HUMAN SURVEILLANCE IN MARITIME ENVIRONMENTS USING DRONE AERIAL FOOTAGE FOR SEARCH AND RESCUE

A PROJECT REPORT

Submitted by

SNEHA S 2022510003

LAVANYA J 2022510004

SURYA K S 2022510022

RAMEEZ AKTHER M 2022510025

in partial fulfilment for the award of the degree

of

BACHELOR OF TECHNOLOGY

in

ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

DEPARTMENT OF INFORMATION TECHNOLOGY

MADRAS INSTITUTE OF TECHNOLOGY

ANNA UNIVERSITY : CHENNAI 600 025

MAY 2025

ANNA UNIVERSITY : CHENNAI 600 025

BONAFIDE CERTIFICATE

Certified that this project report **"HUMAN SURVEILLANCE IN MARITIME ENVIRONMENTS USING DRONE AERIAL FOOTAGE FOR SEARCH AND RESCUE"** is the bonafide work of **"RAMEEZ AKTHER M 2022510025, SURYA K S 2022510022, LAVANYA J 2022510004, SNEHA S 2022510003"** who carried out the project work under my supervision.

SIGNATURE **SIGNATURE**

HEAD OF THE DEPARTMENT **SUPERVISOR**

ABSTRACT

Search and Rescue (SAR) operations in maritime environments are challenging due to vast search areas, rapidly changing sea conditions, and the difficulty of manually monitoring drone footage in real time. To address this, the proposed system automates human surveillance in ocean environments by integrating object detection, tracking, and geo-localization on drone aerial footage. An enhanced version of RT-DETR (Real-Time Detection Transformer) is developed by incorporating a ConvNeXt backbone, Bidirectional Feature Pyramid Network (BiFPN), and Feature Selection Gate (FSG), enabling improved multi-scale feature extraction and strong performance on small and partially occluded objects such as swimmers. The enhanced model achieves a mAP of 49.2%, improving upon the baseline RT-DETR (44.9%) by +4.3%, along with a 2× increase in AP_small, demonstrating superior capability in identifying tiny maritime objects. To maintain consistent identity tracking across frames, a Visually-Augmented Kalman Tracker (VAKT) combining motion prediction and HSV appearance cues is used, reducing ID switches and achieving a MOTA of 79.7%, outperforming traditional trackers such as DeepSORT and ByteTrack (≈77–78%). In addition, a geo-pixel localization module projects image-plane detections into real-world GPS coordinates using drone metadata and camera geometry, enabling location-aware monitoring. The system outputs object IDs, snapshots, and latitude–longitude positions, providing real-time situational awareness to rescue personnel. The results demonstrate that the proposed pipeline significantly enhances detection accuracy, tracking stability, and localization reliability, offering an effective AI-assisted aerial surveillance solution for maritime SAR operations.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVATION

AP	Average Precision
AP@0.50 / AP@0.75	Average Precision at IoU thresholds 0.50 / 0.75
AP (small / medium / large)	Average Precision based on object size
AR	Average Recall
BiFPN	Bidirectional Feature Pyramid Network
CNN	Convolutional Neural Network
ConvNeXt	Modernized CNN architecture used as backbone
DETR	Detection Transformer
DN Loss	Denoising Loss (RT-DETR auxiliary loss branch)
FPN	Feature Pyramid Network
FSG	Feature Selection Gate
FP	False Positive
FN	False Negative
GIoU	Generalized Intersection over Union
GPS	Global Positioning System
HSV	Hue Saturation Value color space
IDF1	Identification F1 Score (track identity consistency metric)
IoU	Intersection over Union
mAP	Mean Average Precision
MOTA	Multiple Object Tracking Accuracy

MOTP	Multiple Object Tracking Precision
MOT	Multiple Object Tracking
OCSORT	Observation-Centric SORT tracker
Re-ID	Re-identification (appearance-based similarity tracking)
RT-DETR	Real-Time Detection Transformer
SAR	Search and Rescue
SORT	Simple Online Real-time Tracking
UAV	Unmanned Aerial Vehicle (Drone)
VAKT	Visually-Augmented Kalman Tracker
VFL	Varifocal Loss (classification loss used in RT-DETR)

CHAPTER 1
INTRODUCTION

1.1 GENERAL

Search and Rescue (SAR) operations in ocean environments are among the most time-critical and resource-intensive rescue missions. Unlike land-based emergencies, maritime distress situations are highly unpredictable due to dynamic environmental factors such as wave turbulence, wind direction, lighting variations, and the constant motion of both victims and rescue assets. Traditionally, surveillance teams rely on manual monitoring of live drone feeds or binocular scanning from rescue vessels or helicopters. These approaches are slow, prone to human fatigue, and inefficient for wide-area search, especially when dealing with small, partially submerged human bodies that are barely distinguishable from the surrounding sea. The emergence of Unmanned Aerial Vehicles (UAVs), or drones, equipped with high-resolution cameras has introduced a breakthrough in SAR missions by enabling continuous aerial monitoring of vast regions at minimal operational cost. However, without automation, the process still depends heavily on human attention, creating a bottleneck when rapid and accurate identification of distressed individuals becomes crucial.

Recent advancements in computer vision and deep learning have made it possible to automatically detect and analyze objects in visual data. State-of-the-art detection models such as YOLO, Faster R-CNN, and DETR have demonstrated remarkable performance in real-time object detection across various applications. Yet, models trained on standard datasets typically struggle in maritime environments due to extremely small object sizes, unstable camera motion caused by drone vibrations, object occlusion behind waves or boats, and the similarity between foreground objects and the background. These challenges significantly

reduce detection accuracy and continuity in tracking. Therefore, a specialized framework capable of addressing these conditions is required.

To bridge this gap, the proposed system introduces an AI-driven pipeline that integrates

1. Real-time object detection using an enhanced RT-DETR model,
2. Robust identity tracking using a Visually-Augmented Kalman Tracker (VAKT), and
3. Geo-pixel localization to estimate the GPS position of detected targets.

By combining computer vision with drone telemetry metadata, the system transforms conventional drone surveillance into an autonomous decision-support tool capable of assisting SAR teams in identifying distressed individuals quickly and accurately.

1.2 PROBLEM DEFINITION

Although modern drones can capture high-resolution aerial footage, the challenge lies in extracting actionable information from the video stream. In most ocean surveillance systems, a human operator watches the video feed and manually identifies objects of interest. This method becomes unreliable due to

- Small object sizes (swimmers appear as tiny pixels from high altitudes),
- Occlusions caused by waves, reflections, and lighting variations,
- Absence of GPS location for detected objects,
- Difficulty in maintaining identity tracking across frames.

Existing object detection and tracking systems mainly function in the image plane; they only reveal where the object is in the frame but not where it exists in the real world. During SAR missions, responders need precise latitude and longitude coordinates to deploy rescue boats or helicopters efficiently. Without

automated geo-localization, the detected object provides limited operational value.

Thus, the core research problem is, “How can a drone autonomously detect, track, and geo-localize swimmers and boats in real-time to support rapid Search and Rescue missions?”

