04 with PyTorch 2.3.1 and CUDA 12.1. The hardware and software configuration used for training and inference is summarized in Table 2. The Grounding DINO models were implemented using the mmdetection framework. The same workstation was used for the inference speed measurements reported in Section 3.6, ensuring a fair comparison between detectors operating on identical hardware. To enable a quantitative comparison with a lightweight CNN-based detector, YOLOv8n was additionally trained and evaluated under an identical dataset configuration. Details of the comparison are provided in Section 3.3. A DETR-based real-time detector, RT-DETR-l (the large variant of RT-DETR) [39], was further trained and evaluated under the same configuration.

Table 2. Hardware and software configuration used for training and inference.

Component	Specification
GPU	NVIDIA GeForce RTX 4080 SUPER (16 GB)
OS	Ubuntu 24.04
CUDA	12.1
PyTorch	2.3.1
Detection framework (Grounding DINO)	mmdetection 3.x
Detection framework (YOLOv8)	Ultralytics 8.x
Input image size	480 × 360 px
Batch size	4
Training epochs	50
Optimizer (Grounding DINO)	AdamW ($\text{lr} = 1 \times 10^{-5}$)
Optimizer (YOLOv8)	AdamW (auto-tuned by Ultralytics)
Confidence threshold	0.25
IoU threshold	0.5

3. Results

3.1. Qualitative Detection Results

Real-world data captured using a UAV-mounted thermal infrared camera were used to validate the effectiveness of the proposed victim detection model. For comparative verification, Figures 4–6 present the detection results obtained in a parking lot, a sports ground, and a sports ground containing debris, respectively. From left to right, each figure shows the ground truth, the detection results of the baseline model without additional training, and the detection results of the proposed model with additional training.

Figure 4 shows the detection results in the parking lot, which featured a uniform background temperature and relatively clear contrast. In this environment, both the baseline and proposed models detected common walking poses, as shown in Figure 4a. However, when persons were captured from a nadir view (i.e., directly overhead) (Figure 4b), the baseline model failed to detect any persons because it could not capture changes in spatial features. The proposed model, in contrast, detected all three persons. Similar results were observed across many other images, in which the proposed model consistently achieved higher detection rates than the baseline. This improvement is likely attributable to the proposed model learning the features of persons represented as small heat sources through the spatial resizing of ground-based imagery.

Furthermore, both models recognized persons in various poses commonly observed at disaster sites, such as sitting with legs extended, lying supine, and crouching (Figure 4c). For the supine person captured from a nadir view in Figure 4d, the baseline model failed to detect the person. The proposed model, by comparison, identified the same individual as two separate bounding boxes corresponding to the upper and lower body parts. Figure 4e contains a person sitting with legs extended and hands on the ground, together with a person curled up on the side. The baseline model detected both persons, whereas the proposed model detected the seated person only at the lower body and the curled-up person as two overlapping boxes covering the entire body. These behaviors can be attributed to the proposed model prioritizing local high- and low-intensity areas characteristic of thermal infrared images over the overall human body shape during training. Because the training data included persons carrying bags or umbrellas whose bodies were partially occluded or fragmented, the proposed model learned to identify fragmented and discontinuous heat-source regions as independent persons. Consequently, it tended to detect individual regions as separate persons when similar discontinuities arose from changes in body pose, as observed in Figure 4d,e.

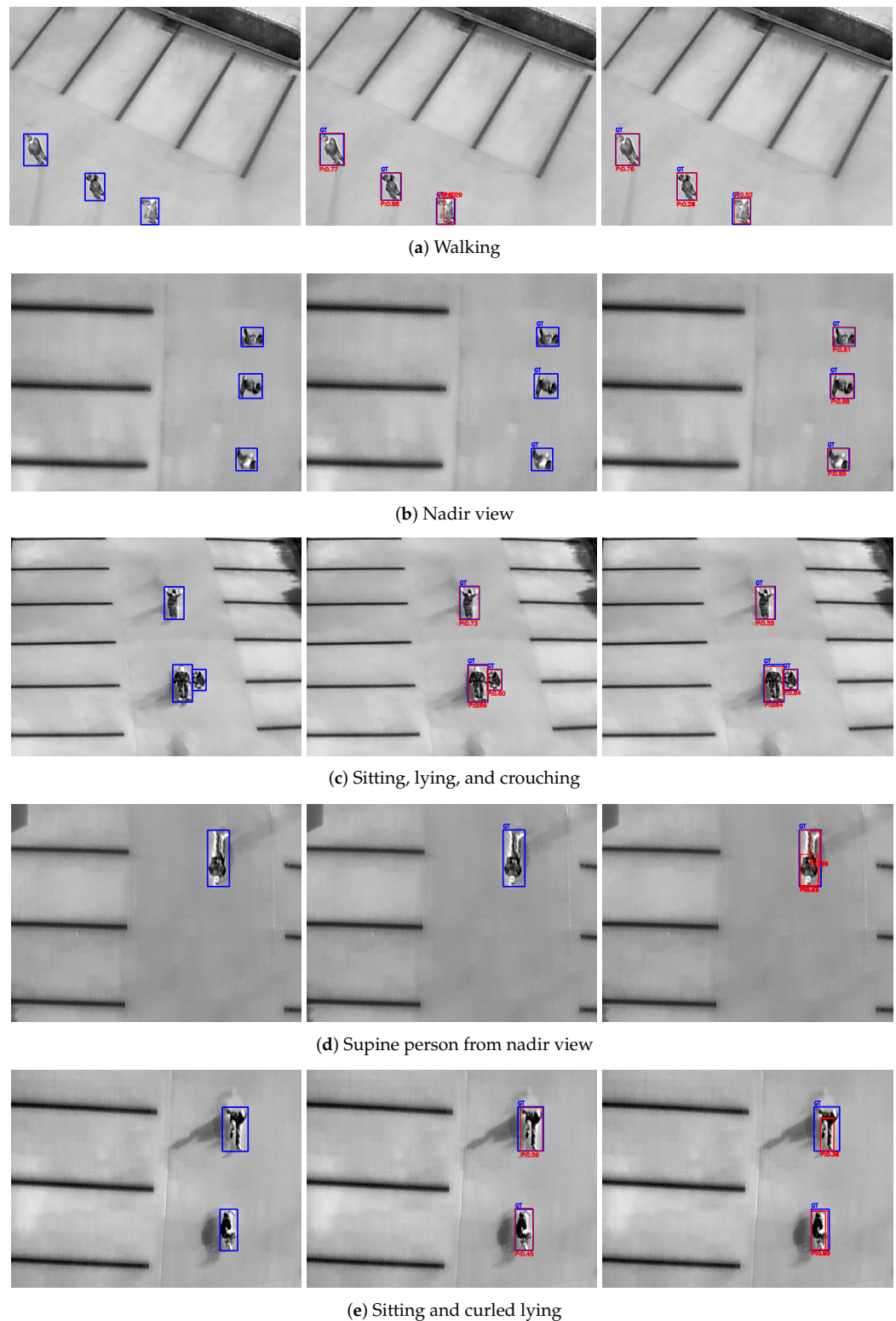


Figure 4. Comparison of person detection results in the parking lot. Columns from left to right show the ground truth, the baseline model (without additional training), and the proposed model (with training data preprocessing). Blue boxes denote ground-truth annotations and red boxes denote model predictions.

Figure 5 shows the detection results in the sports ground, where grass and soil coexist and the temperature distribution is noisy. In this environment, the detection performance of the baseline model deteriorated markedly. As shown in Figure 5a, the baseline model failed to detect either of the two persons and instead produced a single bounding box covering

the entire image. A similar behavior was observed in Figure 5b, where the baseline model failed to detect any of the three walking persons and again produced a single full-image box. In Figure 5c, the baseline model detected one of three persons and additionally produced a single full-image box. This behavior can be attributed to the domain gap that arises when models trained on visible-light images are applied to thermal infrared imagery. This domain gap limits the ability to distinguish human features, particularly under low thermal contrast against the background.

