University of Reading

PETS2025: multi-authority multi-sensor maritime surveillance challenge and evaluation

Conference or Workshop Item

Accepted Version

Markchom, T. ORCID: https://orcid.org/0000-0002-2685-0738, Boyle, J. ORCID: https://orcid.org/0000-0002-5785-8046, Chen, L., Ferryman, J., Marturini, M., Veigl, S., Opitz, A., Kriechbaum-Zabini, A., Bratskas, R., Gkamaris, A., Papachristos, D., Leventakis, G., Fan, W., Huang, H.-W., Hwang, J.-N., Kim, P., Kim, K., Huang, C.-I., Saito, K., Kaneko, S., Sudo, K., Thanh Thien, N., Kao, M.-Y., Hsieh, J.-W., Lilek, T., Pomsuwan, T., Gu, J., Xu, T., Zhu, X., Wu, X., Kittler, J., Stacy, S., Gabaldon, A., Tu, P., Kim, S., Kim, D. and Lee, K. (2025) PETS2025: multi-authority multi-sensor maritime surveillance challenge and evaluation. In: IEEE International Conference on Advanced Visual and Signal-Based Systems (AVSS 2025), 11-13 August 2025, Tainan, Taiwan. doi: 10.1109/AVSS65446.2025.11149786 Available at https://centaur.reading.ac.uk/124177/

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.

To link to this article DOI: http://dx.doi.org/10.1109/AVSS65446.2025.11149786

University of Reading

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the End User Agreement.

www.reading.ac.uk/centaur

CentAUR

Central Archive at the University of Reading

Reading's research outputs online

PETS2025: Multi-Authority Multi-Sensor Maritime Surveillance Challenge and Evaluation

Thanet Markchom[1*], Jonathan Boyle[1], Lulu Chen[1], James Ferryman[1], Matteo Marturini[2], Stephan Veigl[2], Andreas Opitz[2], Andreas Kriechbaum-Zabini[2], Romaios Bratskas[3], Anastasios Gkamaris[3], Dimitris Papachristos[3], George Leventakis[4], Wenjun Fan[5], Hsiang-Wei Huang[5], Jeng-Neng Hwang[5], Pyongkun Kim[6], Kwangju Kim[6], Chung-I Huang[7], Kenta Saito[8], Shunta Kaneko[8], Kyoko Sudo[8], Nguyen Thanh Thien[9], Meng-Yu Kao[10], Jun-Wei Hsieh[10], Teepakorn Lilek[11], Tossapol Pomsuwan[12], Jinjie Gu[13], Tianyang Xu[13], Xuefeng Zhu[13], Xiaojun Wu[13], Josef Kittler[14], Stephanie Stacy[15], Alfredo Gabaldon[15], Peter Tu[15], Sangwon Kim[6], Dongyoung Kim[6], Kyoungoh Lee[6],

[1]University of Reading, UK [2]Austrian Institute of Technology, Austria [3]SKYLD Security And Defence LTD, Cyprus

[4]University of the Aegean, Greece [5]University of Washington, USA

[6]Electronics and Telecommunications Research Institute, Republic of Korea

[7]National Center of High-performance Computing, Taiwan [8]Toho University, Japan

[9]University of Information Technology, Vietnam [10]National Yang Ming Chiao Tung University, Taiwan

[11]National Electronics and Computer Technology Center, Thailand

[12]Digital Government Development Agency, Thailand [13]Jiangnan University, China

[14]University of Surrey, UK [15]GE Aerospace Research, USA

[*]thanet.markchom@reading.ac.uk

Abstract

This paper presents the outcomes of the PETS2025 challenge, held in conjunction with AVSS 2025 and sponsored by the EU-funded EURMARS project. The challenge introduces a novel maritime surveillance dataset comprising image sequences captured by diverse multi-altitude, multimodal sensors, reflecting the real-world multi-authority environment. The key tasks include: (1) object detection using various sensors across different platforms (ground-based and low-altitude aerial) and spectral ranges (visible, thermal, ultraviolet (UV), and short-wave infrared (SWIR)); (2) long-term tracking of targets in maritime environments spanning both sea and land; and (3) approximating target geolocations by using sensor imagery and telemetry data. Performance evaluations of results submitted by 12 international participants are discussed. The results show the effectiveness of these submissions and highlight ongoing challenges posed by heterogeneous sensors and complex environments. These challenges emphasise the need to further improve detection, tracking, and geolocation approximation for maritime and coastal surveillance.

1. Introduction

There has been a notable rise in irregular migration in recent years, accompanied by escalating instances of human trafficking and smuggling [5]. Other illicit activities, such as drug and arms trafficking; and illegal, unreported, and unregulated (IUU) fishing, have also intensified. These evolving challenges call for more coordinated efforts and enhanced surveillance measures among different authorities, particularly through the deployment of various advanced sensing technologies to support wide-area surveillance across maritime and land domains.

Despite advancements in maritime surveillance research, significant challenges remain in developing effective multi-authority, multi-sensor surveillance platforms. A primary challenge arises from the nature of maritime and coastal environments, which span vast and dynamic areas of both land and sea. These settings pose substantial difficulties for surveillance systems due to factors such as fluctuating sea states, variable weather conditions, occlusions from terrain or vessels, and the complexity of tracking multiple target types across heterogeneous landscapes. An additional challenge lies in the multi-authority dimension of the task, which requires coordination among different stakeholders

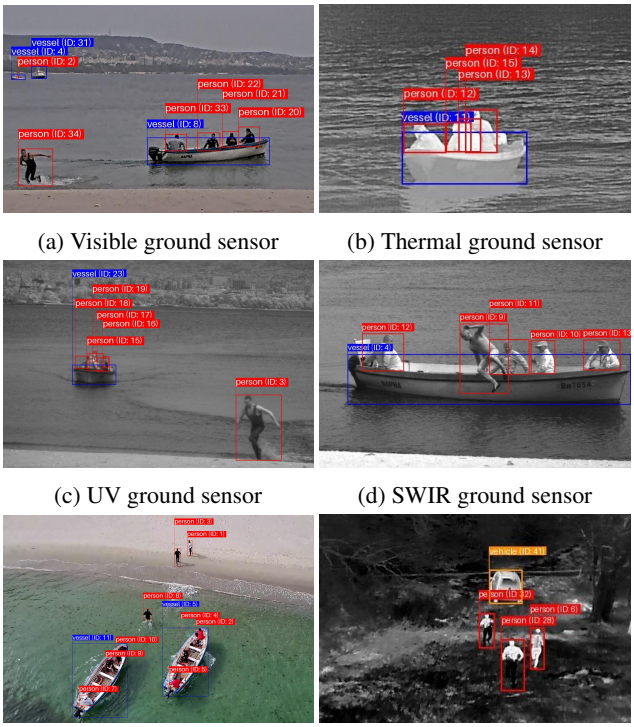
operating a wide range of sensor technologies. In particular, this includes the integration of sensors deployed at various operational levels (e.g., ground-based systems and UAVs) and across diverse spectral ranges (e.g., visible, thermal, ultraviolet (UV), and short-wave infrared (SWIR)). Ensuring interoperability and consistent data integration across various sensor platforms and modalities is inherently complex. Collectively, these challenges underscore the difficulty of designing a robust solution capable of seamlessly integrating a broad array of sensors under a unified surveillance system.

The PETS2025 Challenge is introduced to foster innovation in the development of surveillance systems that integrate multi-authority, multi-sensor capabilities. The challenge focuses on the analytical tasks of detecting and tracking humans, vessels, and vehicles and approximating their geolocations in real-world coordination within multi-authority maritime border surveillance scenarios. These tasks have received relatively little attention in the computer vision community due to the lack of suitable datasets. This initiative aims to raise awareness of these pressing challenges and promote the development of innovative solutions. It will provide border authorities worldwide with advanced tools to enhance cooperation and improve security in multi-authority coastal and maritime border regions.

2. Challenges

Challenge 1: Target detection and classification in maritime and coastal areas using multi-platform and multi-spectral sensors This challenge focuses on detecting and classifying key objects, namely persons (including individuals on deck, on the shore, and a floating mannequin or dummy in the water), vessels, and vehicles, in image sequences. The data consists of image sequences captured from multiple sensor types, simulating a multi-authority maritime surveillance scenario. These sensors include both ground-based cameras and UAVs. Ground-based cameras comprise visible (RGB), thermal, UV, and SWIR modalities, while UAVs provide visible (RGB) and thermal imagery. Figure 1 shows examples of images from different sensors with examples of three object classes: person, vessel, and vehicle.