1.3 OBJECTIVE OF THE PROJECT

The objectives of this project are

1. To develop an AI-based system that automatically detects swimmers, swimmers with life jackets, and boats from drone aerial footage in real time.
2. To enhance small-object detection performance using a custom-modified RT-DETR architecture equipped with ConvNeXt, BiFPN, and Feature Selection Gate (FSG).
3. To track detected objects across frames and maintain consistent identity using a Kalman Filter-based visual tracking approach.
4. To provide real-world location mapping by converting pixel coordinates into GPS latitude and longitude using drone metadata.
5. To generate alert packets containing object IDs, snapshots, counts, and geo-coordinates for rescue teams.

1.4 SCOPE OF THE PROJECT

The scope of the system includes

- Working with aerial maritime datasets (SeaDronesSee and MOBDrone).
- **Detecting three object classes relevant to SAR:** swimmers, swimmers with life jackets, and boats.
- Maintaining identity tracking even when temporary occlusion or scale variation occurs.

- Geo-localizing detected objects and exporting real-world coordinates for operational use.
- Producing outputs such as annotated tracking video, per-frame object logs, and snapshot images.

The project focuses solely on search and surveillance; it does not include autonomous drone navigation or physical rescue operations.

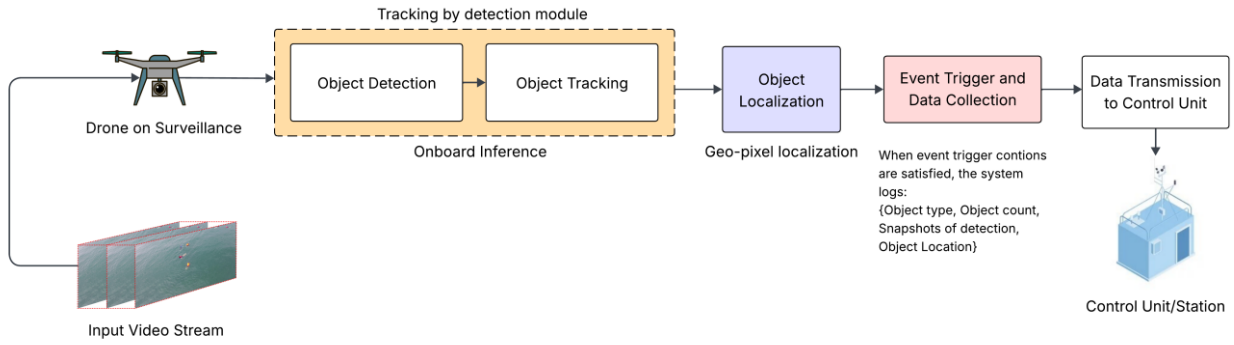

Fig 1. Scope diagram of the proposed work

1.5 SIGNIFICANCE AND CONTRIBUTION

The major contributions of this project are summarized below

1. **Enhanced RT-DETR architecture:** Improved detection performance through cross-scale feature fusion and attention gating mechanisms.
2. **Lightweight tracking system (VAKT):** More efficient than DeepSORT, with lower ID switches and fragmentation.
3. **Geo-pixel localization:** Converts image detections into latitude–longitude, enabling real SAR applicability.
4. **End-to-end integrated framework:** Automated detection, tracking, and localization, transforming raw drone footage into actionable rescue intelligence.

The system improves operational efficiency, reduces manual monitoring effort, and significantly speeds up decision-making during SAR missions, ultimately contributing to saving lives.

1.6 STRUCTURE OF THE THESIS

This thesis is organized into six chapters. Chapter 1 introduces the motivation, problem definition, objectives, scope, and contribution of the work. Chapter 2 presents the literature survey, summarizing existing research in object detection, multi-object tracking, and geo-localization and highlighting the research gaps. Chapter 3 explains the proposed methodology, detailing the enhanced RT-DETR detection model, the Visually Augmented Kalman Tracker (VAKT), and the geo-localization module. Chapter 4 describes the implementation setup, including dataset preprocessing, training configuration, and evaluation metrics. Chapter 5 presents the experimental results and performance analysis of detection, tracking, and localization. Finally, Chapter 6 concludes the thesis and outlines possible future enhancements.

CHAPTER 2
LITERATURE REVIEW

2.1 GENERAL

The evolution of computer vision, particularly through deep learning, has profoundly transformed the capabilities of autonomous drone-based surveillance systems. Early object detection approaches relied heavily on handcrafted features such as SIFT and HOG, which were ineffective in dynamic maritime environments where objects are extremely small and constantly changing due to waves, reflections, and lighting variations. The advent of Convolutional Neural Networks (CNNs) marked a major breakthrough, with models such as Faster R-CNN, SSD, and YOLO demonstrating significant improvements in real-time object detection.

However, these CNN-based detectors continue to face challenges in ocean-based scenarios, especially when targets such as swimmers, floating humans, or life jackets occupy less than 1% of the frame and blend into the cluttered background. Moreover, maritime aerial footage is affected by drone-induced motion, low resolution at high altitudes, and unpredictable occlusions, all of which degrade both detection and tracking accuracy.

In recent years, deep learning research has moved beyond CNN-only architectures toward Transformer-based designs, offering enhanced robustness and contextual understanding. For maritime Search and Rescue (SAR) missions, systems must not only detect but also consistently track individuals across frames and estimate their real-world geographic positions. Therefore, the literature relevant to this project spans three key domains: object detection, multi-object tracking, and geo-localization.

2.2 OBJECT DETECTION IN AERIAL AND MARITIME ENVIRONMENTS

Traditional object detection models such as Faster R-CNN and SSD introduced region-based and single-shot detection strategies, but they are constrained by limited receptive fields and a weak ability to model global context. The YOLO family improved inference speed, enabling real-time deployment, but their accuracy drops sharply when detecting small targets in aerial imagery.

To address this, recent studies have focused on multi-scale feature fusion, resulting in architectures such as the Feature Pyramid Network (FPN) and Bidirectional Feature Pyramid Network (BiFPN) [6]. These frameworks improve contextual understanding by fusing features across scales, which is particularly beneficial for detecting tiny objects in drone imagery. Liu et al. [3] proposed ESOD, an efficient small object detector that enhances contextual features in high-resolution images. Similarly, Jiang et al. [2] introduced MFFSODNet, demonstrating that multi-scale feature aggregation significantly boosts detection performance when objects are nearly indistinguishable from the background.

A major advancement came with Transformer-based detectors. The DETR framework and its successors—Deformable DETR, DINO-DETR, and RT-DETR—replaced anchor-based heuristics and manual NMS tuning with end-to-end attention-based learning. Ma et al. [5] presented OWRT-DETR, a Transformer-based detector optimized for water-based UAV imagery, which achieved superior performance in detecting swimmers in oceanic environments. Cheng et al. [1] developed EF-DETR, a lightweight Transformer model that maintains real-time inference capability by removing redundant encoder layers.