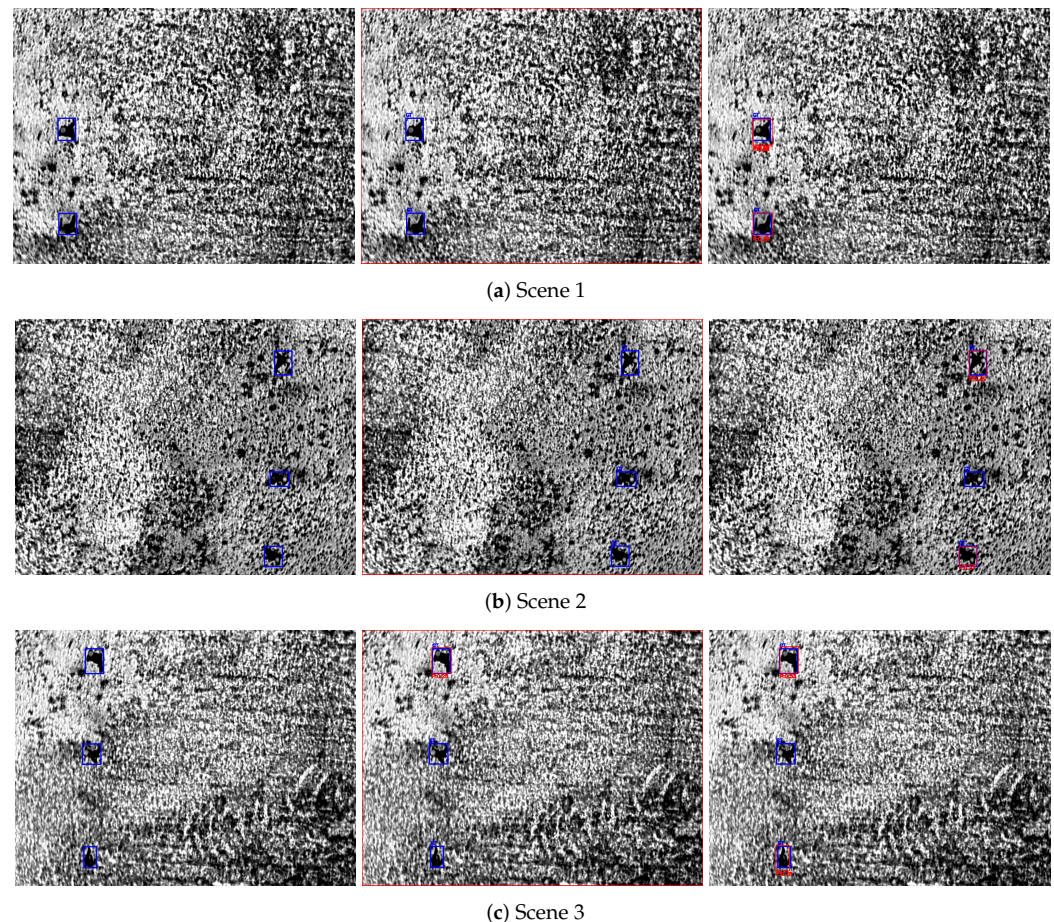


Figure 5. Comparison of person detection results in the sports ground. Columns from left to right show the ground truth, the baseline model, and the proposed model. Blue boxes denote ground-truth annotations and red boxes denote model predictions.

The proposed model, by comparison, detected both persons in Figure 5a, although the upper person was detected twice at the same location. It also detected two of the three persons in both Figure 5b,c without producing any over-segmented box covering the entire image. However, detection often remained challenging under thermal crossover, where the body surface temperature assimilates with the ground temperature, as seen in the undetected persons in Figure 5b,c.

Furthermore, Figure 6 shows a sports ground similar to Figure 5, in which persons are covered by cardboard or placed on top of it. For a person whose lower body was completely covered, the baseline model identified an extended region that included the heat-accumulated obstacles (Figure 6a). The proposed model, in contrast, identified only the exposed upper body once. These results suggest that the baseline model treated low-intensity regions including obstacles as persons, whereas the proposed model distinguished the unique thermal features of each body part and excluded the surrounding obstacles.

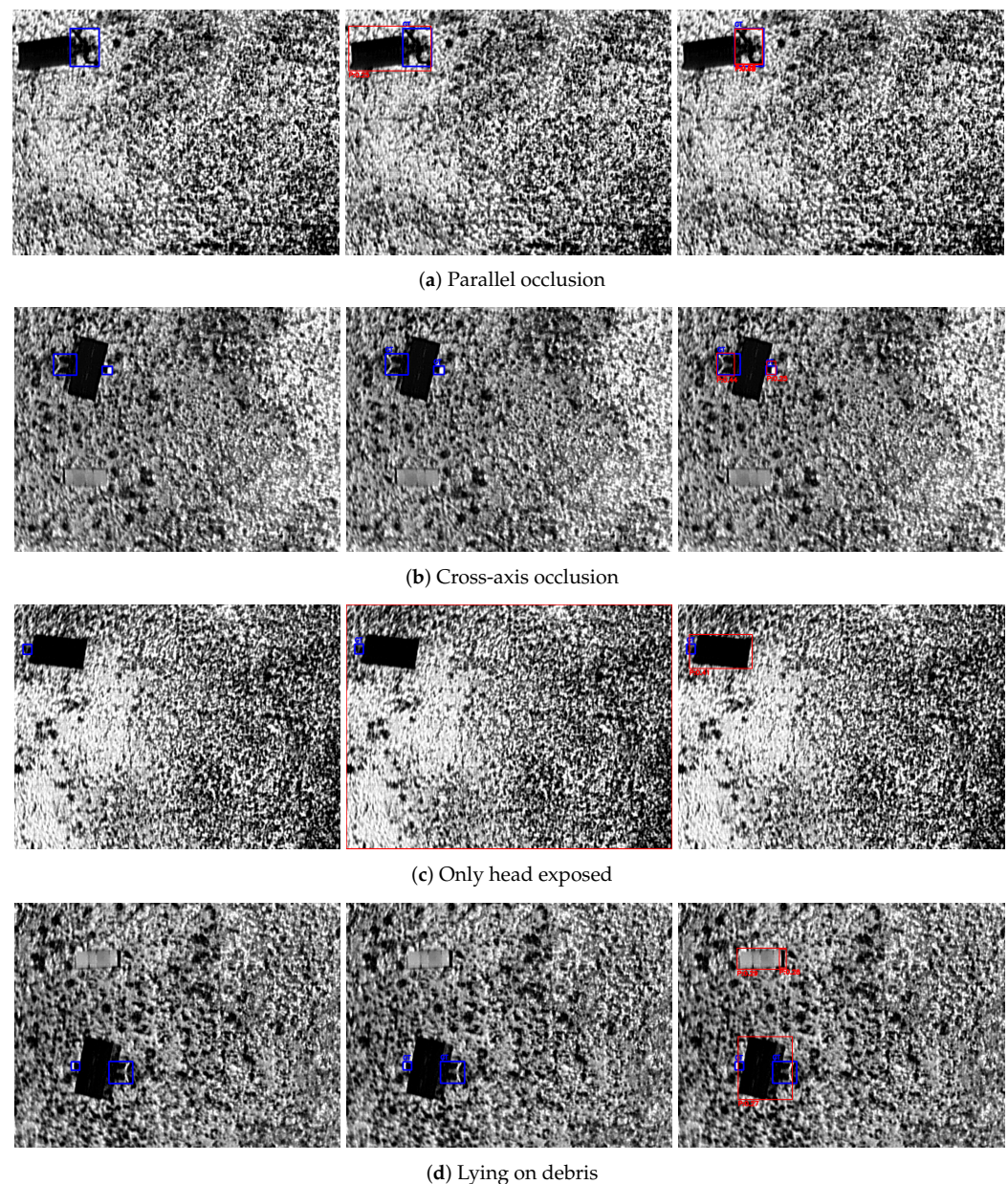


Figure 6. Comparison of person detection results in the sports ground with arranged debris. Columns from left to right show the ground truth, the baseline model, and the proposed model. Blue boxes denote ground-truth annotations and red boxes denote model predictions.

Regarding Figure 6b, where the torso is covered and only the head and feet are exposed, the baseline model failed to detect the person because of the extremely small exposed area. The proposed model, by comparison, detected both the exposed head and the feet as two separate detections with characteristic thermal shapes. In Figure 6c, where the entire body except the head is covered by obstacles, the baseline model failed to detect the person and instead produced a false positive covering the entire image. In this case, the proposed model could not isolate the exposed body part as an independent person. It instead identified a single region encompassing both the exposed head and the overlapping debris as one detection. Although the proposed model detected persons under partial occlusion better than the baseline, Figure 6c shows that precisely isolating individual body parts remains challenging when the exposed area is extremely limited.

In Figure 6d, the upper part of the scene contains cardboard placed alone without any person, whereas the lower part contains a person whose midsection is covered by

debris. The baseline model failed to detect the person. The proposed model detected the partially occluded person together with the surrounding obstacles, but it also produced a false positive on the isolated cardboard. This behavior indicates that the proposed model occasionally responds to obstacle materials whose thermal characteristics resemble the debris associated with persons in the training data. This response represents a residual limitation when isolated obstacles share thermal patterns with debris-occluded persons.

3.2. Quantitative Detection Results

A quantitative evaluation was conducted using the test data to validate the effectiveness of the proposed model. Five standard object-detection metrics were employed, specifically Precision, Recall, F1-score, IoU, and mean Average Precision (mAP). Table 3 shows the detection performance of each model under three different environments, namely a parking lot, a sports ground, and a sports ground with obstacles such as debris.

Table 3. Performance comparison of person detection between the baseline and proposed models across three scenarios.

Test Scenario	Model	Precision	Recall	F1-Score	IoU	mAP@0.5	mAP@0.5:0.95
Parking lot	Baseline	0.9459	0.7609	0.8434	0.9339	0.914	0.791
	Ours	0.6610	0.8478	0.7429	0.8817	0.824	0.649
Sports ground	Baseline	0.1667	0.0800	0.1081	0.8410	0.221	0.125
	Ours	0.4762	0.4000	0.4348	0.6830	0.471	0.246
Sports ground (with debris)	Baseline	0.0741	0.0741	0.0741	0.6267	0.071	0.031
	Ours	0.2500	0.2222	0.2500	0.6597	0.185	0.056

These evaluation metrics were calculated based on the correspondence between bounding boxes predicted by the model and the ground truth annotations. Predictions that sufficiently overlapped with ground truth annotations were defined as True Positives (TP), predictions made in the absence of ground truth annotations were defined as False Positives (FP), and undetected ground truth instances were defined as False Negatives (FN).