Challenge 2: Long-term (LT) target tracking across diverse terrains using multi-platform and multispectral sensors This challenge focuses on long-term target tracking, where the objective is to continuously track persons, vessels, and vehicles across image sequences. Each object must first be detected and classified as a person, vessel, or vehicle, and then assigned a unique identifier (track ID) that remains consistent throughout the sequence. If an object disappears temporarily (due to occlusion or movement), it should be reassigned the same ID when it reappears. Examples of image sequence segments from Challenge 2 are illustrated in Figures 2 and 3.



(a) Visible ground sensor (b) Thermal ground sensor

(c) UV ground sensor (d) SWIR ground sensor

(e) Visible UAV (f) Thermal UAV

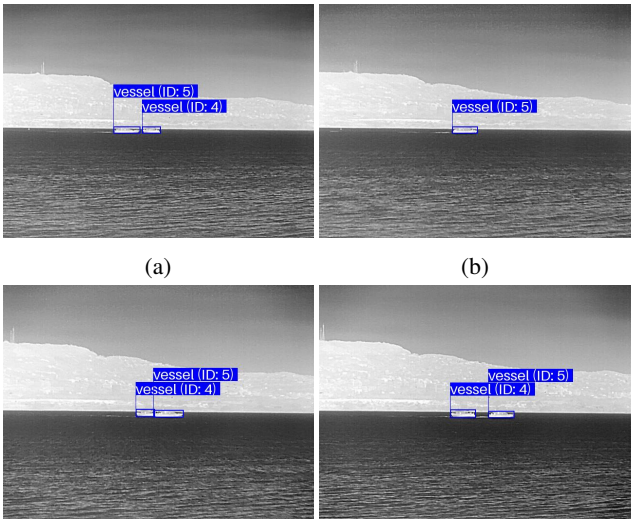
Figure 1: Challenge 1 examples: Images from different sensors, including visible, thermal, UV, and SWIR ground sensors, as well as visible and thermal UAVs, with examples of three object classes: person, vessel, and vehicle.

Challenge 3: Geolocating objects in images captured by a moving sensor This challenge focuses on estimating the geolocations, specifically the longitude and latitude coordinates, of specified objects, either persons or vessels, across sequences of thermal UAV images. Each image sequence is accompanied by ground-truth bounding boxes identifying the objects whose geolocations need to be approximated. For each image, telemetry data of the UAV recorded at the time of capture are provided. These include key parameters useful for the task, such as focal length, digital zoom ratio, latitude, longitude, relative altitude, absolute altitude, gimbal yaw, gimbal pitch, gimbal roll, and the corresponding timestamp.

2.1. Datasets

The dataset provided for this challenge is derived from the EU project EURMARS[1]. It simulates multi-authority

[1] https://eurmars-project.eu/



(a) (b)

(c) (d)

Figure 2: Challenge 2 examples: A thermal ground sensor image sequence (starting from (a) to (d)) showing the tracking of vessels, where "vessel (ID: 4)" is occluded by "vessel (ID: 5)" in (b) and reappears in (c) and (d).



(a) (b)

(c) (d)

Figure 3: Challenge 2 examples: A visible ground sensor image sequence (starting from (a) to (d)) showing the tracking of vessels and persons across different terrains, particularly "person (ID: 34)" from the boat to the shore.

maritime and coastal border scenarios such as vessels navigating coastal waters, individuals on board approaching shorelines, and people disembarking vessels and heading toward vehicles on adjacent land areas. To reflect multi-authority situations, this dataset includes image sequences obtained from various types of sensors at different levels and across the spectral ranges mentioned above.

For Challenges 1 and 2, the dataset consists of 11 training scenarios and 4 test scenarios. Each scenario comprises



Figure 4: Challenge 3 example images showing a detected vessel from a UAV (left) and its ground-truth GNSS data over time, projected onto a map (right).

a varying number of sequences from different sensor types. The training and test sequences are summarised in Tables 1 and 2, respectively. The training set comprises 22,086 images, including 112,755 person annotations, 44,540 vessel annotations, and 903 vehicle annotations, with a total of 678 unique track IDs across all object classes. The test set consists of 14,782 images, containing 77,101 person annotations, 42,971 vessel annotations, and 558 vehicle annotations, along with 311 unique track IDs.

For Challenge 3, the provided dataset contains only thermal UAV image sequences. There are 8 sequences for training and 5 sequences for testing. Sequences "rd1"–"rd7" are from controlled scenarios designed with known conditions for trials and calibration, where the UAV deployment position and the object of interest were at the same altitude, both above sea level. Sequence "bg7" represents a real-world scenario in which the UAV deployment position and the object of interest were also at the same altitude, both at sea level. Sequences "cy1", "cy2", and "cy4"–"cy6" are real-world scenarios where the UAV deployment position and the object of interest were at different altitudes: the UAV was above sea level, while the object was at sea level. Alongside the sequences: telemetry data, including focal length, digital zoom ratio, latitude, longitude, relative altitude, absolute altitude, gimbal yaw, gimbal pitch, gimbal roll, and timestamps corresponding to each image in the image sequence, are also provided. Ground-truth detection (bounding boxes) of objects for which the participant is required to approximate their geolocations is provided.

For all challenges, RGB ground sensor images are Full HD (1920x1080), SWIR images are 1280x1024, UV images are 1416x1420, visible UAV images are 4K (3840x2160), and thermal UAV images are 640x480, all in JPEG format.

3. Challenge Submission

3.1. Requirements

For Challenge 1, participants should provide bounding boxes and class labels for detected objects in the specified format. For Challenge 2, submissions should include

Table 1: Training data for Challenges 1 and 2

Scenario	Sensor	#images	#persons	#vessels	#vehicles	#tracks
bg1	GS-RGB	378	3141	769	76	24
	GS-SWIR	385	2288	822	-	15
	GS-Therm	386	2675	778	-	25
	GS-UV	386	2479	897	-	15
bg3	GS-RGB	728	6791	1523	-	16
	GS-SWIR	800	6693	1846	-	14
	GS-Therm	742	6183	1485	-	19
	GS-UV	702	5117	1657	-	16
bg4	GS-RGB	589	4148	1367	162	42
	GS-SWIR	754	6591	2091	13	32
	GS-Therm	635	4145	1471	-	26
	GS-UV	746	6669	2204	17	30
bg5	GS-RGB	486	1792	676	-	17
	GS-SWIR	583	1751	1165	-	9
	GS-Therm	499	450	861	-	12
	GS-UV	530	879	995	-	6
bg7	GS-RGB	756	2758	1009	69	24
	GS-Therm	810	1165	1869	-	13
	UAV-Therm	808	8801	1395	408	44
bg9	GS-RGB	357	1060	611	14	14
	GS-SWIR	386	581	1094	-	7
	GS-Therm	391	1037	685	-	8
	GS-UV	387	600	1250	-	7
bg10	GS-RGB	459	1798	1051	144	17
	GS-SWIR	449	1297	1559	-	9
	GS-Therm	454	1192	972	-	6
	GS-UV	387	986	1340	-	11
bg11	GS-RGB	391	5203	1255	-	30
	GS-SWIR	409	4085	1325	-	30
	GS-Therm	406	4217	1102	-	33
	GS-UV	410	3771	1356	-	30
bg12	GS-RGB	263	776	958	-	9
	GS-SWIR	300	753	306	-	7
	GS-Therm	297	652	36	-	5
	GS-UV	301	790	416	-	7
cy1	UAV-RGB	970	2115	399	-	10
	UAV-Therm	969	1795	906	-	5
cy2	UAV-RGB	1204	1569	888	-	12
	UAV-Therm	1193	3962	2151	-	22
Total		22086	112755	44540	903	678