Meanwhile, Liu et al. [4] proposed ConvNeXt, a CNN architecture modernized with Transformer-inspired design principles—such as depthwise convolutions, GELU activations, and layer normalization—achieving improved feature

representation with lower memory cost. In addition, Tan et al. [6] introduced EfficientDet, incorporating BiFPN to refine multi-scale feature fusion through learnable weights, which further enhances tiny-object detection performance.

Collectively, these studies indicate that integrating Transformer-based detectors with advanced multi-scale fusion modules remains the most effective approach for small-object detection in SAR scenarios. This insight directly motivates our proposed Enhanced RT-DETR model, which combines ConvNeXt, BiFPN, and FSG to achieve superior detection accuracy in drone-based ocean surveillance.

2.3 MULTI-OBJECT TRACKING IN UAV/MARITIME VIDEOS

In multi-object tracking (MOT), the tracking-by-detection paradigm remains the dominant framework. The SORT algorithm introduced a simple yet efficient tracking mechanism that combines Kalman filtering with IoU-based matching, achieving real-time performance. However, SORT struggles with frequent ID switches when objects experience occlusions or cross paths—issues that are particularly common in ocean footage, where swimmers can temporarily disappear behind waves.

The DeepSORT method [7] improved identity preservation by integrating appearance-based features (Re-ID embeddings), but this comes with higher computational overhead, making it unsuitable for real-time UAV applications. Later models such as StrongSORT and ByteTrack refined association logic to enhance robustness, yet they still depend heavily on deep appearance models.

Xue et al. [8] demonstrated that lightweight visual and motion cues can achieve comparable robustness without expensive feature extraction. This is particularly important for SAR environments, where targets are small and exhibit limited visual variation. Motivated by this insight, our work introduces the Visually-Augmented Kalman Tracker (VAKT), which combines Kalman filtering with HSV-based appearance cues—striking a balance between computational

efficiency and identity consistency. This hybrid tracking approach aligns with trends in recent literature that emphasize lightweight, real-time tracking strategies.

2.4 GEO-LOCALIZATION IN UAV DETECTION SYSTEMS

While detection and tracking systems operate in the image plane, effective SAR operations require the transformation of object coordinates into real-world geographic locations. The literature identifies two main approaches for this purpose

1. **Metadata and Ray Projection Technique**

   This method uses the drone's intrinsic parameters, altitude, orientation, and GPS metadata to project pixel coordinates into geographic coordinates. It is the most accurate method and is widely used in professional UAV mapping systems.

2. **Visual Projection Approximation**

   This alternative is applied when telemetry data is unavailable. It estimates the ground distance per pixel based on the drone's altitude and the camera's field of view. Although less precise, it remains practical for scenarios lacking detailed metadata.

Prior studies emphasize that localization accuracy largely depends on the availability of drone pose data, camera calibration, and sensor metadata. However, most existing works focus on mapping and remote sensing applications rather than real-time integration with detection and tracking modules.

Our system bridges this gap by incorporating both strategies—using metadata-based ray projection when available and visual projection estimation when metadata is absent—allowing seamless and flexible geo-localization in real-time SAR applications.

2.5 LITERATURE GAP AND MOTIVATION

From the above studies, several research gaps are evident

- Object detection models still struggle to identify extremely small objects such as swimmers in drone footage.
- Trackers frequently lose object identities when targets undergo occlusions or visual distortions in ocean environments.
- Geo-localization is rarely integrated in real time alongside detection and tracking components.

These limitations highlight the need for a unified framework that can

- Detecting ocean-based targets,
- Tracking them consistently across frames, and
- Converting image-plane detections into geographic coordinates.

To address this need, the present work integrates Enhanced RT-DETR (for small-object detection), VAKT (for robust tracking), and Geo-Localization (for real-world coordinate mapping) into a single end-to-end Search and Rescue (SAR) support system optimized for drone-based operations.

CHAPTER 3

PROPOSED WORK

3.1 OVERALL SYSTEM FLOW

The proposed system automates human surveillance in maritime environments by performing real-time object detection, multi-object tracking, and geo-localization on aerial drone footage. The end goal is to identify swimmers or boats in distress and provide their GPS coordinates for Search and Rescue (SAR) teams.

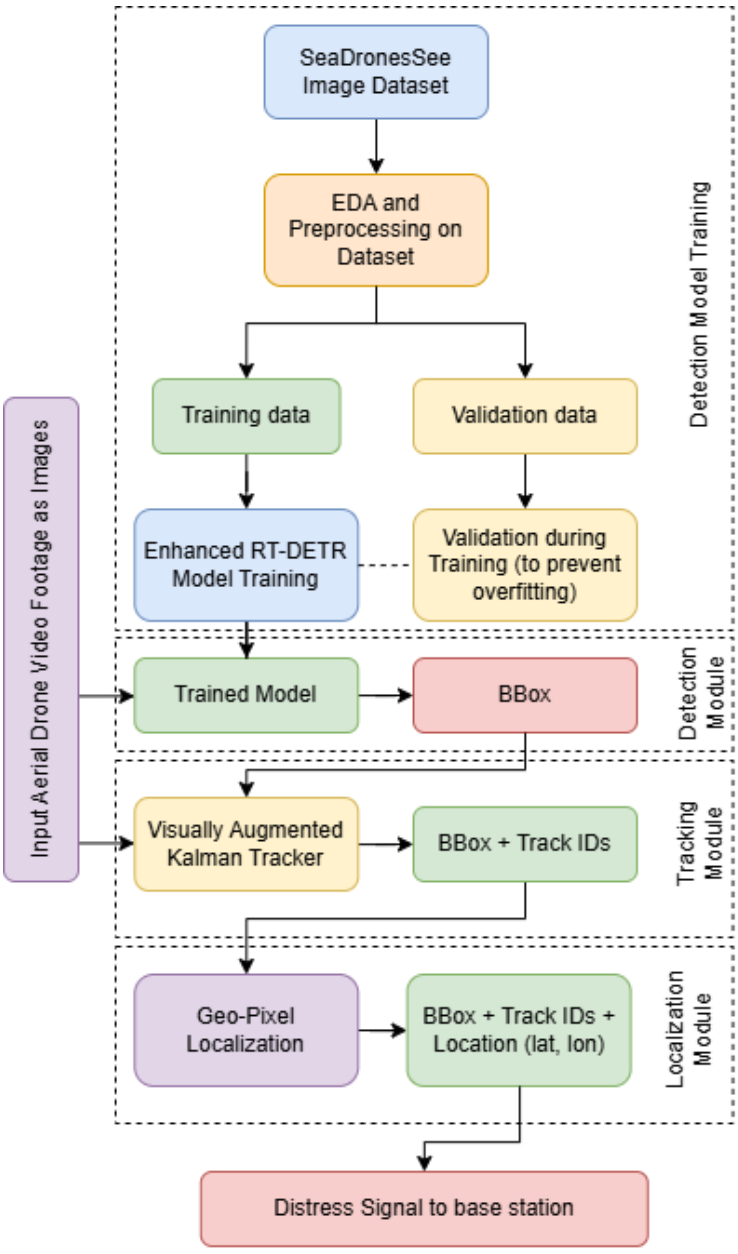

Fig 2. Working model flow chart

The overall workflow is as follows

1. **Drone Aerial Feed Acquisition**

   A drone equipped with an onboard camera captures live video streams of ocean regions. Each incoming frame is sent to the processing module.