Precision represents the proportion of correctly detected persons among all detected regions and is calculated as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (4)$$

Recall represents the proportion of existing persons that were correctly detected and is defined as

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (5)$$

The F1-score is the harmonic mean of Precision and Recall and is used to evaluate the balance between these two metrics, which is calculated as

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6)$$

In addition, IoU quantifies the degree of overlap between predicted and ground truth bounding boxes and is defined as the area of their intersection divided by the area of their union. As described in Section 2.4, predictions with an IoU of 0.5 or higher were regarded as True Positive, and only detections with a confidence score of 0.25 or higher were considered valid predictions in the evaluation.

In the parking lot, which features a relatively uniform background composed primarily of asphalt, the baseline model achieved a Precision of 0.9459 and a Recall of 0.7609, whereas the proposed model improved Recall to 0.8478. The improvement in Recall demonstrates

that the use of preprocessed training data reduced missed detections for persons with altered spatial features, including cases captured from a nadir viewpoint (Section 3.1). The baseline model achieved a mAP@0.5 of 0.914 and a mAP@0.5:0.95 of 0.791, exceeding the corresponding values of 0.824 and 0.649 for the proposed model. This trend is consistent with the lower Precision of the proposed model in this environment. In this simple scene, where targets are well separated from the background, the conservative detection style of the baseline yields a higher mAP. The proposed model, in contrast, trades some Precision for a substantially improved Recall.

Next, in the sports ground, where grass and soil are intermixed and thermal-distribution noise is present, the Precision and Recall of the baseline model decreased markedly to 0.1667 and 0.0800, respectively. Consistent with the observations in Section 3.1, the baseline model frequently failed to distinguish persons because of the low thermal contrast against the background. This limitation appeared as the misclassification of the entire image as a single bounding box and led to many missed detections. The proposed model, by comparison, achieved a Precision of 0.4762 and a Recall of 0.4000, which represents an improvement in Recall of approximately five times that of the baseline. The mAP@0.5 also increased from 0.221 to 0.471, and the mAP@0.5:0.95 increased from 0.125 to 0.246, indicating consistent improvements across IoU thresholds. These results indicate that the proposed model performs stable person detection by exploiting local high- and low-intensity regions in thermal infrared imagery, even under challenging conditions such as thermal crossover.

Furthermore, in the sports-ground cases where debris was positioned, the Precision, Recall, and F1-score of the baseline model were all extremely low at 0.0741. As described in Section 3.1, the baseline model frequently recognized both persons and rectangular obstacles as human features. This produced detections that covered areas including the obstacles or, in some cases, misidentified the entire image as a single bounding box. These remarkably low values demonstrate the vulnerability of the baseline model in failing to distinguish person-specific geometric characteristics. By comparison, the proposed model maintained a certain level of accuracy even where persons were partially occluded. This performance can be attributed to the model's ability to recognize discontinuous heat-source regions individually, such as limbs and the head, and to exclude obstacles appropriately. The mAP@0.5 increased from 0.071 to 0.185, and the mAP@0.5:0.95 increased from 0.031 to 0.056, confirming the same trend in challenging environments where partial occlusions and thermal noise dominate.

However, the IoU for the proposed model was 0.6597, only slightly exceeding the 0.6267 recorded by the baseline. This small difference can be attributed to the inclusion relationship between the bounding boxes produced by the two models. The baseline model's large false detections inevitably overlapped with the ground truth regions, leading to IoU values that exceeded the threshold. The proposed model identified only localized heat sources consistent with the ground truth, where even slight coordinate displacements substantially reduced the IoU. As a consequence, the numerical IoU values of the two models did not differ substantially despite their qualitative differences.

3.3. Comparison with CNN-Based and DETR-Based Detectors

To quantitatively compare the proposed approach with a widely used lightweight detector, we additionally trained YOLOv8n [31] under the same dataset configuration as Grounding DINO. The YOLOv8n model was trained on the proposed preprocessed dataset and evaluated on the same three test environments as in Section 3.1. These environments were the parking lot, the sports ground, and the sports ground containing debris. The evaluation metrics included Precision, Recall, mAP@0.5, and mAP@0.5:0.95 [36].

Table 4 presents the quantitative results. The confidence threshold was set to 0.25, which is the default value commonly used in object detection benchmarks and balances detection sensitivity against the false-positive rate. In the parking lot, YOLOv8n attained a Precision of 0.66 and a Recall of 0.58, whereas Grounding DINO reached a Precision of 0.66 and a Recall of 0.85. In the sports ground with mixed grass and soil, YOLOv8n yielded a mAP@0.5 of 0.00 and localized no persons above the confidence threshold. The same outcome was observed in the sports ground containing debris. By comparison, Grounding DINO achieved mAP@0.5 values of 0.47 and 0.19 in these two environments, respectively.

These results indicate that the lightweight CNN-based detector cannot reliably localize persons in thermal infrared imagery containing complex background patterns. The Transformer-based architecture of Grounding DINO, combined with its language-grounded feature representation, provides robustness against thermal background noise that the lightweight CNN does not match.

Table 4. Quantitative comparison among YOLOv8n, RT-DETR-l, and Grounding DINO under the proposed preprocessing scheme.

Test Scenario	Model	Precision	Recall	mAP@0.5	mAP@0.5:0.95
Parking lot	YOLOv8n	0.66	0.58	0.50	0.27
	RT-DETR-l	0.51	0.09	0.08	0.06
	Grounding DINO	0.66	0.85	0.82	0.65
Sports ground	YOLOv8n	0.00	0.00	0.00	0.00
	RT-DETR-l	0.00	0.00	0.00	0.00
	Grounding DINO	0.48	0.40	0.47	0.25
Sports ground (with debris)	YOLOv8n	0.00	0.00	0.00	0.00
	RT-DETR-l	0.00	0.00	0.00	0.00
	Grounding DINO	0.25	0.22	0.19	0.06

To examine whether this limitation is specific to CNN-based architectures, RT-DETR-l [39], a representative real-time DETR-based detector, was trained and evaluated under conditions identical to those of YOLOv8n. As reported in Table 4, RT-DETR-l attained a mAP@0.5 of only 0.08 in the parking lot and failed to detect any person in the sports ground and debris-covered environments. The closed-set detectors therefore exhibited the same failure pattern regardless of whether the backbone was convolutional or Transformer-based. This observation indicates that the robustness of Grounding DINO derives primarily from the semantic priors acquired through large-scale image–text pre-training, rather than from the Transformer architecture itself.

Figure 7 presents the qualitative comparison between YOLOv8n and the proposed Grounding DINO model under the same proposed preprocessing scheme. In the parking lot (a), YOLOv8n detected only one of the three persons captured from a nadir view, whereas the proposed model detected two of the three. On the open sports ground (b), YOLOv8n produced dense false positives across the grass area, whereas the proposed model maintained stable detection of the persons. On the sports ground with arranged debris (c, d), YOLOv8n generated dense false positives across the entire scene, whereas the proposed model detected the persons with only minor bounding-box deviations. This behavior indicates that the lightweight CNN-based detector relies more heavily on full-body spatial features and produces scattered false detections in scenes containing complex thermal background patterns. By comparison, the Transformer-based architecture of the proposed model handles partial occlusions and thermal distractors more robustly under the same preprocessing scheme.