Table 2: Test data for Challenges 1 and 2

Scenario	Sensor	#images	#persons	#vessels	#vehicles	#tracks
bg2	GS-RGB	871	8233	2506	-	26
	GS-SWIR	846	6069	2197	-	14
	GS-Therm	845	5679	1723	-	20
	GS-UV	845	5131	2318	-	18
	UAV-Therm	842	6780	1947	-	17
bg6	GS-RGB	628	3536	1485	133	28
	GS-SWIR	569	3531	2116	-	14
	GS-Therm	565	3851	1823	-	15
	GS-UV	571	3818	2038	-	16
bg8	GS-RGB	1526	7434	5345	-	24
	GS-SWIR	1515	5093	4609	-	25
	GS-Therm	1515	6337	4046	-	22
	UAV-Therm	1513	9455	7426	425	47
cy3	UAV-RGB	1072	440	1552	-	16
	UAV-Therm	1059	1714	1840	-	9
Total		14782	77101	42971	558	311

Table 3: Training set for Challenge 3

Scenario	#images	#GNSS data points
cy1	969	906
cy2	1193	1134
rd1	514	415
rd2	588	527
rd3	1388	1388
rd4	322	322
rd5	395	337
rd6	589	529
Total	5958	5558

Table 4: Test set for Challenge 3

Scenario	#images	#GNSS data points
bg7	808	697
cy4	351	343
cy5	225	225
cy6	201	201
rd7	561	464
Total	2146	1930

the track ID, along with the bounding box and class label for each detected object, following the specified format. For Challenge 3, the geolocation coordinates of specified objects in each image should be submitted as (longitude, latitude) pairs, using the provided format. Challenge 1 is mandatory, while Challenges 2 and 3 are optional. For each challenge, results should be submitted for all test sequences (if possible, for comprehensive evaluation) or at least two test sequences from different sensor types. For example, one sequence could be from the RGB ground sensor in scenario "bg2" and another from the thermal UAV in scenario "cy3". The submission to the challenge consists of the generated XML files (in the provided format) and a brief description (less than 500 words) of the methods used.

3.2. Submissions

In total, 12 teams participated in Challenge 1 (Table 5), and 4 teams participated in Challenge 2 (Table 6). For Challenge 3, there is only one submission from the PETS2025 organisers. Details on the approaches used by each team are provided in the following sections.

3.2.1 University of Reading, UK (UoR)

As the challenge organisers, UoR provided baseline results for all sequences. Several models were employed for object detection: YOLOv5 [9], YOLOv8 [10], YOLOv11 [11], and RT-DETR [23]. In addition, a YOLOv11 model was retrained using the thermal UAV training set to better handle thermal images. For Challenge 2, UoR also provided tracking baselines using three existing trackers: DeepSORT [21], ByteTrack [22], and BoT-SORT [1]. Each tracker used detections from RT-DETR as input for generating tracks.

For Challenge 3, geolocations were estimated using a method based on geometric projection and homography transformation. Specifically, the UAV's field of view (FOV) was projected onto the ground using its altitude, pitch, and FOV angles. Ground distances to the image centre, front, and back were computed, followed by slant ranges and horizontal extents. These were combined with the UAV's global position and yaw to derive the ground-projected image cor-

Table 5: Challenge 1 submissions (bracketed numbers indicate the count of separate submissions per sequence; absence implies a single submission.)

Participant	GS-RGB			GS-SWIR			GS-Therm			GS-UV		UAV-RGB	UAV-Therm		
	bg2	bg6	bg8	bg2	bg6	bg8	bg2	bg6	bg8	bg2	bg6	cy3	bg2	bg8	cy3
UoR	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
AIT				✓	✓	✓	✓	✓	✓						
SKYLD	✓	✓										✓			
UWIPL_ETRI				✓	✓	✓	✓	✓	✓			✓	✓	✓	✓
Toho U.	✓	✓	✓				✓	✓	✓			✓	✓	✓	✓
UIT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
NYCU	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
NECTEC	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
DGA	✓	✓	✓										✓	✓	✓
JU & U. Surrey	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
GE Aerospace		✓(2)				✓(2)									
ETRI-Vision	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 6: Challenge 2 submissions

Participant	GS-RGB			GS-SWIR			GS-Therm			GS-UV		UAV-RGB	UAV-Therm		
	bg2	bg6	bg8	bg2	bg6	bg8	bg2	bg6	bg8	bg2	bg6	cy3	bg2	bg8	cy3
UoR	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
AIT				✓	✓	✓	✓	✓	✓						
JU & U. Surrey	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
ETRI-Vision	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

ners, from which a homography matrix was calculated to map image pixels to geospatial coordinates. To address low-pitch scenarios where the image may contain sky, a horizon-aware correction was applied. If the pitch exceeded a certain threshold, the front ground distance was capped, and the horizon line in the image was estimated. A new homography was then computed using only the region below the horizon. Additionally, vertical pixel scaling was applied to mitigate distortion in oblique views: upper pixels were compressed and lower pixels were stretched, improving localisation accuracy near the horizon.

3.2.2 Austrian Institute of Technology, Austria (AIT)

AIT provided comprehensive results covering all GS-SWIR and GS-Therm sequences for both Challenge 1 and Challenge 2. In Challenge 1, their detection method utilised an enhanced YOLO-based algorithm, specifically adapted to the unique environmental conditions and task requirements. They developed two distinct detection pipelines for each sensor: one for land-based targets and another for maritime targets. The land-based detector focused on identifying persons and vehicles, while the maritime detector focused on detecting ships and smaller vessels.

For Challenge 2, AIT implemented a detect-and-track framework for object tracking. In this approach, tracked objects were initialised directly from the outputs of a detection model, as opposed to methods that require manual bounding box initialisation. Their tracking approach leveraged a YOLO-X backbone [7] for robust object detection, which provided the initial detections used to start tracks in the first frame. In subsequent frames, tracking was accomplished by associating new detections with existing tracks. This association was performed by computing a matrix of Euclidean distances between detected objects and active tracks. The optimal assignment between tracks and detections was then determined using the Hungarian algorithm, a well-established method for solving global assignment problems efficiently.

3.2.3 SKYLD Security And Defence LTD, Cyprus (SKYLD)

This submission includes results for both UAV-RGB and GS-RGB sequences in Challenge 1. SKYLD utilised the SSD-MobileNet v2 architecture. This approach combined the speed and accuracy of the Single Shot MultiBox Detector (SSD) with the efficiency of the lightweight MobileNet v2 backbone, enabling real-time inference on the NVIDIA Jetson Orin platform. The model was integrated through NVIDIA's real-time vision DNN library tailored for Jetson devices. Execution was optimised using TensorRT, providing high-performance GPU-accelerated inference accessible from both C++ and Python environments, achieving

more than 200 frames per second with very good accuracy. Model training was conducted in PyTorch on a UAV-RGB dataset. Although trained specifically for object detection on UAV-RGB images, the model was also applied to the GS-RGB test sequences to explore its performance in detecting objects from different perspectives (ground-based instead of aerial) within the same modality.

3.2.4 University of Washington, USA & Electronics and Telecommunications Research Institute, Republic of Korea & National Center of High-performance Computing, Taiwan (UWIPL_ETRI)

This submission is a collaboration between researchers from the University of Washington, the Electronics and Telecommunications Research Institute (ETRI), and the National Center for High-performance Computing as participants in Challenge 1. The submission consists of results across all sequences from the GS-RGB, GS-SWIR, GS-Therm, UAV-RGB, and UAV-Therm sensors. The proposed approach used a YOLOv11 object detector, trained on each modality, to perform object detection and classification.

3.2.5 Toho University, Japan (Toho U.)

The team from Toho University participated in Challenge 1, submitting results across all test scenarios for the GS-RGB, GS-Thermal, UAV-RGB, and UAV-Thermal sequences. They proposed a YOLOv11n-based system with an auxiliary detection module. YOLOv11n was pre-trained on the COCO dataset and then fine-tuned for each sensor using the PET2025 training dataset. The auxiliary detection module detected objects whose training data did not cover enough appearance variations, resulting in YOLOv11n false negatives. This module first semantically segmented the image using the Segment Anything Model 2 (SAM2), then classified each segment into the target classes or none according to the bounding-box (BBox) parameters.