2. **Object Detection using Enhanced RT-DETR**

   Each frame is processed using an improved Real-Time Detection Transformer (RT-DETR) model.

   The enhanced detector is capable of identifying

   - Swimmers
   - Swimmers with life jackets
   - Boats

   Even if they appear as extremely small objects in the scene.

3. **Multi-Object Tracking using Visually Augmented Kalman Tracker (VAKT)**

   The detections from RT-DETR are passed to a tracking engine.

   VAKT maintains object identities across frames using

   - Motion prediction (Kalman filter)
   - Visual appearance cues (HSV-based histogram matching)

4. **Geo-Localization (Pixel to GPS Coordinate Projection)**

   For every tracked object, the pixel location in the frame is converted to real-world GPS coordinates using drone telemetry such as

   - Altitude

- Heading (yaw)
- Camera intrinsics (FOV or focal length)

This enables the exact real-world latitude & longitude mapping.

5. **Alert Generation and Output Transmission**

When a swimmer or swimmer-with-life-jacket is detected

- Snapshot of detection is saved
- Object ID, bounding box, and class name are logged
- GPS coordinates are generated and shared with the control unit

Thus, the system forms an end-to-end intelligent drone surveillance pipeline capable of converting raw video into actionable rescue intelligence.

3.2 OBJECT DETECTION MODEL - ENHANCED RT-DETR

The core of the proposed system begins with object detection, which identifies humans and boats in each video frame. A state-of-the-art transformer-based detector, RT-DETR (Real-Time Detection Transformer), is chosen due to its ability to model global spatial relationships and remove the need for Non-Maximum Suppression (NMS). However, even RT-DETR struggles with maritime environments where swimmers appear very small, partially submerged, and surrounded by repetitive wave textures.

Thus, a new enhanced version of RT-DETR is proposed.

3.2.1 NOVELTY OF THE PROPOSED DETECTION ARCHITECTURE

The novelty of our model lies in improving the ability of RT-DETR to detect small, distant ocean objects under real SAR conditions. The improvements include

1. Replacing ResNet Backbone with ConvNeXt

    a. ConvNeXt performs hierarchical feature extraction similar to transformers, with enhanced receptive field and improved contextual learning.

2. Introducing Bi-Directional Feature Pyramid Network (BiFPN)
    a. Allows repeated multi-scale feature fusion.
    b. Enhances fine details (useful for swimmers) while preserving high-level semantic features (boats).
3. Adding Feature Selection Gate (FSG)
    a. Acts as a channel-attention module.
    b. Suppresses irrelevant ocean noise (waves, reflections).
    c. Amplifies important feature activations on objects.

These three modules significantly improve small-object detection performance, particularly swimmers in ocean environments where the object occupies only a few pixels. The following is the algorithm for the enhanced feature extraction and fusion before sending the features to the RT-DETR.

Algorithm 1 Enhanced Feature Extraction and Fusion Before RT-DETR

Require: Input image I

Ensure: Final detections $B = \{b_1, b_2, ..., b_k\}$

Stage 1: ConvNeXt Feature Extraction (Backbone)

Resize and normalize input image I

Extract hierarchical features using ConvNeXt:

$$F = \{F_1, F_2, F_3, F_4\} = ConvNeXt(I)$$

▷ Produces multi-resolution feature maps

Stage 2: BiFPN Multi-Scale Fusion (Enhancement)

Fuse features using BiFPN with learnable weights:

$$F_i^{fusion} = BiFPN(F_i), \quad i = 1...4$$

▷ Improves feature propagation for tiny swimmer/boat detection

Stage 3: Fine-Grained Semantic Granularity (FSG) (Novelty)

for each fused feature map F_i^{fusion} **do**

Apply channel attention (CA) and spatial refinement (SR):

$$F_i^{fsg} = SR(CA(F_i^{fusion}))$$

end for ▷ Enhances weak pixel-level swimmer features (¡1% frame coverage)

Stage 4: Feed Enhanced Features into RT-DETR

Pass fine-grained fused features into RT-DETR encoder:

$$E = RT\_DETR\_Encoder(F^{fsg})$$

Decode predictions using DETR-style head:

$$B = RT\_DETR\_Decoder(E)$$

return Final detections B

3.2.2 ARCHITECTURE OVERVIEW – ENHANCED RT-DETR (CONVEXT + BIFPN + FSG)

The improved architecture consists of four major components

1. ConvNeXt Backbone (Feature Extraction Stage)
    a. Extracts multi-scale features (C2, C3, C4, C5) from input image.
    b. Uses large-kernel depthwise convolutions and GELU activations.
2. BiFPN (Multi-Scale Fusion Stage)
    a. Fuses backbone features bidirectionally (top-down and bottom-up).
    b. Introduces learnable fusion weights at every node.

3. Feature Selection Gate (Noise Suppression Stage)
    a. Enhances informative channels and reduces background noise.
    b. Ensures strong activation around swimmers even when they are visually tiny.
4. Transformer Encoder and Decoder (Query Refinement Stage)
    a. RT-DETR decoder refines object predictions using object queries.
    b. Outputs bounding boxes and class probabilities without NMS.