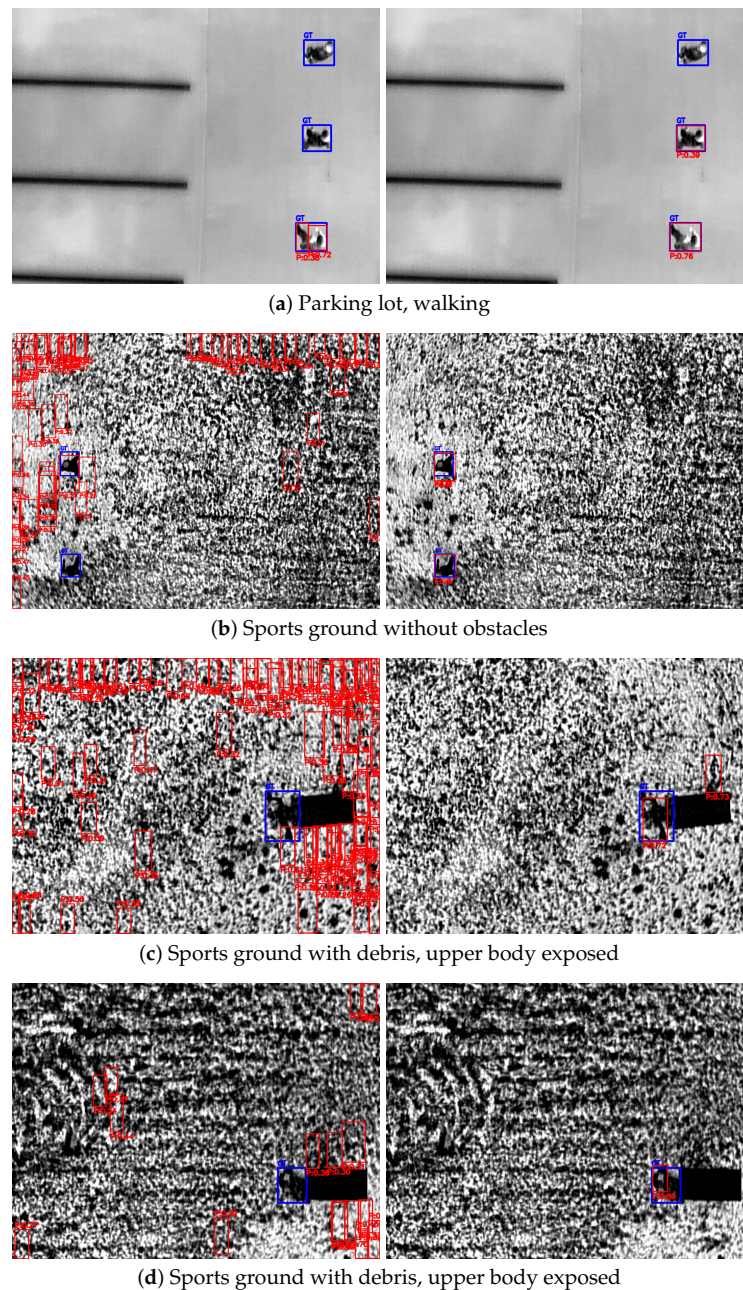


Figure 7. Qualitative comparison between YOLOv8n (**left**) and the proposed Grounding DINO model (**right**) under the same proposed preprocessing scheme. The corresponding ground truth is shown in Figures 4–6. Red boxes denote model predictions.

3.4. Ablation Study

To quantitatively evaluate the contribution of each component in the proposed preprocessing scheme, we conducted an ablation study under four settings defined as follows. None used the original ground-based imagery. Resize applied the spatial resizing factor s to the ground-based imagery, and the resulting smaller images were stored without padding. Resize + Padding applied both spatial resizing and padding to a 480×360 px black background, with the resized image placed at the center only. Full corresponded to the complete proposed scheme, in which the resized image was placed at the center and at the four corners of the background, yielding five training samples per source image. These settings were chosen to isolate the contribution of three preprocessing components, namely spatial resizing alone, the addition of padding, and the addition of position-based augmentation.

Both Grounding DINO and YOLOv8n were trained under each of the four settings, together with RT-DETR-l. The validation set, the test set, and the training hyperparameters were kept identical across all settings to isolate the effect of preprocessing. Table 5 reports the resulting mAP@0.5 values across all combinations.

Table 5. Ablation study results (mAP@0.5) across four preprocessing settings and three test environments. None uses the original ground-based imagery. Resize applies spatial resizing only, without padding. Resize + Pad applies spatial resizing and center padding to 480×360 px. Full is the proposed scheme, which adds five-position augmentation.

Test Scenario	Model	Preprocessing Setting (mAP@0.5)			
		None	Resize	Resize + Pad	Full
Parking lot	YOLOv8n	0.66	0.52	0.51	0.50
	RT-DETR-l	0.83	0.70	0.73	0.08
	Grounding DINO	0.88	0.90	0.77	0.82
Sports ground	YOLOv8n	0.00	0.00	0.00	0.00
	RT-DETR-l	0.00	0.00	0.00	0.00
	Grounding DINO	0.11	0.14	0.29	0.47
Sports ground (with debris)	YOLOv8n	0.00	0.00	0.00	0.00
	RT-DETR-l	0.00	0.00	0.00	0.00
	Grounding DINO	0.05	0.04	0.10	0.19

For Grounding DINO in the parking lot, mAP@0.5 showed only minor variation across settings, ranging from 0.77 to 0.90, which reflects the relative ease of detection in this uniform-background environment. In the sports ground environment, mAP@0.5 increased from 0.11 under None to 0.47 under Full, corresponding to a 4.3-fold improvement. In the sports ground containing debris, the corresponding improvement was 3.9-fold, with mAP@0.5 rising from 0.05 to 0.19. The transition from None to Resize yielded only a minor change, whereas the subsequent transitions to Resize + Pad and to Full produced the largest improvements.

These results indicate that padding and position-based augmentation are the two dominant contributors to the observed improvement under challenging environmental conditions. The behavior of YOLOv8n further indicates that the proposed preprocessing alone is insufficient when the backbone lacks the capacity to model global context across complex thermal scenes.

Figure 8 illustrates the qualitative effects of each preprocessing component across three environments. The parking lot contained three persons in diverse poses, including curled-up, legs-extended seated, and supine postures. Conditions None and Resize detected all three persons with closely matched bounding boxes, whereas Resize + Pad and Full produced over-segmented detections for some persons. This trend is consistent with the closely matched mAP@0.5 scores reported in Table 5, which indicates that the proposed preprocessing has limited effect in simple scenes. On the open sports ground, conditions None and Resize failed to detect most persons and produced a single bounding box covering the entire image, whereas Resize + Pad introduced a false positive on the grass area that was subsequently suppressed in Full. On the sports ground with arranged debris, None failed entirely, Resize produced a single full-image box, Resize + Pad detected the upper body together with a false positive on adjacent grass, and Full constrained the box around the exposed body parts. These observations indicate that the contribution of preprocessing components becomes pronounced in environments with occlusions, scale variation, or thermal distractors, while remaining limited in simple scenes.

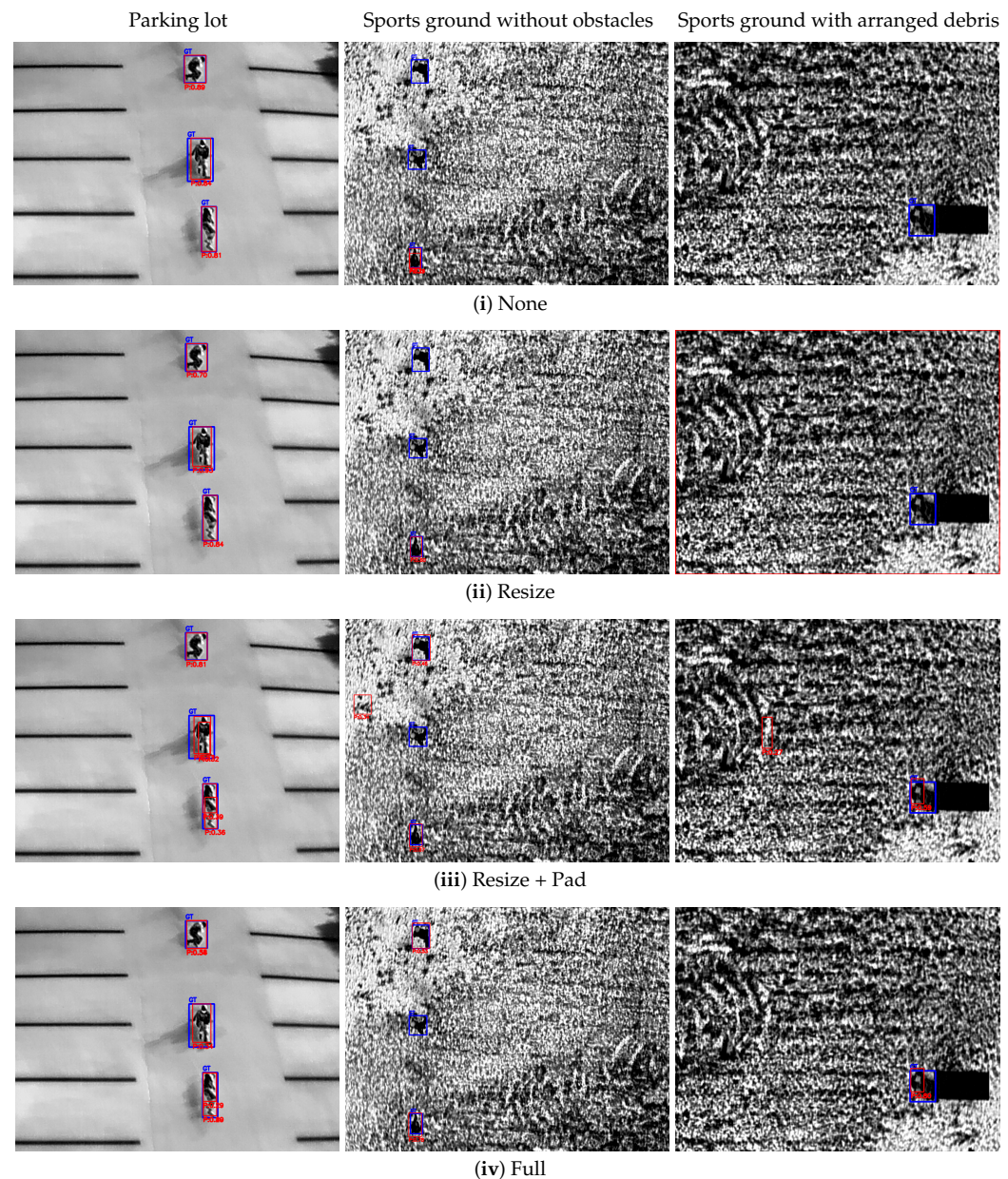


Figure 8. Qualitative comparison of the four ablation conditions for the proposed Grounding DINO model. From top to bottom, the rows show the None, Resize, Resize + Pad, and Full settings, and the columns correspond to the parking lot, the open sports ground, and the sports ground with arranged debris. Blue boxes denote ground truth and red boxes denote model predictions.