Based on preliminary experiments using the PET2025 training dataset, they modelled the BBox parameters (width, height) for each sensor as a Gaussian Mixture Model (GMM). The number of components in the GMM was determined such that all classes were assigned to at least one cluster, and the increase in log-likelihood fell below a predefined threshold. They also applied class-specific weights using the ratio of the number of training samples for class imbalance correction. Each GMM cluster was associated with a class that had the maximum number of training samples and the highest likelihood. For multi-class classification using the GMM, they assigned each input to the class corresponding to the cluster with the highest likelihood. After eliminating BBoxes that did not belong to any object class, the result of the auxiliary detection module was fused with the result of the YOLOv11n module. They obtained the fused result by comparing the sets of BBoxes from each module, adopting the auxiliary detection result only when the IoU was below 0.5. Otherwise, they treated the detection as overlapping.

In their proposed auxiliary module, a class was assigned based solely on the similarity of the BBox size, regardless of the texture within the BBox detected by SAM2. This module was partially effective, especially on UAV RGB, since SAM2 could cover many objects that YOLOv11n missed. However, this resulted in incorrect classifications when different classes had the same BBox size. To solve these problems, they planned to introduce a classifier based on an element distribution model that also incorporated texture information in a future challenge.

3.2.6 University of Information Technology, Vietnam (UIT)

This submission includes results for all sequences of Challenge 1. The proposed method employs three YOLO models: YOLOv8m [12], YOLOv8m-SA [3], and YOLOv8m-ResCBAM [3]. For each sensor type, the training datasets were randomly split into two subsets with a ratio of 8:2; these subsets were used for training these models. All models were trained from scratch for 300 epochs with an image size of 640. After the training process, for each sensor type, the best model was selected based on the mAP50 metric and used later for inference on the test set. For the GS-RGB sensor, YOLOv8m-SA was selected, while YOLOv8m was used for the remaining sensors. To improve the detection results, the inference process used an image size of 960 instead of 640 in the training process.

3.2.7 National Yang Ming Chiao Tung University, Taiwan (NYCU)

The team from NYCU participated in Challenge 1 and submitted results for all sequences in the challenge. The proposed approach employed the Parallel Residual Bi-Fusion Feature Pyramid Network (PRB-FPN) detector, implemented as per `https://github.com/pingyang1117/PRBNet_PyTorch`. The model was trained on the provided dataset, with the last 10% of each sequence reserved as a validation set to select the best weights based on validation performance. Training data from all sensor types (Visible, Thermal, UV, SWIR for ground sensors; Visible, Thermal for UAV sensors) were combined to enhance robustness across diverse conditions. The PRB-FPN6-MSP architecture, initialised with MS-COCO pretrained weights, was trained for 100 epochs at an image resolution of 1280x1280 pixels. Training was performed on a single NVIDIA V100 GPU with a batch size of

8. The best-performing weights were used for inference on test sequences.

3.2.8 National Electronics and Computer Technology Center, Thailand (NECTEC)

NECTEC's submission includes results for all sequences across all sensors in Challenge 1. They fine-tuned the RT-DETR (Large) model [23] using only the provided dataset, starting from COCO-pretrained weights. Initially, they experimented with lightweight models, such as YOLOv8 and YOLOv12 [20], but these models struggled with detecting small objects in this dataset. Therefore, they switched to a transformer-based model, RT-DETR, which is known to perform well on small object detection tasks. Its attention mechanism helps capture fine details and global context much more effectively than traditional CNNs. For data augmentation, they applied both colour space and geometric techniques, such as HSV adjustments, rotation, translation, scaling, shearing, perspective transformation, and mosaic augmentation. These helped improve the model's robustness. The model was trained using the AdamW optimiser with a learning rate of 1e-3 for 60 epochs (including 5 warmup epochs) and a batch size of 16.

3.2.9 Digital Government Development Agency, Thailand (DGA)

The participant conducted an evaluation of two YOLOv8 models on the GS-RGB and UAV-Therm datasets as part of Challenge 1. The models tested were YOLOv8x, an extra-large version pretrained on the COCO dataset, and YOLOv8s, a smaller variant also pretrained on COCO but further fine-tuned on data from all sensor types. Despite the additional fine-tuning, the YOLOv8s model was consistently outperformed by the off-the-shelf YOLOv8x across both sensor modalities. Based on these results, the participant selected the YOLOv8x model to generate the final outputs for the test sequences from the GS-RGB and UAV-Therm sensors.

3.2.10 Jiangnan University, China & University of Surrey, UK (JU & U. Surrey)

This submission includes results for all sequences of Challenges 1 and 2. Using the MOTIP framework [6], they fine-tuned the deformable_detr [24] (initialised with COCO pretrained weights) for 10 epochs to perform end-to-end multi-object tracking on each modality. The training was based solely on the data of each individual modality within the dataset, without incorporating any additional datasets.

3.2.11 GE Aerospace Research, USA (GE Aerospace)

GE Aerospace Research submitted two submissions for Challenge 1. Each submission contains results for scenarios "bg6" and "bg8", from GS-RGB and GS-SWIR sensors, respectively. In this challenge, GE Aerospace Research investigated the utility of applying Visual Language Models (VLMs) for the purpose of object detection with respect to people, vessels and vehicles using different sensing modalities. They considered two VLMs, a large state-of-the-art private VLM (Gemini 2.5) [8] and a smaller open-source VLM (pali-gemma) [2]. While the size of Gemini 2.5 is not publicly known, the size of pali-gemma is on the order of 28 billion parameters. They considered both RGB and SWIR imagery. No training was performed in advance, so the results of these experiments represent a zero-shot analysis of the testing data. The testing data was composed of two scenarios. For the first scenario, they used RGB data, resulting in 628 images that were processed by both VLMs. The second scenario had 1515 images, and they focused on the SWIR sensing modality. They considered each image in isolation; no tracking was attempted. Analysis by the VLMs was based on a text-based prompt requesting bounding boxes for the three object classes along with the raw image under consideration, which was either RGB or SWIR. They observed that the large private VLM significantly outperformed the smaller publicly available VLM. However, due to computational limitations, they were only able to process the first 1103 images of the second scenario using the large private VLM. They considered this effort as a benchmarking exercise so that they could better understand the capacity of VLMs for basic object detection on an independent frame-by-frame basis.

3.2.12 Electronics and Telecommunications Research Institute, Republic of Korea (ETRI-Vision)

ETRI-Vision proposed an approach called Target Perception in Multi-Sensor Surveillance: A Coarse-to-Fine Tracking Framework for both Challenges 1 and 2. This approach was applied to all sequences from all sensors in both challenges. They proposed a three-stage pipeline for robust target detection and tracking. Their method was designed to handle visually ambiguous scenes with small-scale objects, heavy occlusions, and diverse modalities by combining coarse-to-fine localisation, contextual classification, and appearance-based tracking.

Stage 1 Multi-Scale Class-Agnostic Detection with Co-DETR + WBF: They first trained Co-DETR [25] as a class-agnostic detector, labelling all targets as a single "object" class. This simplification allowed the model to focus purely on accurate localisation without early misclassification. To improve the detection of small and large objects simultaneously, they ran inference on multiple input

scales (e.g., 640×640 and 2048×1280). The outputs were then fused using Weighted Boxes Fusion (WBF), enhancing precision and recall by combining complementary box predictions across resolutions. From these fused detections, region-of-interest (RoI) crops were extracted for the next stage.

Stage 2 Fine-Grained RoI Classification with Contextual Fusion: Given the low resolution and strong inter-class confusion (e.g., vessel vs. vehicle), fine-grained classification was non-trivial. For each detected RoI, they extracted two views: 1. A cropped view tightly centred on the detected object, and 2. A global scene view for contextual information. Both views were passed through a shared backbone (e.g., Swin Transformer [14]), and their features were combined using a gated attention fusion module. This helped the model leverage both fine details and global spatial cues. They adopted a HERBS-style [4] classification pipeline to enhance performance in this fine-grained, low-resolution regime.

Stage 3 Appearance-Based Tracking with DINO Features: To ensure temporal consistency across frames, they employed an appearance-based tracker built on top of features extracted from a DINO [13] backbone. By leveraging visual embeddings rather than relying solely on spatial proximity, the tracker achieved robust association even in scenes with occlusion, dense object clusters, or abrupt camera motion. This component stabilised identity assignments and complemented their detection–classification pipeline effectively.