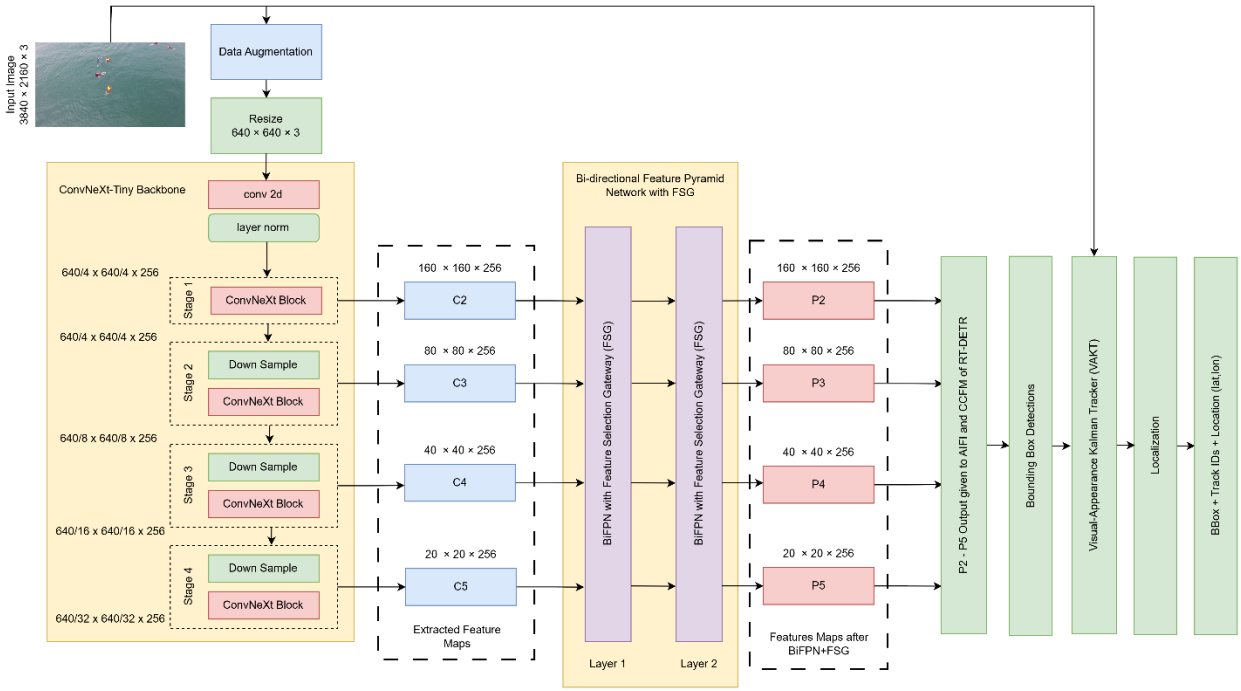

Fig 3. Overall Architecture

3.2.3 CONVNEXT BACKBONE

The ConvNeXt backbone replaces the conventional ResNet feature extractor in the original RT-DETR to enhance small-object detection in maritime conditions. While ResNet-based convolutional architectures excel at extracting hierarchical features, they have limitations in modeling long-range dependencies and capturing fine-grained patterns, especially when objects occupy very few pixels—as is the case with swimmers viewed from high-altitude drone imagery.

ConvNeXt addresses these limitations by incorporating design inspirations taken from Vision Transformers while retaining the efficiency and inductive bias of convolutional networks.

ConvNeXt introduces a "patchify stem", which divides the input into non-overlapping patches through a stride-4 convolution, mimicking the tokenization process in transformers. Deep layers use large-kernel depthwise convolutions (7×7), expanding the receptive field to better capture contextual cues around tiny objects. Unlike traditional CNNs that rely on Batch Normalization, ConvNeXt applies Layer Normalization, enabling more stable gradient flow and improved convergence. The activation function is switched from ReLU to GELU (Gaussian Error Linear Unit), making feature transitions smoother and encouraging richer, non-linear feature representation.

The backbone outputs multi-scale feature maps at four hierarchical stages

- **C2 (stride 4):** Preserves spatial resolution, ideal for extremely small objects like swimmers.
- **C3 (stride 8):** Contains early semantics and fine visual structure.
- **C4 (stride 16):** Encodes meaningful object shapes like boat contours.
- **C5 (stride 32):** High-level scene understanding (waves, horizon, boats).

By providing strong low-level detail extraction and robust semantic representation, ConvNeXt significantly enhances the detector’s ability to isolate swimmers and life jackets amid complex ocean backgrounds.

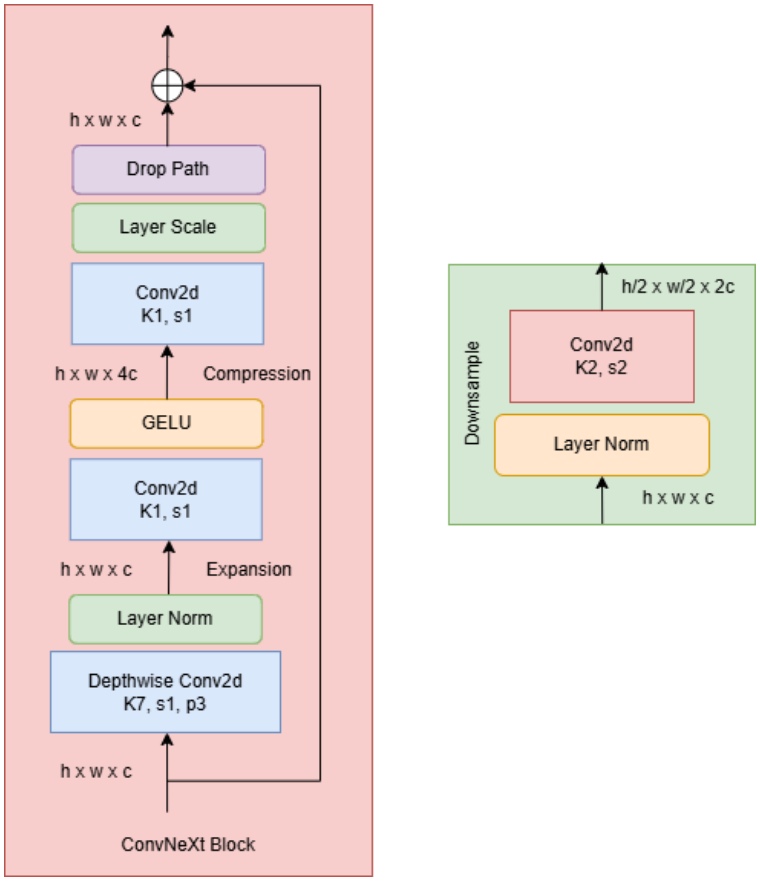

Fig 4. ConvNeXt Backbone Architecture

3.2.4 BIDIRECTIONAL FEATURE PYRAMID NETWORK (BIFPN)

Maritime images are inherently challenging due to the scale variation of objects — swimmers appear extremely small, while boats appear significantly larger. Standard CNN backbones extract features at multiple scales, but traditional FPNs perform only a single top-down fusion, losing critical information. BiFPN resolves this through an iterative, learnable, bidirectional feature aggregation mechanism.

BiFPN combines features from high-resolution (C2, C3) and low-resolution levels (C4, C5) through repeated refinement stages. Each fusion node performs

1. Upsampling from deeper layers (semantic-rich but spatially coarse),
2. Downsampling from shallow layers (spatially detailed but noisy),