3.5. Extended Evaluation Under Nighttime Conditions

To evaluate the generalization capability of the proposed approach to nighttime acquisition, we constructed an additional test set captured at night. The test set comprised 30 thermal infrared images acquired by the UAV-mounted thermal camera at night in a sports ground environment. The captured scenes included diverse poses such as walking and lying down, as well as persons partially occluded by objects or by other body parts. The persons in each frame were annotated using axis-aligned bounding boxes following the procedure described in Section 2.3. The proposed Grounding DINO model and the YOLOv8n comparison model were both evaluated on this nighttime test set under all four ablation settings, without any additional fine-tuning. Table 6 presents the results.

Table 6. Extended evaluation under nighttime conditions (mAP@0.5). The test set consists of 30 thermal infrared images captured at night in a sports ground environment with partial occlusions and diverse poses. The four preprocessing settings (None, Resize, Resize + Pad, Full) follow the same definitions as in Table 5.

Model	Preprocessing Setting (mAP@0.5)			
	None	Resize	Resize + Pad	Full
YOLOv8n	0.00	0.07	0.04	0.03
RT-DETR-l	0.03	0.03	0.00	0.00
Grounding DINO	0.17	0.22	0.18	0.32

For Grounding DINO, the Full setting yielded a mAP@0.5 of 0.32 on the nighttime test set, which exceeded the corresponding value of 0.19 on the daytime sports ground containing debris. This result can be attributed to the fact that nighttime thermal imaging enhances the temperature contrast between cooled background surfaces and warm body parts. The enhanced contrast supports stable person localization under nighttime acquisition. The same monotonic improvement across preprocessing settings observed in the daytime experiments was also reproduced, with mAP@0.5 increasing from 0.17 under None to 0.32 under Full.

For YOLOv8n, all settings yielded mAP@0.5 below 0.07. This outcome is consistent with the limitations of the lightweight CNN-based detector observed in the daytime sports ground and debris-covered environments. RT-DETR-l likewise remained at or below a mAP@0.5 of 0.03 across all four settings, consistent with its failure in the daytime sports-ground and debris-covered environments (Section 3.4).

Together with the qualitative observations described above, these results indicate that the proposed preprocessing scheme enables effective generalization to nighttime acquisition. The proposed approach demonstrates a certain level of effectiveness in scenarios that combine nighttime acquisition, partial occlusions, and pose variation.

Due to logistical constraints, the nighttime data acquisition was conducted only at the sports ground. The captured scenes covered three representative situations, namely two persons lying on cardboard with partial occlusion by debris, three persons walking captured from a nadir view, and a person lying on the ground. The ground thermal appearance differed slightly between the daytime and nighttime acquisitions, which indicates a residual domain gap. Figure 9 presents the qualitative detection results across these scenes. In (a), the baseline model detected only a small body part of one person, whereas the proposed model detected both persons on the cardboard. In (b), the baseline model detected only one of three walking persons, whereas the proposed model detected two of the three. In (c), the baseline model produced a single bounding box covering the entire image, whereas the proposed model detected the lying person with a tight bounding box. These observations indicate that the advantage of the proposed scheme over the baseline is maintained under nighttime acquisition. This suggests that the preprocessing components mitigate not only spatial domain gaps but also thermal variations between daytime and nighttime acquisitions.

3.6. Inference Speed and Real-Time Feasibility

The computational efficiency of the proposed approach was evaluated by measuring the per-frame inference time of both Grounding DINO and YOLOv8n on the same hardware described in Section 2.4. The measurement followed a standard benchmarking protocol that included 20 warmup iterations followed by 100 timed iterations on the same thermal infrared image with a single inference batch. The timing window covered preprocessing, the model forward pass, and postprocessing for both models. Image

disk read time became negligible after warmup due to operating-system file caching, and `torch.cuda.synchronize()` was invoked before and after each timed iteration to ensure accurate GPU timing.

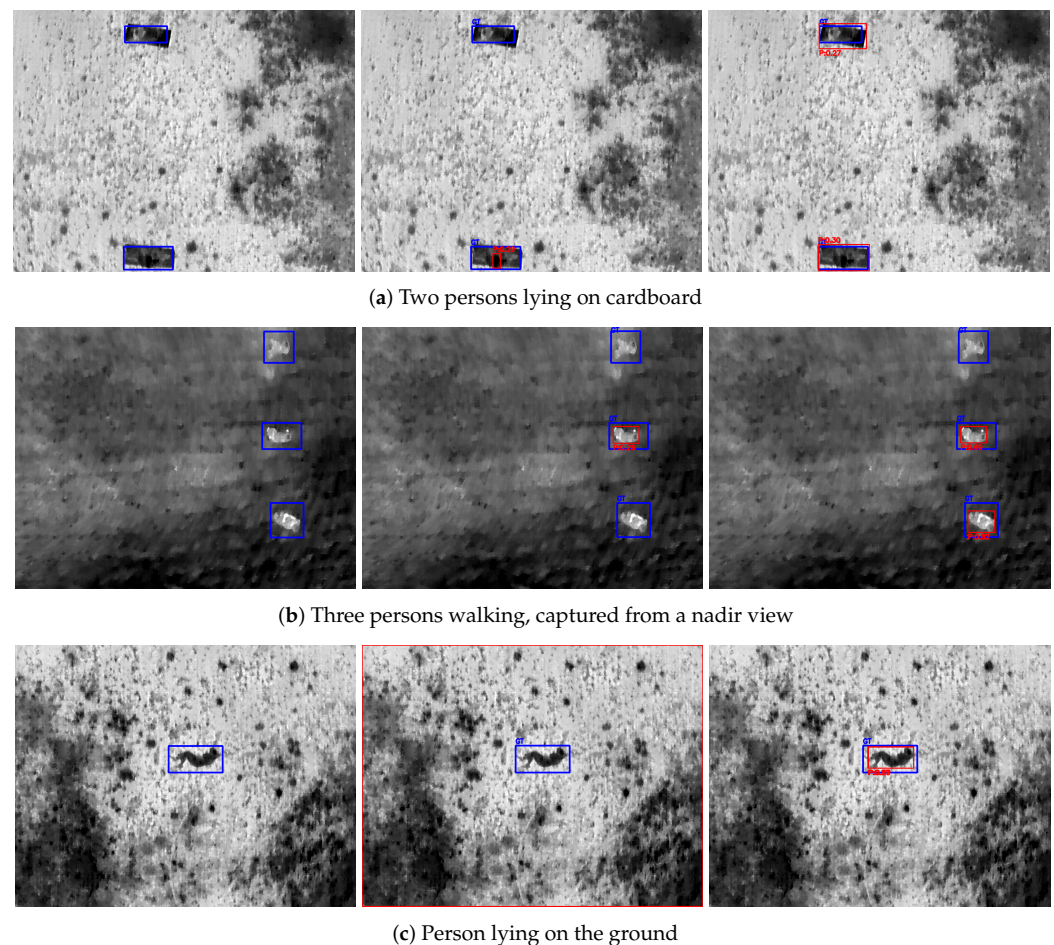


Figure 9. Qualitative detection results under nighttime conditions at the sports ground. Columns from left to right show the ground truth, the baseline model, and the proposed model. Blue boxes denote ground-truth annotations and red boxes denote model predictions.

Table 7 reports the mean inference time and the corresponding throughput in frames per second (FPS). Grounding DINO required approximately 41.3 ms per frame, corresponding to 24.2 FPS. YOLOv8n required approximately 2.2 ms per frame, corresponding to 457.4 FPS. The throughput of YOLOv8n was approximately 19 times higher than that of Grounding DINO on identical hardware. RT-DETR-l required approximately 21.1 ms per frame, corresponding to 47.4 FPS, under the same protocol. Table 7 additionally reports the number of parameters, the floating point operations (FLOPs) computed at the inference input resolution, and the model size of each detector. RT-DETR-l incurs a computational cost of the same order as Grounding DINO. Nevertheless, only Grounding DINO maintained detection capability in the challenging environments (Section 3.3), which indicates that detection robustness in this task is not determined by computational budget.

These results indicate that the proposed Grounding DINO-based system can support real-time processing on a desktop-class GPU. The lightweight YOLOv8n provides a substantial throughput margin suitable for embedded deployment on UAV platforms equipped with edge GPUs. The choice between the two models therefore depends on the operational priorities of the disaster-response scenario. Grounding DINO is preferable when detection performance is prioritized, whereas YOLOv8n is preferable when inference speed is prioritized.

The proposed method is intended for near-real-time processing at a ground station, in which thermal infrared video transmitted from the UAV is processed on a GPU-equipped ground station. The measured throughput of approximately 24 FPS satisfies this requirement. Onboard processing on resource-constrained edge devices is not assumed in the present study. It would require model compression techniques such as knowledge distillation or quantization, which remain for future work.