4. Results and Analysis

For Challenges 1 and 2, evaluation was performed separately for each sensor type. All submitted sequences were grouped according to sensor type, and evaluation metrics were calculated using all sequences within each group. Each submission was ranked for every metric within its respective sensor group. An average of these ranks across all metrics was then computed to produce a single overall rank for each submission. To ensure fairness when submissions included different numbers of sequences, the number of submitted sequences was also ranked and factored into the final average.

4.1. Evaluation Metrics

For Challenge 1, standard object detection metrics were used: these include the number of true positives (TP), false positives (FP), false negatives (FN), false negative rate (FNR), precision, recall, and F1-score (F1). Moreover, the quality of true positive detections was evaluated by adopting the Multiple Object Tracking Precision (MOTP), computed as the average bounding box overlap between all correctly matched hypotheses and their respective ground-truth objects [16, 18]. Although often used for tracking evaluation, it mainly reflects detector localisation accuracy. Thus, it was included in Challenge 1 alongside standard detection metrics. To summarise the overall performance across all object classes: sums of TP, FP and FN across all classes were used together with mean values of FNR, precision, recall, F1-score, and MOTP.

For Challenge 2, the metrics used in MOTChallenge [16] to assess both localisation accuracy and identity preservation were adopted including: ID switches (IDSW) [16], ID F1-Score (IDF1) [19], Multiple Object Tracking Accuracy (MOTA) [16], Multiple Object Tracking Precision (MOTP) [16], and Higher Order Tracking Accuracy (HOTA) [15]. For each sensor group: the sum of IDSWs and mean values of IDF1, MOTA and HOTA were computed across all sequences for comparison.

For Challenge 3, geolocation accuracy was evaluated using the geolocation estimation deviation metric [17]. This involves computing the minimum, maximum, Mean Absolute Error (MAE), and Root Mean Square Error (RMSE) of the Haversine distance between the estimated and ground-truth coordinates.

4.2. Results Discussion

4.2.1 Challenge 1

The results for each sensor group are discussed in this section. Table 7 shows the results for **GS-RGB**. Among the **UoR** models, **YOLOv8s** slightly outperformed **YOLOv5s**, while **YOLOv11s** produced more detections but with significantly higher FP and FN. **RT-DETR Large** achieved the best F1 among UoR models by favouring recall, though at the cost of higher FN. **NYCU** stood out as the top performer overall, with the highest TP, recall, F1, and lowest FNR. In contrast, **Toho U.** produced results with the lowest F1, likely due to class confusion when different objects shared similar bounding box sizes. **NECTEC** and **ETRI-Vision** also performed strongly; NECTEC achieved the highest precision and MOTP, while ETRI-Vision ranked second across several metrics, including TP, FNR, recall, and MOTP. **JU & U. Surrey** demonstrated reliable performance with a more cautious detection style, as reflected by their lower recall. **UIT** showed moderate performance, with F1 scores slightly below top models. **SKYLD** underperformed in this setting, though their method was originally designed for UAV-RGB imagery and included here for exploratory evaluation on GS-RGB. **GE Aerospace (Gemini)** and (**pali-gemma**) had low FN but were only tested on a single sequence. Overall, **NYCU** achieved the highest average rank across all metrics.

As for the results on **GS-UV** (Table 8), **NYCU** achieved the best overall performance, with the highest F1 and the second-highest MOTP. **ETRI-Vision** followed closely, showing the best performance in terms of TP, FN, FNR, re-

Table 7: Challenge 1 GS-RGB results. For each metric, the best value is shown in bold, and the second-best is underlined. The best overall performance (based on the lowest average rank across all metrics) is highlighted in yellow.

Participant	Submitted/Total sequences	TP	FP	FN	FNR	Precision	Recall	F1-score	MOTP
UoR (YOLOv5)	3/3	4479	205	24193	0.869	0.632	0.131	0.207	0.627
UoR (YOLOv8)	3/3	4945	**195**	23727	0.861	0.639	0.139	0.221	0.627
UoR (YOLOv11)	3/3	7100	1090	21572	0.815	0.578	0.185	0.276	0.630
UoR (RT-DETR)	3/3	11927	3492	16745	0.691	0.515	0.309	0.378	0.638
SKYLD	2/3	1554	1671	14339	0.934	0.324	0.066	0.109	0.579
Toho U.	3/3	276	3207	28396	0.990	0.032	0.010	0.015	0.316
UIT	3/3	14253	2381	14419	0.612	0.572	0.388	0.451	0.644
NYCU	3/3	**21348**	2749	7324	**0.476**	0.605	**0.524**	**0.842**	0.652
NECTEC	3/3	18060	6837	10612	0.537	**0.684**	0.463	0.495	**0.961**
DGA	3/3	10558	1487	18114	0.729	0.588	0.271	0.551	0.635
JU & U. Surrey	3/3	17393	2995	11279	0.563	0.588	0.437	0.751	0.647
GE Aerospace (Gemini)	1/3	3016	392	**2138**	0.540	0.603	0.460	0.771	0.625
GE Aerospace (pali-gemma)	1/3	1047	9081	4107	0.882	0.067	0.118	0.127	0.561
ETRI-Vision	3/3	21053	7517	7619	0.493	0.517	0.507	0.768	0.652

Table 8: Challenge 1 GS-UV results. For each metric, the best value is shown in bold, and the second-best is underlined. The best overall performance (based on the lowest average rank across all metrics) is highlighted in yellow.

Participant	Submitted/Total sequences	TP	FP	FN	FNR	Precision	Recall	F1-score	MOTP
UoR (YOLOv5)	2/2	774	**8**	12531	0.916	**0.994**	0.084	0.146	0.966
UoR (YOLOv8)	2/2	848	25	12457	0.906	0.985	0.094	0.160	0.966
UoR (YOLOv11)	2/2	1263	25	12042	0.858	0.990	0.142	0.222	0.936
UoR (RT-DETR)	2/2	2712	361	10593	0.710	0.874	0.290	0.374	0.955
UIT	2/2	4085	365	9220	0.590	0.940	0.410	0.496	0.974
NYCU	2/2	9252	974	4053	0.264	0.914	0.736	**0.813**	0.982
NECTEC	2/2	6535	1726	6770	0.423	0.789	0.577	0.639	0.979
JU & U. Surrey	2/2	7977	1545	5328	0.358	0.861	0.642	0.734	0.979
ETRI-Vision	2/2	**9289**	3013	**4016**	**0.254**	0.758	**0.746**	0.747	**0.983**

call, and MOTP. However, it performed significantly worse than the top model in terms of FP and precision, placing it in second overall. **JU & U. Surrey** also performed robustly, maintaining good precision while keeping FP manageable. **NECTEC** delivered moderate results with a lower F1-score, while **UIT** achieved high precision but relatively low recall. **RT-DETR**, **YOLOv11**, **YOLOv8**, and **YOLOv5** all underperformed, with particularly low recall and F1-score values, despite having high precision. These models produced too few detections, missing a substantial number of true positives.

Table 9 shows results on **GS-SWIR**, **NYCU** again led the overall ranking, achieving the best TP, FN, FNR, recall, F1-score and the second-best MOTP. **ETRI-Vision** followed closely, obtaining the second-best in multiple metrics and the highest MOTP. **NECTEC** and **JU & U. Surrey** also demonstrated similar solid results, balancing precision and recall. **UIT**, **AIT**, **UWIPL_ETRI**, **GE Aerospace (Gemini)**, and all **UoR's models** showed moderate performance with decent precision but significantly lower recall. This resulted in low F1-score values among these models. Lastly, **GE Aerospace (pali-gemma)** produced results with substantially higher FP, compared to the others. This suggests the limitation of this open-source VLM when applied to non-RGB images, such as SWIR images in this case.

The **GS-Therm** results in Table 10 mirror the GS-UV results, where **NYCU** achieved the best overall ranking, even though **ETRI-Vision** led in more individual metrics. The reason is the same: **ETRI-Vision**'s significantly higher FP and lower precision placed it second overall. **JU & U. Surrey** also performed robustly, attaining the highest precision but with relatively lower recall due to higher FN. **NECTEC** and **UIT** remained consistent performers, with F1-scores above 0.85. The models from **AIT**, **UWIPL_ETRI**, and **UoR (RT-DETR)** demonstrated moderate detection performance with low recall, but still showed a good trade-off between precision and recall, resulting in moderate F1-scores. In contrast, **UoR (YOLOv5, YOLOv8, and YOLOv11)** exhibited notably high precision with extremely low recall, indicating a poor trade-off between these metrics. **Toho U.** also showed severe under-detection, limiting its performance for this sensor type.