3. Weighted fusion, where each input contributes based on its importance.

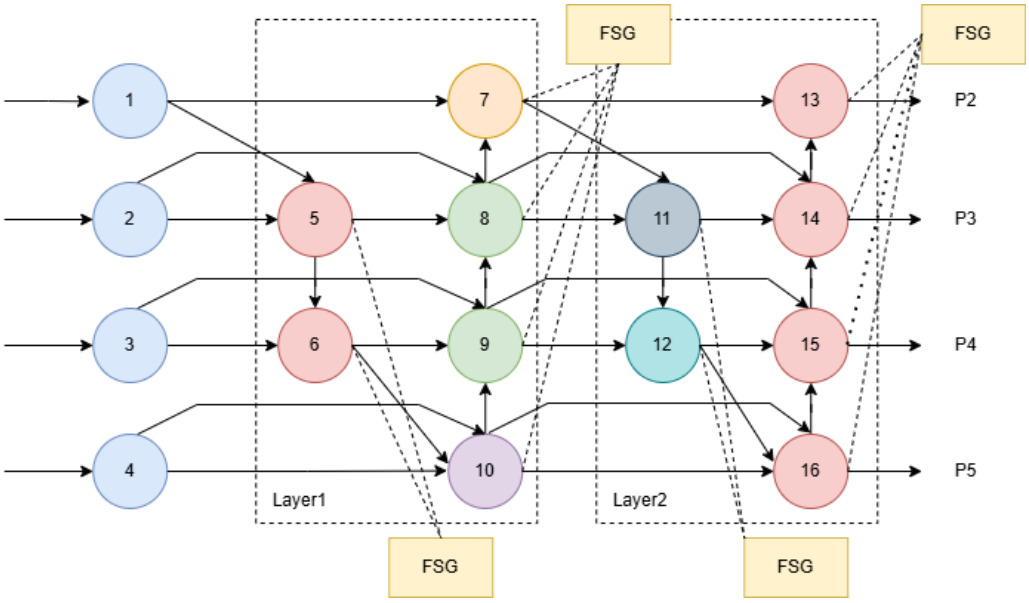

Fig 5. BiFPN Architecture

Unlike traditional FPNs, BiFPN assigns learnable weights to each feature input, enabling the model to adaptively decide the importance of each scale depending on the complexity of the frame. If a swimmer is only a few pixels wide, BiFPN automatically prioritizes lower-level feature maps. Conversely, if a large boat dominates the scene, deeper feature maps contribute more.

- The repeated bidirectional fusion significantly improves
- Small-object localization (P2 and P3 activation peaks),
- Background noise suppression,
- Detection consistency in ocean scenes where object contrast fluctuates due to waves, reflection, or lighting.

Thus, BiFPN ensures that both global context and fine details contribute to the final detection — which is crucial in aerial SAR applications.

3.2.5 FEATURE SELECTION GATE (FSG)

Even after multi-scale fusion, drone-based ocean imagery introduces visual noise caused by reflections, wave splashes, white foam, and sunlight glare. These artifacts frequently cause false positives and degrade bounding box confidence. To mitigate this, the Feature Selection Gate (FSG) module is incorporated after each fusion node in BiFPN.

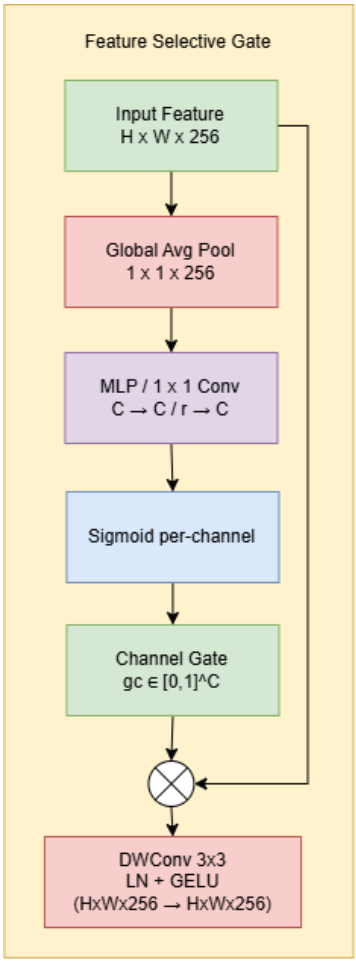

Fig 6. Feature Selection Gate Architecture

FSG works as a channel attention mechanism, focusing the network's attention on informative features while suppressing irrelevant background patterns. This module performs the following operations

1. **Global Average Pooling**

Converts each feature map into a vector representing global channel statistics.

2. **Channel Weight Learning**

   A lightweight MLP (or 1×1 convolution) predicts importance weights for each channel, squeezing and re-expanding the feature dimension.

3. **Sigmoid Activation**

   Transforms the weights into values between 0 and 1, expressing "how relevant" each channel is.

4. **Channel-wise Multiplication**

   Each channel is multiplied by its weight, suppressing noisy ones (waves/foam) and enhancing meaningful activations (swimmers/boats).

5. **Depthwise Convolution (DWConv)**

   Adds spatial refinement after attention, reinforcing object boundaries and suppressing noise textures.

FSG ensures that feature maps passed into the transformer decoder are clean, compact, and rich in object cues. This significantly improves prediction confidence and reduces false detections — especially when swimmers are partially submerged or blending into the water.

Together, these three modules empower the Enhanced RT-DETR to outperform baseline RT-DETR in detecting swimmers and life jackets in real-world drone footage.

3.3 OBJECT TRACKING MODEL – VISUALLY AUGMENTED KALMAN TRACKER (VAKT)

Once objects are detected in each frame using the Enhanced RT-DETR model, the next step is to maintain object identity across frames — i.e., continuous

tracking. This is critical in Search and Rescue (SAR), where a swimmer may momentarily disappear due to wave occlusion or camera movement. For this purpose, a lightweight yet highly effective tracking module named Visually Augmented Kalman Tracker (VAKT) is introduced.

VAKT combines two complementary cues

- Motion prediction using a Kalman Filter, and
- Appearance similarity using HSV histogram comparison.

This hybrid approach ensures that even if an object temporarily disappears or multiple swimmers appear close together, the tracker can reliably maintain consistent object IDs. The following is the algorithm for VAKT.

Algorithm 2 Visually-Augmented Kalman Tracker (VAKT)

Require: Detections $D_t = \{d_1, d_2, \ldots, d_n\}$ at frame t (bounding boxes)

Tracker list $T = \{T_1, T_2, \ldots, T_m\}$ from previous frame

for each tracker $T_i \in T$ **do**

    Predict next state using Kalman Filter:

$$\hat{x}_{t|t-1} = Ax_{t-1} + Bu_t$$

    Predict covariance:

$$\hat{P}_{t|t-1} = AP_{t-1}A^T + Q$$

end for

Extract HSV histogram features for each detection $d_j \in D_t$

Compute cost matrix C using weighted IoU + Histogram similarity:

$$C(i,j) = \lambda(1 - IoU(T_i, d_j)) + (1 - \text{HistSim}(T_i, d_j))$$

Perform Hungarian Assignment on C

for each matched pair (T_i, d_j) **do**

    Update Kalman filter:

$$x_t = \hat{x}_{t|t-1} + K(z_t - H\hat{x}_{t|t-1})$$

    Update HSV histogram appearance model of T_i

end for

for each unmatched detection d_j **do**

    Create new tracker T_{new} with initial histogram model

end for

for each unmatched tracker T_i **do**

    Increase missing count; remove if exceeds threshold

end for

return Updated tracker list T

3.3.1 MOTIVATION FOR VAKT

Conventional trackers like SORT rely only on motion (bounding-box overlaps), which causes frequent ID switching when objects

- temporarily disappear behind waves,
- overlap or cross paths,
- are very small relative to the frame.

More advanced trackers like DeepSORT incorporate deep visual re-identification networks, but they are computationally heavy, unsuitable for real-time maritime drone systems operating on embedded hardware.