Table 7. Efficiency comparison on NVIDIA RTX 4080 SUPER (input size 480×360 px, batch size 1). FLOPs were computed at the same input resolution. The value for Grounding DINO is a lower-bound estimate because several custom operators are not counted by the profiler.

Model	Parameters	FLOPs (G)	Model Size	Mean Time (ms)	FPS
YOLOv8n	3.0 M	4.6	5.9 MB	2.19	457.4
RT-DETR-l	32.8 M	62.2	63.1 MB	21.11	47.4
Grounding DINO	173.5 M	≈ 77.1	660 MB	41.26	24.2

3.7. Generalization to an Unseen External Dataset

To examine the generalization capability beyond the data collected in this study, both the baseline and the proposed models were evaluated on the publicly available HIT-UAV dataset [40]. HIT-UAV is a high-altitude infrared thermal dataset acquired by an independent group using a different UAV platform and thermal sensor over schools, parking lots, roads, and playgrounds, at flight altitudes of 60–130 m under both daytime and nighttime conditions. The test split comprises 579 images containing 2611 person instances. Neither model was retrained or tuned on this dataset, and the evaluation pipeline, an input resolution of 480×360 px, the text prompt, and the confidence threshold were identical to those of the main evaluation. All images were therefore completely unseen by both models.

Table 8 summarizes the results stratified by flight altitude and illumination. The baseline achieved a mAP@0.5 of 0.281 on the full test split, and its performance decreased monotonically with flight altitude, from 0.567 at 60 m to 0.024 at 100–130 m. Nighttime imagery yielded substantially higher performance than daytime imagery, with mAP@0.5 values of 0.356 and 0.060, respectively. This difference is attributable to the enhanced thermal contrast between cooled background surfaces and warm bodies, and it supports the nighttime orientation of this study. On the subset closest to the operational assumption of this study, namely 60 m at night, the baseline reached 0.635. By contrast, the proposed model degraded severely under the default input scale, reaching only 0.021 on the full split and thereby exhibiting a specialization–generalization trade-off. Specifically, fine-tuning with the proposed preprocessing specializes the model to the apparent person scale of the assumed flight altitude of approximately 20 m, which is approximately one third of the lowest altitude in HIT-UAV.

To isolate the cause of this degradation, the input resolution was varied while keeping all other settings fixed (Table 9). The performance of the proposed model recovered monotonically with the input scale, from 0.021 to 0.214 on the full split and from 0.093 to 0.474 on the 60 m nighttime subset at a $2.0\times$ input. At this input scale, it exceeded the baseline on the same subset, which decreased from 0.635 at the default scale to 0.458 under the same upscaling. The baseline likewise did not benefit from input upscaling on the full split, changing only from 0.281 to 0.228, which confirms that the recovery is specific to the scale prior learned through the proposed preprocessing rather than a generic effect of higher resolution. At a $3.0\times$ input, the performance of the proposed model collapsed to a mAP@0.5 of 0.013, which is attributable to resolution extrapolation beyond the trained positional-encoding range. These results identify the apparent person scale, expressed in

Pixels on Target, as the dominant factor governing cross-altitude generalization. Because the flight altitude is known during UAV operations, the input scale can be adjusted accordingly. This altitude-aware rescaling recovers most of the scale-induced degradation without any retraining. Figure 10 illustrates these effects qualitatively. The fine-tuned proposed model misses most persons at the default input but recovers them when the input is rescaled, whereas Figure 10c shows a case in which the recovery remains incomplete.

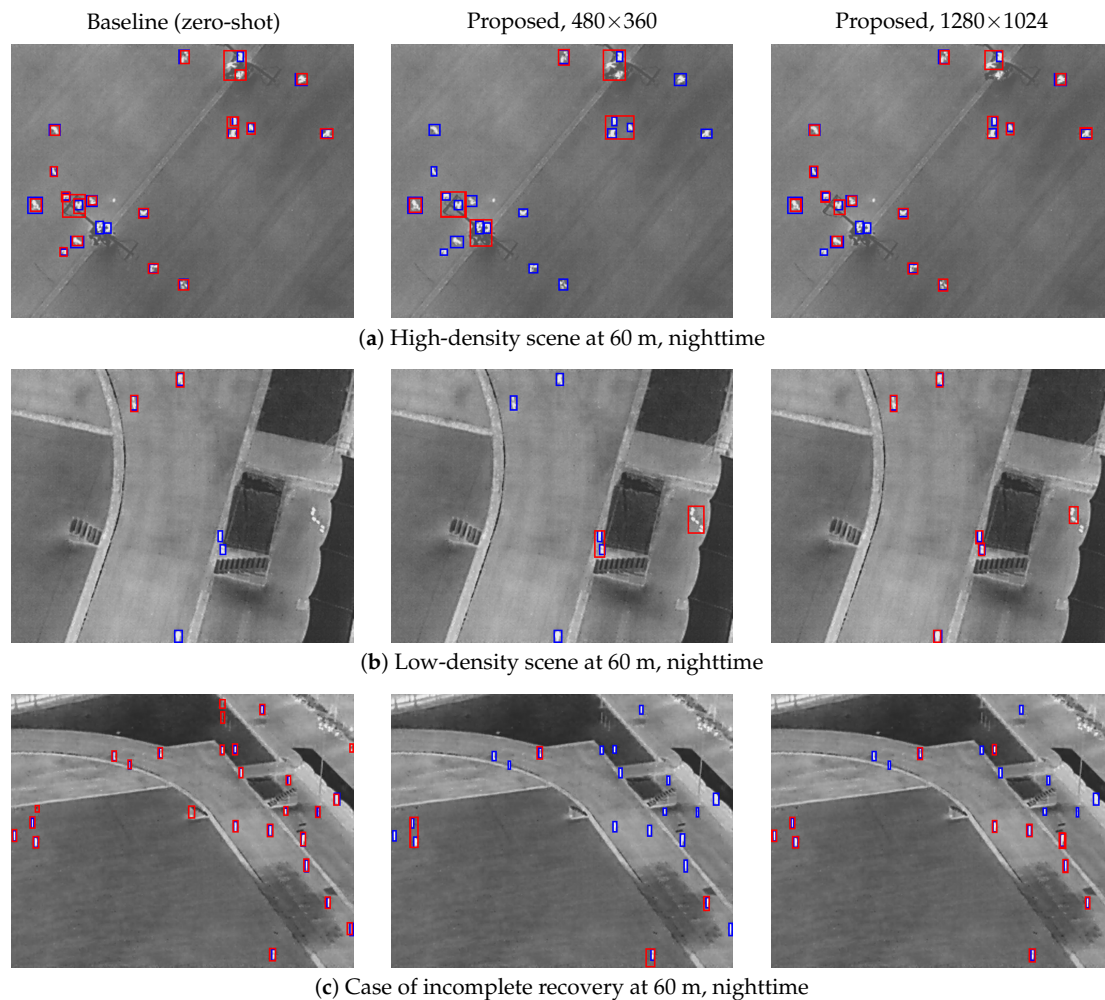


Figure 10. Qualitative detection results on the HIT-UAV dataset at 60 m under nighttime conditions. From left to right, each row shows the zero-shot baseline and the proposed model at input sizes of 480×360 px and 1280×1024 px. Blue boxes denote ground-truth annotations and red boxes denote model predictions. The proposed model misses most persons at the default input but recovers them after input rescaling, except in (c), where the recovery remains incomplete.

Table 8. External evaluation on the HIT-UAV test split (mAP@0.5, input size 480×360 px). Neither model was trained on this dataset.

Subset	Images	Baseline	Proposed
Full test split	579	0.281	0.021
Altitude 60 m	170	0.567	0.067
Altitude 70–90 m	135	0.116	0.012
Altitude 100–130 m	274	0.024	0.002
Daytime	183	0.060	0.003
Nighttime	396	0.356	0.033
60 m, nighttime	137	0.635	0.093

Table 9. Effect of input scale on the HIT-UAV evaluation (mAP@0.5). The default input of 480×360 px corresponds to a $0.70 \times$ scale of the original 640×512 px images. The dash (–) indicates a configuration that was not evaluated.

Subset	Input Scale	Baseline	Proposed
Full test split	480×360	0.281	0.021
	960×720	0.228	0.132
	1280×1024	–	0.214
Altitude 60 m	480×360	0.567	0.067
	1280×1024	0.388	0.422
60 m, nighttime	480×360	0.635	0.093
	1280×1024	0.458	0.474

4. Discussion

The detection results indicate that training on thermal infrared imagery captured by fixed-point ground cameras and processed through spatial resizing and padding improved person detection performance for UAV-based thermal infrared imagery. In the sports ground environment where grass and soil were mixed, the proposed model achieved a markedly higher Recall than the baseline. This improvement can be attributed to the baseline model failing to learn person-specific features under low thermal contrast. The proposed model captured spatial features related to person scale and contextual features derived from relative intensity distributions, maintaining stable detection even when persons appeared as small thermal sources.

The ablation study (Section 3.4) revealed that padding and position-based augmentation are the dominant contributors to detection performance under challenging environments. Padding introduces black background regions that emulate the low-temperature surroundings observed from a UAV, which encourages the model to learn person-specific local features. Position-based augmentation exposes the network to varied spatial locations of the target, which reduces the dependency on fixed central placement. These contributions become pronounced in environments with occlusions and scale variation.