Table 9: Challenge 1 GS-SWIR results. For each metric, the best value is shown in bold, and the second-best is underlined. The best overall performance (based on the lowest average rank across all metrics) is highlighted in yellow.

Participant	Submitted/Total sequences	TP	FP	FN	FNR	Precision	Recall	F1-score	MOTP
UoR (YOLOv5)	3/3	4204	222	19411	0.801	0.958	0.199	0.319	0.942
UoR (YOLOv8)	3/3	4518	**92**	19097	0.785	**0.982**	0.215	0.342	0.940
UoR (YOLOv11)	3/3	5364	251	18251	0.740	0.961	0.260	0.389	0.945
UoR (RT-DETR)	3/3	9271	1554	14344	0.537	0.864	0.463	0.544	0.946
AIT	3/3	5531	575	18084	0.735	0.895	0.265	0.396	0.942
UWIPL_ETRI	3/3	6978	4048	16637	0.667	0.625	0.333	0.419	0.915
UIT	3/3	10869	1686	12746	0.497	0.867	0.503	0.617	0.957
NYCU	3/3	**19922**	2235	**3693**	**0.147**	0.908	**0.853**	**0.880**	0.975
NECTEC	3/3	17999	3664	5616	0.219	0.838	0.781	0.808	0.964
JU & U. Surrey	3/3	17480	3794	6135	0.254	0.856	0.746	0.796	0.970
GE Aerospace (Gemini)	1/3	2329	313	7373	0.750	0.918	0.250	0.343	0.940
GE Aerospace (pali-gemma)	1/3	1182	11046	8520	0.876	0.103	0.124	0.112	0.841
ETRI-Vision	3/3	19912	6767	3703	0.155	0.763	0.845	0.801	**0.980**

Table 10: Challenge 1 GS-Therm results. For each metric, the best value is shown in bold, and the second-best is underlined. The best overall performance (based on the lowest average rank across all metrics) is highlighted in yellow.

Participant	Submitted/Total sequences	TP	FP	FN	FNR	Precision	Recall	F1-score	MOTP
UoR (YOLOv5)	3/3	5508	**338**	17951	0.726	0.945	0.274	0.411	0.933
UoR (YOLOv8)	3/3	5228	358	18231	0.737	0.940	0.263	0.396	0.938
UoR (YOLOv11)	3/3	6867	755	16592	0.626	0.923	0.374	0.481	0.931
UoR (RT-DETR)	3/3	11228	2848	12231	0.418	0.816	0.582	0.621	0.943
AIT	3/3	8091	2430	15368	0.550	0.780	0.450	0.503	0.939
UWIPL_ETRI	3/3	13949	2501	9510	0.361	0.855	0.639	0.727	0.937
Toho U.	3/3	1667	1759	21792	0.901	0.595	0.099	0.147	0.913
UIT	3/3	18275	1788	5184	0.209	0.921	0.791	0.851	0.966
NYCU	3/3	19933	1584	3526	0.153	0.945	0.847	**0.893**	0.971
NECTEC	3/3	19807	3832	3652	0.162	0.879	0.838	0.855	0.969
JU & U. Surrey	3/3	18290	996	5169	0.211	**0.956**	0.789	0.865	0.967
ETRI-Vision	3/3	**20336**	4425	**3123**	**0.140**	0.866	**0.860**	0.859	**0.975**

For **UAV-RGB**, the results are presented in Table 11. The top performer was **NYCU**, achieving the best average ranking result. **ETRI-Vision** followed second, due to high FP and low precision, similar to the GS-UV and GS-SWIR results. **SKYLD** produced the highest precision and F1 scores, as well as the second-best TP and FN. Unlike its results on GS-RGB, the results on UAV-RGB were significantly better, as it was designed for UAV imagery, not ground-sensor imagery. **JU & U. Surrey** also performed well, especially in terms of recall and FNR. In contrast, **NECTEC** underperformed on this sensor despite performing well on others. One possible reason could be the lower number of training samples in UAV-RGB, which may have reduced its performance. The rest of the models, **UoR (YOLOv5, YOLOv8, YOLOv11, and RT-DETR)**, **UIT**, **UWIPL_ETRI**, and **Toho U.**, all struggled significantly with F1. Although they produced fewer false positives compared to the top-performing models, these models missed most detections, resulting in F1 scores below 0.27. Notably, all methods performed worse on the GS-RGB test sequence compared to other sensor types. This may be attributed to the presence of a floating mannequin, which only partially resembles a real person, and distant, blurry vessels as shown in Figure 5. These pose challenges for person and vessel detection and contribute to the overall lower performance across all models.

As for UAV-Therm, as shown in Table 12, the top performance was achieved by **NYCU**. **NECTEC** performed best in terms of TP, FN, FNR, and recall; however, it suffered from high FP and low precision, which caused it to be ranked second overall. **JU & U. Surrey** followed closely, with the second-best FNR and recall. For **UoR (retrained YOLOv11)** and **UIT**, despite having lower TP, these models produced substantially fewer FP compared to the top-performing models, making them more precise at the cost of missing some ground-truth detections. As for **ETRI-Vision**, its performance on this sensor was not as strong as on others due to a large number of false posi-

Table 11: Challenge 1 UAV-RGB results. For each metric, the best value is shown in bold, and the second-best is underlined. The best overall performance (based on the lowest average rank across all metrics) is highlighted in yellow.

Participant	Submitted/Total sequences	TP	FP	FN	FNR	Precision	Recall	F1-score	MOTP
UoR (YOLOv5)	1/1	226	108	1766	0.914	0.551	0.086	0.147	0.956
UoR (YOLOv8)	1/1	212	**19**	1780	0.927	0.677	0.073	0.130	0.974
UoR (YOLOv11)	1/1	248	33	1744	0.905	0.687	0.095	0.167	0.960
UoR (RT-DETR)	1/1	578	362	1414	0.810	0.426	0.190	0.263	0.969
SKYLD	1/1	769	237	1223	0.610	**0.784**	0.390	**0.521**	0.910
UWIPL_ETRI	1/1	107	130	1885	0.959	0.354	0.041	0.073	0.943
Toho U.	1/1	95	284	1897	0.969	0.138	0.031	0.050	0.485
UIT	1/1	109	114	1883	0.894	0.524	0.106	0.174	0.933
NYCU	1/1	617	243	1375	0.524	0.781	0.476	0.486	0.970
NECTEC	1/1	603	648	1389	0.647	0.467	0.353	0.384	0.966
JU & U. Surrey	1/1	641	351	1351	0.487	0.736	0.513	0.471	0.953
ETRI-Vision	1/1	**848**	1643	**1144**	**0.465**	0.449	**0.535**	0.377	**0.983**

Table 12: Challenge 1 UAV-Therm results. For each metric, the best value is shown in bold, and the second-best is underlined. The best overall performance (based on the lowest average rank across all metrics) is highlighted in yellow.