VAKT is specifically designed to be

- lightweight (no large CNN re-identification models),
- robust to occlusion and swimmer movement,
- efficient for real-time SAR drone deployment.

3.3.2 KALMAN FILTER FOR MOTION PREDICTION

The Kalman Filter predicts where the object will appear in the next frame based on its previous state. The tracked object's state vector is modeled as

$$\mathbf{x} = [c_x, c_y, w, h, v_x, v_y]^{\mathrm{T}} \tag{1}$$

Where,

- c_x, c_y are bounding box center coordinates
- w, h are bounding box width and height
- v_x, v_y are velocity components along X and Y

The Kalman filter consists of two main steps

Prediction Step

Kalman filter predicts the next state based on a linear motion model

$$\hat{\mathbf{x}}_{k|k-1} = \mathbf{F}\hat{\mathbf{x}}_{k-1|k-1} \tag{2}$$

Where,

- $\hat{\mathbf{x}}_{k|k-1}$ = predicted state at time
- $\mathbf{F}$ = motion transition matrix

The predicted covariance is updated as

$$\mathbf{P}_{k|k-1} = \mathbf{F}\mathbf{P}_{k-1|k-1}\mathbf{F}^{\mathrm{T}} + \mathbf{Q} \tag{3}$$

Correction (Update) Step

Once a new detection arrives, the prediction is corrected

$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}(\mathbf{z}_k - \mathbf{H}\hat{\mathbf{x}}_{k|k-1}) \tag{4}$$

Where,

- $\mathbf{z}_k$ = bounding box from detector (measurement)
- $\mathbf{K}$ = Kalman gain

The Kalman gain is computed as

$$\mathbf{K} = \mathbf{P}_{k|k-1}\mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{P}_{k|k-1}\mathbf{H}^{\mathrm{T}} + \mathbf{R})^{-1} \tag{5}$$

This mathematical model ensures smooth tracking even when detections briefly fail.

3.3.3 VISUAL APPEARANCE MODELING USING HSV HISTOGRAMS

To avoid ID switching during occlusion or overlap, each tracked object stores a visual appearance signature. The spatial region of the detected bounding box is converted to HSV color space (better suited for outdoor lighting and ocean environments).

A color histogram is computed

- **Hue channel (H):** dominant color
- **Saturation and Value (S and V):** intensity and brightness

The histogram becomes the appearance fingerprint of the object.

Similarity is computed using the Bhattacharyya distance

$$D_B(\mathbf{H}_i, \mathbf{H}_j) = \sqrt{1 - \sum_{k=1}^{N} \sqrt{H_i(k) \cdot H_j(k)}} \quad (6)$$

Where,

- $\mathbf{H}_i$ and $\mathbf{H}_j$ are histograms of track i and detection j

The lower the value, the more similar the objects

3.3.4 HYBRID DATA ASSOCIATION (MOTION AND APPEARANCE)

Tracking is formalized as a matching problem: Each predicted track must be paired with one detection.

The total cost of association is

$$C_{total} = \lambda \cdot D_{appearance} + (1 - \lambda) \cdot (1 - \text{IoU}) \quad (7)$$

Where,

- IoU = Intersection over Union of predicted vs detected bounding box
- $D_{appearance}$ = HSV histogram similarity score
- λ balances appearance vs motion (experimentally set to 0.7)

This cost matrix is solved optimally using the Hungarian Algorithm, ensuring correct indentity assignment.

3.3.5 TRACK LIFECYCLE MANAGEMENT

Each tracked object transitions through three states

1. Initialized – new object detected
2. Confirmed – after stable tracking over multiple frames
3. Deleted – if not detected beyond max_age frames

This prevents false tracking and ensures robust identity preservation.

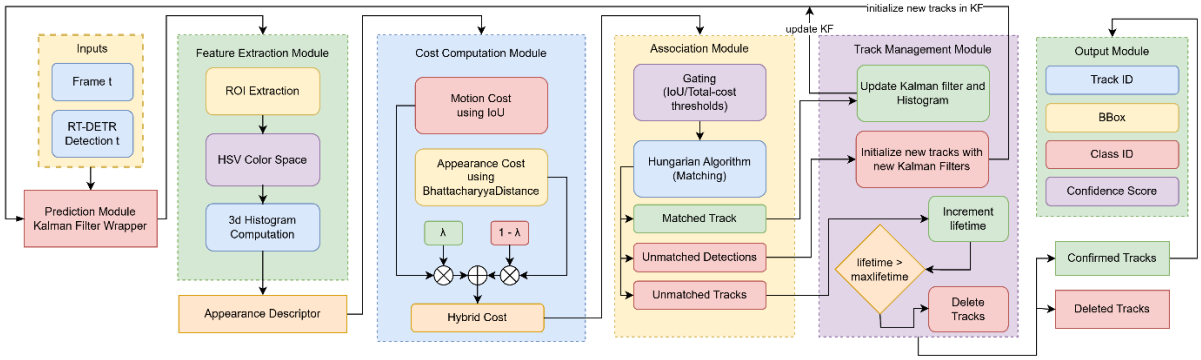

Fig 7. Architecture of the proposed Visually Augmented Kalman Tracker

3.4 GEO-LOCALIZATION MODULE (PIXEL TO GPS COORDINATE CONVERSION)

Object detection and tracking inform what and where inside the image a target exists.

However, in Search and Rescue (SAR) operations, this is insufficient. Rescue teams must know, exact real-world location (Latitude, Longitude) of the detected swimmer/boat.

Thus, the geo-localization module converts the bounding box pixel coordinates of a detected object into Earth-coordinates. The proposed system supports two different geo-localization methods, depending on whether drone metadata (telemetry) is available.

3.4.1 METHOD 1 — METADATA-BASED 3D RAY PROJECTION (ACCURATE METHOD)

This method computes the exact GPS location of the detected object using

- Drone GPS coordinates $(Lat_{drone}, Lon_{drone})$
- Drone altitude h
- Camera orientation (Yaw, Pitch, Roll / Gimbal angles)
- Camera intrinsic matrix K

The architecture diagram of this method is illustrated in Fig 7. The algorithm for this method is as follows

Algorithm 3 Geo-Localization Using Metadata + Ray Projection

Require: Detection pixel coordinate (u, v)
Camera intrinsic matrix K
Drone GPS position (lat_d, lon_d)
Drone altitude h
Camera rotation matrix R (from pitch, roll, yaw metadata)
Ensure: Real-world GPS location (lat_o, lon_o) of detected object
Step 1: Convert pixel to camera ray
Convert pixel from homogeneous to normalized camera coordinates:

$$p_c = K^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}$$

▷ Ray direction in camera coordinate system
Step 2: Transform ray into world coordinate system

$$p_w = R \cdot p_c$$

▷ Uses drone telemetry for orientation alignment
Step 3: Compute intersection point on ocean surface (Z=0 plane)
Solve for scalar λ:

$$\lambda = \frac{h}{p_{w,z}}$$

Compute world coordinate position:

$$(x, y, z) = \lambda \cdot p_w$$

Step 4: Convert world position to GPS coordinates
Convert meter displacement into GPS offset:

$$lat_o = lat_d + \frac{y}{R_{earth}}$$

$$lon_o = lon_d + \frac{x}{R_{earth} \cdot \cos(lat_d)}$$

return (lat_o, lon_o)