The decrease in Precision observed for the proposed model can be attributed to over-segmentation. The model relied more heavily on the distribution of localized high- and low-intensity thermal regions than on the global human body shape, treating spatially discontinuous heat sources caused by occlusions or pose changes as independent detection targets. Post-processing strategies such as modified non-maximum suppression with a higher IoU threshold or spatial clustering of detected regions could mitigate this behavior, and remain a direction for future investigation.

The high sensitivity to localized thermal information can function effectively under specific conditions, as observed in the sports ground environment with debris. Whereas the baseline model frequently detected persons and occluding objects as a single entity, the proposed model identified person regions by recognizing thermal characteristics specific to the head and limbs even when heat sources were spatially discontinuous. This behavior suggests that the proposed model relies on contextual information incorporating relationships with surrounding regions rather than absolute intensity differences alone.

The detection failure of YOLOv8n in the sports ground and debris-covered environments (Section 3.3) further illustrates the importance of the Transformer-based architecture. The convolutional feature extractor of YOLOv8n encodes features through local kernels with limited receptive fields, which constrains its ability to model long-range dependencies across complex thermal scenes. The Transformer-based encoder of Grounding DINO captures global contextual relationships through self-attention, and the language-grounded text encoder provides additional semantic priors that stabilize detection under low thermal

contrast. The failure of RT-DETR-l under identical conditions (Section 3.3) reinforces this interpretation. The Transformer architecture alone is not sufficient, and the robustness of Grounding DINO derives primarily from the semantic priors acquired through large-scale image–text pre-training.

The proposed approach also generalized to nighttime acquisition (Section 3.5). The proposed model achieved a mAP@0.5 of 0.32 on the nighttime test set, exceeding its performance on the daytime debris-covered environment. This result indicates that the dominant failure mode in challenging scenes is physical occlusion rather than the time of day, because nighttime acquisition can enhance thermal contrast between cooled background surfaces and warm body parts.

The external evaluation on HIT-UAV (Section 3.7) revealed a specialization–generalization trade-off introduced by fine-tuning. The proposed model is specialized to the apparent person scale of its intended operational altitude and degrades on imagery acquired at three to six times that altitude, whereas the zero-shot baseline retains broader generality. The scale-matching analysis demonstrated that this degradation is governed primarily by the apparent person scale and is largely recoverable through altitude-aware input rescaling, which is practical because the flight altitude is known during UAV operations. This finding reframes the proposed preprocessing as an explicit scale-prior injection and suggests that maintaining altitude-specific input scaling, or altitude-stratified training data, is the principal requirement for extending the proposed approach to other operational altitudes.

The inference speed measurements (Section 3.6) indicate that Grounding DINO supports real-time processing on desktop-class GPUs at approximately 24 FPS, whereas YOLOv8n offers approximately 19 times higher throughput. Deployment on resource-constrained edge devices may require model compression techniques such as knowledge distillation or quantization, depending on whether detection performance or inference speed is prioritized.

In environments characterized by both thermal crossover and substantial background noise, detection remained difficult even for the proposed model. Further improvements are likely to require additional information sources such as temporal information from video data or complementary sensor modalities. The evaluation was also conducted using a single UAV platform and thermal camera model, with the nighttime evaluation limited to a single sports ground location. Generalization to other UAV altitudes, sensor resolutions, postures, and broader nighttime environments remains for future work. The present validation therefore remains limited in scale. Further testing with larger datasets, additional disaster-like environments, different UAV platforms, and different flight altitudes is required before operational deployment. The present system is designed for a flight altitude of approximately 20 m, assuming residential areas in Japan where utility poles of 10–15 m constrain the practical flight altitude. Applications in areas with high-rise buildings, which require higher flight altitudes, remain for future work.

5. Conclusions

This study developed a method for automatically detecting persons from UAV-acquired thermal infrared imagery, in support of victim search under nighttime disaster conditions. Thermal infrared imagery captured by fixed-point ground cameras was used to construct the training data. Spatial resizing and padding were applied to emulate the scale of persons observed from UAV viewpoints, and the resulting data were used to train the person detection model.

Quantitative evaluation showed that the proposed model achieved a substantial improvement in Recall compared with the baseline, particularly in environments where grass and soil were mixed and the thermal contrast was reduced. This improvement can

be attributed to the proposed model learning spatial features related to person scale and contextual features based on relative intensity distributions, even when persons appeared as small heat sources.

The novelty of this study lies in a training data construction strategy that addresses the limited availability of UAV-collected thermal infrared aerial imagery. Specifically, this study uses thermal infrared images collected by fixed-point ground cameras and adapts them to a UAV viewpoint for training. This strategy enhances the practical applicability of UAV-based thermal infrared remote sensing for nighttime disaster response.

The comprehensive evaluation included an ablation study, a comparison with YOLOv8n, and an extended evaluation under nighttime acquisition. The results indicate that the proposed preprocessing scheme generalizes across diverse environments and that Grounding DINO enables reliable detection in scenes with complex thermal noise. The inference speed measurements further indicate that real-time processing is feasible on desktop-class GPUs. The comparison with the DETR-based RT-DETR-l showed that closed-set detectors fail in complex thermal environments regardless of architecture, which attributes the robustness of the proposed system to the language-grounded semantic priors of Grounding DINO. The external evaluation on the HIT-UAV dataset quantified the generalization boundary of the fine-tuned model and demonstrated that altitude-aware input rescaling substantially mitigates the scale-induced degradation without retraining.

The proposed model exhibited a tendency toward over-segmentation due to its high sensitivity to localized thermal information, which resulted in a decrease in Precision. Detection also remained challenging under conditions involving thermal crossover and substantial background noise. Further improvements are expected to require the incorporation of temporal information from video data or other complementary sources. Future work will extend the evaluation to other UAV platforms, sensor resolutions, and broader nighttime environments, including higher flight altitudes required in areas with high-rise buildings. It will also investigate post-processing strategies to mitigate over-segmentation.

Author Contributions: Conceptualization, H.Y., S.K. and K.Y.; methodology, S.K. and K.Y.; software, S.K. and K.Y.; validation, K.Y.; formal analysis, K.Y.; investigation, S.K.; resources, K.Y.; data curation, S.K. and K.Y.; writing—original draft preparation, S.K.; writing—review and editing, S.K. and H.Y.; visualization, S.K. and K.Y.; supervision, S.K. and H.Y.; project administration, S.K. and H.Y.; funding acquisition, S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by JSPS KAKENHI Grant Number JP24K17486.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Ethics Committee of Kagawa University (protocol code 07-001, 29 July 2025).

Informed Consent Statement: Informed consent was obtained from all subjects whose identifiable images are presented in this study. For the data collection involving a large number of participants on campus, informed consent was obtained through prior public notification via the university website, signage at the site, and bulk email, providing individuals the option to opt-out or request data deletion.

Data Availability Statement: The data presented in this study are not publicly available due to privacy and ethical restrictions, as the datasets contain thermal infrared imagery of human subjects acquired in public spaces and private laboratory environments where individual privacy must be maintained.

Acknowledgments: The authors would like to thank the students and faculty members of Kagawa University who participated as subjects for the thermal infrared data collection.

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AdamW	Adaptive Moment Estimation with decoupled weight decay
BERT	Bidirectional Encoder Representations from Transformers
CNN	Convolutional Neural Network
FDMA	Fire and Disaster Management Agency
GIoU	Generalized Intersection over Union
GSD	Ground Sampling Distance
GSI	Geospatial Information Authority of Japan
IoU	Intersection over Union
INSARAG	International Search and Rescue Advisory Group
LWIR	Long-Wave Infrared
OCHA	United Nations Office for the Coordination of Humanitarian Affairs
PoT	Pixels on Target
SAR	Synthetic Aperture Radar
SEM	Satellite-based Emergency Mapping
UAV	Unmanned Aerial Vehicle