Participant	Submitted/Total sequences	TP	FP	FN	FNR	Precision	Recall	F1-score	MOTP
UoR (YOLOv5)	3/3	3641	61	25946	0.852	0.936	0.148	0.250	0.949
UoR (YOLOv8)	3/3	3142	**55**	26445	0.848	0.960	0.152	0.252	0.947
UoR (YOLOv11)	3/3	4502	86	25085	0.820	**0.977**	0.180	0.282	0.946
UoR (RT-DETR)	3/3	9440	773	20147	0.628	0.827	0.372	0.483	0.951
UoR (retrained YOLOv11s)	3/3	12080	1926	17507	0.506	0.891	0.494	0.629	0.956
UWIPL_ETRI	3/3	6417	9287	23170	0.852	0.287	0.148	0.195	0.632
Toho U.	3/3	1924	907	27663	0.943	0.227	0.057	0.091	0.326
UIT	3/3	15625	2555	13962	0.513	0.813	0.487	0.608	0.971
NYCU	3/3	19718	4234	9869	0.340	0.873	0.660	0.739	0.977
NECTEC	3/3	**20051**	10650	**9536**	**0.268**	0.784	**0.732**	0.729	0.968
DGA	3/3	6077	284	23510	0.778	0.955	0.222	0.343	0.964
JU & U. Surrey	3/3	18727	5557	10860	0.299	0.845	0.701	**0.741**	0.967
ETRI-Vision	3/3	19075	9528	10512	0.541	0.844	0.459	0.515	**0.978**

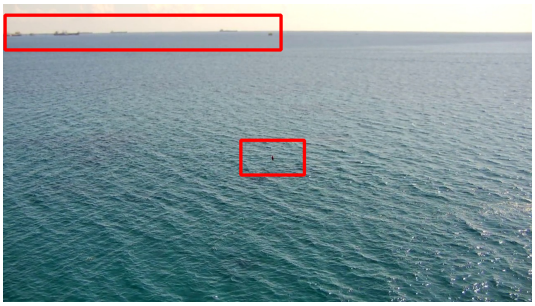


Figure 5: Sample image from the UAV-RGB test sequence that shows a floating mannequin (representing a person in water) and distant vessels, both posing detection challenges.

tives; however, it achieved the highest MOTP. Interestingly, **UoR (YOLOv5, YOLOv8, YOLOv11, and RT-DETR)** and **DGA**, although not retrained on the UAV-Therm data, achieved highly precise predictions with low FP. Nonetheless, they missed many ground-truth objects. This demonstrates the ability of pretrained YOLO models to detect objects in thermal UAV imagery, despite being originally trained on the COCO dataset, which mainly contains RGB images. **UWIPL_ETRI** detected a large number of objects; however, its number of false positives exceeded true positives, indicating a precision issue. **Toho U.**, on the other hand, detected far fewer objects compared to other models, reflecting severe under-detection that may be attributed to the BBox size problem.

Overall, the results highlight key challenges and observations in multi-sensor object detection. Fine-tuning deep learning models on sensor-specific data improved object detection in corresponding image sequences, but often

led to a significant increase in the number of false positives. This poses concerns for some real-world applications, where false alarms might be undesirable. Another point is that performance on UAV images was generally worse than on ground sensor data, as most models rely on pretrained backbones trained on datasets with perspectives similar to ground-sensor images. This emphasises the challenge in developing detection models for sensor platforms across various altitudes. Finally, deep learning-based detection models demonstrated superior performance compared to VLMs in this challenge. However, VLM-based methods in this challenge were limited to zero-shot prompting, and their performance could improve with further prompt tuning or more advanced techniques.

4.2.2 Challenge 2

Tables 13 - 18 show the results of Challenge 2 for GS-RGB, GS-UV, GS-SWIR, GS-Therm, UAV-RGB and UAV-Therm, respectively. From these tables, both **JU & U. Surrey** and **ETRI-Vision** were top performers across all sensor groups. **ETRI-Vision** achieved first place in GS-SWIR, GS-Therm, UAV-RGB, and UAV-Therm, while **JU & U. Surrey** ranked first in GS-UV. For GS-RGB, both methods achieved a joint first place. These results underscore the effectiveness of their approaches, particularly in terms of the association and re-identification strategies employed. **ETRI-Vision** used appearance-based tracking with DINO features, while **JU & U. Surrey** adopted the MOTIP framework, which trains an ID prediction model using image features and ground-truth tracks.

Considering **AIT**'s tracker on the **GS-SWIR** and **GS-Therm** datasets, the method performed well, particularly in terms of IDSW. It consistently maintained strong identity consistency for successfully tracked objects, as reflected by its low IDSW count. However, it missed more detections than the top two methods, indicated by lower IDF1 and MOTA scores. These results suggest that the tracker prioritised high-confidence, stable tracks while discarding lower-confidence or ambiguous detections. This trade-off reflects a conservative strategy that favours precision and ID stability over broader coverage. This could be beneficial in applications where reliable individual tracks are critical.

Comparing tracking methods from **UoR** for all sensor groups, **RT-DETR + BoT-SORT** generally performed better than the method using **RT-DETR + ByteTrack**, while **RT-DETR + DeepSORT** performed the worst. This difference likely stems from the underlying association and motion modelling techniques. BoT-SORT incorporates appearance features and camera-motion compensation, leading to improved handling of occlusions, identity switches, and constantly moving sensors. ByteTrack, although effective, relies more heavily on motion cues, which can be less reliable in scenarios where the sensors are also in motion. DeepSORT is an older method with simpler appearance modelling and association strategies. Therefore, it is likely to underperform the other two methods.

Table 13: Challenge 2 GS-RGB results. For each metric, the best value is shown in bold, and the second-best is underlined. The best overall performance (based on the lowest average rank across all metrics) is highlighted in yellow.

Participant	IDSW	IDF1	MOTA	HOTA
UoR (RT-DETR + BoT-SORT)	1098	0.266	0.257	0.279
UoR (RT-DETR + ByteTrack)	1289	0.201	0.187	0.237
UoR (RT-DETR + DeepSORT)	**650**	0.144	0.120	0.187
JU & U. Surrey (MOTIP)	864	**0.572**	0.495	**0.517**
ETRI-Vision (DINO)	710	0.501	**0.557**	0.490

Table 14: Challenge 2 GS-UV results. For each metric, the best value is shown in bold, and the second-best is underlined. The best overall performance (based on the lowest average rank across all metrics) is highlighted in yellow.

Participant	IDSW	IDF1	MOTA	HOTA
UoR (RT-DETR + BoT-SORT)	169	0.262	0.160	0.247
UoR (RT-DETR + ByteTrack)	313	0.222	0.147	0.209
UoR (RT-DETR + DeepSORT)	175	0.138	0.086	0.177
JU & U. Surrey (MOTIP)	**141**	**0.596**	0.458	0.418
ETRI-Vision (DINO)	277	0.533	**0.576**	**0.454**

Table 15: Challenge 2 GS-SWIR results. For each metric, the best value is shown in bold, and the second-best is underlined. The best overall performance (based on the lowest average rank across all metrics) is highlighted in yellow.

Participant	IDSW	IDF1	MOTA	HOTA
UoR (RT-DETR + BoT-SORT)	915	0.284	0.263	0.278
UoR (RT-DETR + ByteTrack)	800	0.271	0.260	0.256
UoR (RT-DETR + DeepSORT)	399	0.209	0.180	0.209
AIT	**106**	0.267	0.194	0.298
JU & U. Surrey (MOTIP)	826	**0.656**	0.504	0.510
ETRI-Vision (DINO)	477	0.574	**0.675**	**0.542**

4.2.3 Challenge 3

Table 19 presents the results for Challenge 3 using **UoR**'s method. In scenarios "cy4" and "rd7", the method achieved low MAE and RMSE (below 11 meters), as the target vessels were captured at low pitch angles and relatively short distances. In "bg7" and "cy6", a combination of near and distant vessels, along with varied pitch angles, resulted in moderate errors. Scenario "cy5" proved the most challenging, with the target captured from a significant distance and a high pitch angle (close to 0). Consequently, the method yielded MAE and RMSE values exceeding 40 meters.

Table 16: Challenge 2 GS-Therm results. For each metric, the best value is shown in bold, and the second-best is underlined. The best overall performance (based on the lowest average rank across all metrics) is highlighted in yellow.

Participant	IDSW	IDF1	MOTA	HOTA
UoR (RT-DETR + BoT-SORT)	1502	0.275	0.275	0.253
UoR (RT-DETR + ByteTrack)	1527	0.253	0.272	0.227
UoR (RT-DETR + DeepSORT)	757	0.186	0.178	0.176
AIT	**258**	0.297	0.273	0.240
JU & U. Surrey (MOTIP)	799	**0.537**	0.589	**0.492**
ETRI-Vision (DINO)	534	0.474	**0.698**	0.474

Table 17: Challenge 2 UAV-RGB results. For each metric, the best value is shown in bold, and the second-best is underlined. The best overall performance (based on the lowest average rank across all metrics) is highlighted in yellow.

Participant	IDSW	IDF1	MOTA	HOTA
UoR (RT-DETR + DeepSORT)	59	0.131	0.104	0.173
UoR (RT-DETR + ByteTrack)	83	0.166	0.158	0.215
UoR (RT-DETR + BoT-SORT)	66	0.165	0.078	0.233
JU & U. Surrey (MOTIP)	119	0.296	0.115	0.411
ETRI-Vision (DINO)	**38**	**0.485**	**0.243**	**0.469**

Table 18: Challenge 2 UAV-Therm results. For each metric, the best value is shown in bold, and the second-best is underlined. The best overall performance (based on the lowest average rank across all metrics) is highlighted in yellow.