References

1. Koresawa, A. Evidence-Based Analysis of Search and Rescue Operations Following the Great East Japan Earthquake. *J. Disaster Res.* **2013**, *8*, 746–755. (In Japanese) [CrossRef]
2. Matsuyama, M.; Kimura, R.; Hayashi, H. Organizational Structure and Institutions for Disaster Prevention: Research on the 1995 Great Hanshin-Awaji Earthquake in Kobe City. *J. Disaster Res.* **2015**, *10*, 1051–1066. [CrossRef]
3. INSARAG (International Search and Rescue Advisory Group). *INSARAG Guidelines 2020: Volume II—Preparedness and Response, Manual B: Operations*; United Nations Office for the Coordination of Humanitarian Affairs (OCHA): Geneva, Switzerland, 2020. Available online: <https://insarag.org/methodology/insarag-guidelines/> (accessed on 1 May 2026).
4. Fire and Disaster Management Agency (FDMA). *Safety Management Manual for Firefighting and Emergency Operations*; Fire and Disaster Management Agency: Tokyo, Japan, 2016. Available online: https://www.fdma.go.jp/singi_kento/kento/items/kento002_01_manual00.pdf (accessed on 1 May 2026). (In Japanese)
5. Araiba, K.; Doshida, S. Technical Support for Search and Rescue Operations at a Landslide Disaster Site. *Fire Res. Inst. Rep.* **2020**, *70*, 22–27. (In Japanese)
6. National Police Agency of Japan. *Survey and Analysis of Police Rescue Activities During the 2016 Kumamoto Earthquake*; National Police Agency: Tokyo, Japan, 2017. Available online: <https://www.npa.go.jp/bureau/security/kumamotojishin/3-3-1.pdf> (accessed on 1 May 2026). (In Japanese)
7. Ministry of Land, Infrastructure, Transport and Tourism. *Report on Damage Situation Regarding the 2018 Hokkaido Eastern Iburi Earthquake (Report No. 24)*; MLIT: Tokyo, Japan, 2018. Available online: <https://www.mlit.go.jp/common/001255085.pdf> (accessed on 1 May 2026). (In Japanese)
8. Cabinet Office, Government of Japan. *Report on Verification of Response to the July 2018 Heavy Rain Disaster (Heisei 30-nen 7-Gatsu Gouu Saigai Taiou Kenshou Houkokusho)*; Cabinet Office: Tokyo, Japan, 2018. Available online: https://www.bousai.go.jp/fusuigai/suigai_dosyaworking/pdf/kurashikikensyou.pdf (accessed on 1 May 2026). (In Japanese)
9. Cabinet Office, Government of Japan. *Report on the Verification of Initial Response to the 2016 Kumamoto Earthquake*; Cabinet Office: Tokyo, Japan, 2016. Available online: <https://www.bousai.go.jp/updates/h280414jishin/h28kumamoto/index.html> (accessed on 1 May 2026).
10. Cabinet Office, Government of Japan. *Regarding the Response to the 2024 Noto Peninsula Earthquake (Verification Team Material)*; Cabinet Office: Tokyo, Japan, 2024. Available online: https://www.bousai.go.jp/updates/r60101notojishin/kensho_team.html (accessed on 1 May 2026). (In Japanese)
11. Ishikawa Prefecture. *2024 Noto Peninsula Earthquake Countermeasures Verification Report*; Ishikawa Prefecture: Kanazawa, Japan, 2024. Available online: https://www.pref.ishikawa.lg.jp/bousai/bousai_g/bousaikaigi/20241028/kensho-houkokusho.html (accessed on 1 May 2026). (In Japanese)

12. Voigt, S.; Giulio-Tonolo, F.; Lyons, J.; Kučera, J.; Jones, B.; Schneiderhan, T.; Platzeck, G.; Kaku, K.; Hazarika, M.K.; Czarán, L.; et al. Global trends in satellite-based emergency mapping. *Science* **2016**, *353*, 247–252. [CrossRef] [PubMed]
13. Twele, A.; Kandarr, J.; Brinkmann, S.; Martinis, S. Sentinel-1-based flood mapping: A fully automated processing chain. *Int. J. Remote Sens.* **2016**, *37*, 2990–3004. [CrossRef]
14. Wang, D.; Ma, G.; Wang, X.; Yang, R.; Zhang, Y. Few-Shot Change Detection in Optical and SAR Remote Sensing Images for Disaster Response. *Int. J. Appl. Earth Obs. Geoinf.* **2026**, *146*, 105100. [CrossRef]
15. Arai, K. Modified pix2pixHD for Enhancing Spatial Resolution of Image for Conversion from SAR Images to Optical Images in Application of Landslide Area Detection. *Information* **2025**, *16*, 163. [CrossRef]
16. Restas, A. Drone Applications for Supporting Disaster Management. *World J. Eng. Technol.* **2015**, *3*, 316–321. [CrossRef]
17. Turner, D.; Lucieer, A.; de Jong, S.M. Time Series Analysis of Landslide Dynamics using an Unmanned Aerial Vehicle (UAV). *Remote Sens.* **2015**, *7*, 1736–1757. [CrossRef]
18. VanExel, K.; Sherchan, S.; Liu, S. Optimizing Deep Learning Models for Climate-Related Natural Disaster Detection from UAV Images and Remote Sensing Data. *J. Imaging* **2025**, *11*, 32. [CrossRef] [PubMed]
19. Shukla, P.; Shukla, S. Unmanned Aerial Vehicle (UAV) Based Disaster Detection and Crowd Sensing Using Deep Learning Models. In *Proceedings of the 26th International Conference on Distributed Computing and Networking (ICDCN 2025), Hyderabad, India, 4–7 January 2025*; ACM: New York, NY, USA, 2025; pp. 414–419. [CrossRef]
20. Wang, Z.; Gao, Q.; Xu, J.; Li, D. A Review of UAV Power Line Inspection. In *Advances in Guidance, Navigation and Control*; Yan, L., Duan, H., Yu, X., Eds.; Lecture Notes in Electrical Engineering; Springer: Singapore, 2022; Volume 644, pp. 3147–3159. [CrossRef]
21. Bouguettaya, A.; Zarzour, H.; Kechida, A.; Taberkit, A.M. Deep learning techniques to classify agricultural crops through UAV imagery: A review. *Neural Comput. Appl.* **2022**, *34*, 9511–9536. [CrossRef] [PubMed]
22. Mohammadi, M.E.; Watson, D.P.; Wood, R.L. Deep Learning-Based Damage Detection from Aerial SfM Point Clouds. *Drones* **2019**, *3*, 68. [CrossRef]
23. Lygouras, E.; Santavas, N.; Taitzoglou, A.; Tarchanidis, K.; Mitianoudis, N.; Gasteratos, A. Unsupervised Human Detection with an Embedded Vision System on a Fully Autonomous UAV for Search and Rescue Operations. *Sensors* **2019**, *19*, 3542. [CrossRef] [PubMed]
24. Sambolek, S.; Ivasic-Kos, M. Automatic Person Detection in Search and Rescue Operations Using Deep CNN Detectors. *IEEE Access* **2021**, *9*, 37905–37922. [CrossRef]
25. Bejiga, M.B.; Zeggada, A.; Nouffidj, A.; Melgani, F. A Convolutional Neural Network Approach for Assisting Avalanche Search and Rescue Operations with UAV Imagery. *Remote Sens.* **2017**, *9*, 100. [CrossRef]
26. Qu, C.; Hao, X.; Ma, F.; Hui, J.; Guo, W.; Wang, H. Enhancing Human Detection in Post-Disaster Scenarios Through Generative Adversarial Networks. In *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management*; Duffy, V.G., Ed.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2025; Volume 15791, pp. 257–272. [CrossRef]
27. Munir, A.; Siddiqui, A.J.; Hossain, M.S.; El-Maleh, A. YOLO-RAW: Advancing UAV Detection with Robustness to Adverse Weather Conditions. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 7857–7873. [CrossRef]
28. Li, Y.; Bao, G.; Hu, J.; Zhi, X.; Hu, T.; Wang, J.; Wu, W. A Ship Incremental Recognition Framework via Unknown Extraction and Joint Optimization Learning. *Remote Sens.* **2026**, *18*, 149. [CrossRef]
29. Geospatial Information Authority of Japan. GSI Maps. Available online: <https://maps.gsi.go.jp/> (accessed on 1 May 2026).
30. Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. In *Computer Vision—ECCV 2024*; Springer: Cham, Switzerland, 2024; pp. 38–55. [CrossRef]
31. Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLOv8, Version 8.4.41. Software. 2023. Available online: <https://github.com/ultralytics/ultralytics> (accessed on 1 May 2026).
32. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99.
33. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In *European Conference on Computer Vision (ECCV)*; Springer: Cham, Switzerland, 2020; pp. 213–229. [CrossRef]
34. Ren, T.; Jiang, Q.; Liu, S.; Zeng, Z.; Liu, W.; Gao, H.; Huang, H.; Ma, Z.; Jiang, X.; Chen, Y.; et al. Grounding DINO 1.5: Advance the “Edge” of Open-Set Object Detection. *arXiv* **2024**, arXiv:2405.10300. [CrossRef]
35. CVAT.ai Corporation. Computer Vision Annotation Tool (CVAT). Available online: <https://www.cvat.ai/> (accessed on 1 May 2026).
36. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In *Computer Vision—ECCV 2014*; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2014; Volume 8693, pp. 740–755. [CrossRef]
37. Zhao, H.; Ji, Z.; Li, N.; Gu, J.; Li, Y. Target Detection over the Diurnal Cycle Using a Multispectral Infrared Sensor. *Sensors* **2017**, *17*, 56. [CrossRef] [PubMed]

38. Singh, R.; Puhl, R.B.; Dhakal, K.; Sornapudi, S. Few-Shot Adaptation of Grounding DINO for Agricultural Domain. *arXiv* **2025**, arXiv:2504.07252. [[CrossRef](#)]
39. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs Beat YOLOs on Real-Time Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2024*; IEEE: Piscataway, NJ, USA, 2024; pp. 16965–16974. [[CrossRef](#)]
40. Suo, J.; Wang, T.; Zhang, X.; Chen, H.; Zhou, W.; Shi, W. HIT-UAV: A High-Altitude Infrared Thermal Dataset for Unmanned Aerial Vehicle-Based Object Detection. *Sci. Data* **2023**, *10*, 227. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.