Participant	IDSW	IDF1	MOTA	HOTA
UoR (RT-DETR + BoT-SORT)	1131	0.296	0.234	0.292
UoR (RT-DETR + ByteTrack)	1349	0.211	0.203	0.221
UoR (RT-DETR + DeepSORT)	**495**	0.133	0.125	0.176
JU & U. Surrey (MOTIP)	2639	0.529	0.398	**0.479**
ETRI-Vision (DINO)	946	**0.534**	**0.452**	0.473

5. Conclusions

PETS2025 introduces three challenges aimed at advancing maritime and coastal surveillance using multi-authority, multi-platform, multi-spectral sensors. Challenge 1 focuses on detecting and classifying persons, vessels, and vehicles across RGB, thermal, UV, and SWIR imagery from ground and UAV sensors. Challenge 2 addresses long-term tracking of these targets, requiring consistent object IDs across occlusions and diverse terrains. Challenge 3 involves geolocating persons and vessels in thermal UAV images using provided telemetry data. Together, these challenges promote robust multimodal detection, tracking, and geolocation approximation in maritime and coastal environments.

A total of 12 teams joined Challenge 1, 4 teams participated in Challenge 2, and Challenge 3 received a single submission. The majority of object detection methods centred

Table 19: Challenge 3 results.

Scenario	Min	Max	MAE	RMSE
bg7	9.03	77.86	33.37	40.56
cy4	2.16	19.73	5.8	6.62
cy5	20.2	75.08	42.93	44.83
cy6	15.48	29.14	21.57	21.99
rd7	4.52	16.33	9.79	10.42

on using deep learning-based detectors, particularly variants of the YOLO family (e.g., YOLOv5, YOLOv8, YOLOv11), which were the most widely adopted across participants. Several teams trained or fine-tuned these detectors on the PETS2025 dataset, often tailoring them to different sensor modalities (RGB, thermal, SWIR) or target domains (land vs maritime). Transformer-based models such as RT-DETR and deformable DETR were also explored, especially by teams aiming to improve performance on small or occluded objects. A few groups incorporated more advanced components, such as semantic segmentation (e.g., SAM2), bounding box modelling (e.g., GMM), or auxiliary modules to compensate for false negatives.

The best-performing model was obtained using PRB-FPN, fine-tuned on a combined dataset from all sensors to ensure robustness across multiple modalities. The second-best approach employed a three-stage coarse-to-fine pipeline, which initially detects objects in a class-agnostic manner at multiple scales using Co-DETR and Weighted Boxes Fusion, followed by fine-grained classification that leverages both local and contextual features through a Swin Transformer backbone. Another effective method utilised the MOTIP framework, integrating object detection and object association into a single end-to-end trainable model. Additionally, zero-shot object detection with vision-language models (VLMs) was investigated, highlighting the potential of prompt-driven multimodal reasoning without the need for dataset-specific training.

For object tracking, common strategies involved detect-and-track pipelines using established tracking methods based on the Hungarian algorithm, with detection results from YOLO or transformer-based backbones serving as inputs. To maintain consistent tracking through occlusions and temporary disappearances, some methods incorporated advanced techniques for enhanced object association and re-identification. These included appearance-based tracking using DINO embeddings to ensure robust temporal consistency across challenging multi-sensor scenes, as well as MOTIP, which employs image features as input to an ID prediction model trained on ground-truth tracks. This enables the model to associate detected objects across frames effectively based on both appearance and motion cues.

For the geolocation approximation task, a method was

submitted that projects the UAV's FOV onto the ground using its altitude, orientation, and FOV to calculate ground distances and image corners, from which a homography matrix maps image pixels to geospatial coordinates. To handle low-pitch angles with sky regions, horizon-aware correction and vertical pixel scaling techniques were applied to improve the accuracy. The results show that the method achieved an error of less than 11 metres in less challenging cases (closer targets with an optimal UAV camera pitch angle), and under 45 metres in more challenging cases (distant targets with the UAV camera pitch angle nearly horizontal).

All submissions demonstrate that detection, tracking, and geolocation approximation remain challenging tasks, particularly due to varying sensor modalities, platforms, and complex maritime and coastal backgrounds. Future work will focus on advancing these challenges further to enable the development of more effective detection, tracking, and geolocation methods in maritime and coastal environments.

6. Acknowledgements

This challenge was funded by the European Union's Horizon Europe Research and Innovation Programme under grant agreement No 101073985 (EURMARS).

References

[1] N. Aharon, R. Orfaig, and B.-Z. Bobrovsky. Bot-sort: Robust associations multi-pedestrian tracking. arXiv preprint arXiv:2206.14651, 2022.

[2] L. Beyer, A. Steiner, A. S. Pinto, A. Kolesnikov, X. Wang, D. Salz, M. Neumann, I. Alabdulmohsin, M. Tschannen, E. Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. arXiv preprint arXiv:2407.07726, 2024.

[3] C.-T. Chien, R.-Y. Ju, K.-Y. Chou, E. Xieerke, and J.-S. Chiang. YOLOv8-AM: YOLOv8 based on effective attention mechanisms for pediatric wrist fracture detection. IEEE Access, 13:52461–52477, 2025.

[4] P.-Y. Chou, Y.-Y. Kao, and C.-H. Lin. Fine-grained visual classification with high-temperature refinement and background suppression. arXiv preprint arXiv:2303.06442, 2023.

[5] European Border and Coast Guard Agency – Frontex. Annual Risk Analysis 2025/2026. Publications Office of the European Union, 2025.

[6] R. Gao, J. Qi, and L. Wang. Multiple object tracking as id prediction, 2025.

[7] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun. Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430, 2021.

[8] Google DeepMind. Gemini 2.5: Pushing the frontier with advanced reasoning and agentic ai. Technical report, Google DeepMind, 2025.

[9] G. Jocher. Ultralytics yolov5, 2020.

[10] G. Jocher, A. Chaurasia, and J. Qiu. Ultralytics yolov8, 2023.

[11] G. Jocher and J. Qiu. Ultralytics yolo11, 2024.

[12] G. Jocher, J. Qiu, and A. Chaurasia. Ultralytics yolo (version 8.0.0), 2023. Computer software.

[13] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In European Conference on Computer Vision, pages 38–55. Springer, 2024.

[14] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international conference on computer vision, pages 10012–10022, 2021.

[15] J. Luiten, A. Osep, P. Dendorfer, P. Torr, A. Geiger, L. Leal-Taixé, and B. Leibe. Hota: A higher order metric for evaluating multi-object tracking. International journal of computer vision, 129:548–578, 2021.

[16] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. Mot16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831, 2016.

[17] E. Namazi, R. Mester, C. Lu, and J. Li. Geolocation estimation of target vehicles using image processing and geometric computation. Neurocomputing, 499:35–46, 2022.

[18] L. Patino, J. Boyle, J. Ferryman, J. Auer, J. Pegoraro, R. Pflugfelder, M. Cokbas, J. Konrad, P. Ishwar, G. Slavic, L. Marcenaro, Y. Jiang, Y. Jin, H. Ko, G. Zhao, G. Ben-Yosef, and J. Qiu. Pets2021: Through-foliage detection and tracking challenge and evaluation. In 2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pages 1–10, 2021.

[19] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In European conference on computer vision, pages 17–35. Springer, 2016.

[20] Y. Tian, Q. Ye, and D. Doermann. Yolov12: Attention-centric real-time object detectors. arXiv preprint arXiv:2502.12524, 2025.

[21] N. Wojke, A. Bewley, and D. Paulus. Simple online and realtime tracking with a deep association metric. In 2017 IEEE International Conference on Image Processing (ICIP), pages 3645–3649. IEEE, 2017.

[22] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, and X. Wang. Bytetrack: Multi-object tracking by associating every detection box. 2022.

[23] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen. Detrs beat yolos on real-time object detection, 2023.

[24] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai. Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159, 2020.

[25] Z. Zong, G. Song, and Y. Liu. Detrs with collaborative hybrid assignments training. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 6748–6758, October 2023.