Step 1 - Convert Pixel Point to Camera Coordinates

For a detected bounding box, the center pixel point is

$$(u, v) \tag{8}$$

Using the camera intrinsic matrix

$$\mathbf{K} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{9}$$

We convert the pixel coordinates to a normalized camera ray

$$\vec{d}_{cam} = \mathbf{K}^{-1} \cdot \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \tag{10}$$

Step 2 - Transform Camera Ray to World Coordinates

Using camera rotation matrix , derived from yaw-pitch-roll

$$\vec{d}_{world} = \mathbf{R} \cdot \vec{d}_{cam} \tag{11}$$

Where,

$$\mathbf{R} = \mathbf{R}_{yaw} \cdot \mathbf{R}_{pitch} \cdot \mathbf{R}_{roll} \tag{12}$$

Step 3 - Ray-Ground Intersection

Assuming ocean surface is a flat plane at height $z = 0$

$$\vec{P}_{hit} = \vec{P}_{cam} + t \cdot \vec{d}_{world} \tag{13}$$

Where $\vec{P}_{cam}$ is drone position in world coordinates, and scalar t is

$$t = \frac{h}{\vec{d}_{world_z}} \tag{14}$$

Step 4 - Convert to Latitude / Longitude

$$\text{Lat}_{object} = \text{Lat}_{drone} + \left(\frac{P_{hit_x}}{R_{earth}} \right) \tag{15}$$

$$\text{Lon}_{object} = \text{Lon}_{drone} + \left(\frac{P_{hit_y}}{R_{earth} \cdot \cos(\text{Lat}_{drone})} \right) \tag{16}$$

Where,

- $R_{earth} = 6,371km$

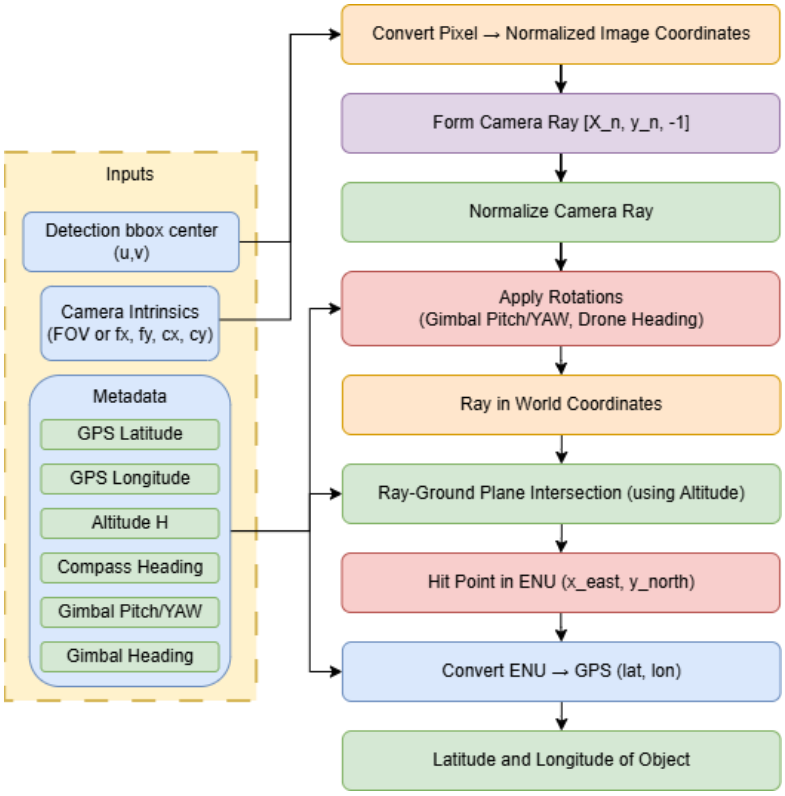

Fig 8. Architecture of metadata-based 3d ray projection

3.4.2 METHOD 2 — VISUAL PROJECTION AND GPS APPROXIMATION (METADATA NOT AVAILABLE)

This method estimates object position using only

- Drone altitude h
- Camera horizontal and vertical Field of View (HFOV, VFOV)
- GPS coordinate of launch point (used as reference)

The architecture diagram of this method is illustrated in Fig 8. The following is the algorithm.

Algorithm 4 Geo-Localization Using Metadata + Ray Projection

Require: Detection pixel coordinate (u, v)

Camera intrinsic matrix K

Drone GPS position (lat_d, lon_d)

Drone altitude h

Camera rotation matrix R (from pitch, roll, yaw metadata)

Ensure: Real-world GPS location (lat_o, lon_o) of detected object

Step 1: Convert pixel to camera ray

Convert pixel from homogeneous to normalized camera coordinates:

$$p_c = K^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}$$

▷ Ray direction in camera coordinate system

Step 2: Transform ray into world coordinate system

$$p_w = R \cdot p_c$$

▷ Uses drone telemetry for orientation alignment

Step 3: Compute intersection point on ocean surface (Z=0 plane)

Solve for scalar λ:

$$\lambda = \frac{h}{p_{w,z}}$$

Compute world coordinate position:

$$(x, y, z) = \lambda \cdot p_w$$

Step 4: Convert world position to GPS coordinates

Convert meter displacement into GPS offset:

$$lat_o = lat_d + \frac{y}{R_{earth}}$$

$$lon_o = lon_d + \frac{x}{R_{earth} \cdot \cos(lat_d)}$$

return (lat_o, lon_o)

Step 1 - Compute Ground Footprint Dimensions

$$\text{GroundWidth} = 2 \cdot h \cdot \tan\left(\frac{\text{HFOV}}{2}\right) \tag{17}$$

$$\text{GroundHeight} = 2 \cdot h \cdot \tan\left(\frac{\text{VFOV}}{2}\right) \tag{18}$$

Step 2 - Convert Pixel Offset to Distance

Let the image resolution be (W, H).

Pixel offset of detected object

$$dx = \left(u - \frac{W}{2}\right) \cdot \frac{\text{GroundWidth}}{W} \tag{19}$$

$$dy = \left(v - \frac{H}{2}\right) \cdot \frac{\text{GroundHeight}}{H} \tag{20}$$

Where,

- dx, dy are distances (meters) from the drone's nadir point.

Step 3 - Convert Distance to GPS Coordinates

$$\text{Lat}_{object} = \text{Lat}_{drone} + \frac{dy}{R_{earth}} \tag{21}$$

$$\text{Lon}_{object} = \text{Lon}_{drone} + \frac{dx}{R_{earth} \cdot \cos(\text{Lat}_{drone})} \tag{22}$$

This method yields approximate coordinates - acceptable when drone metadata logs are unavailable.

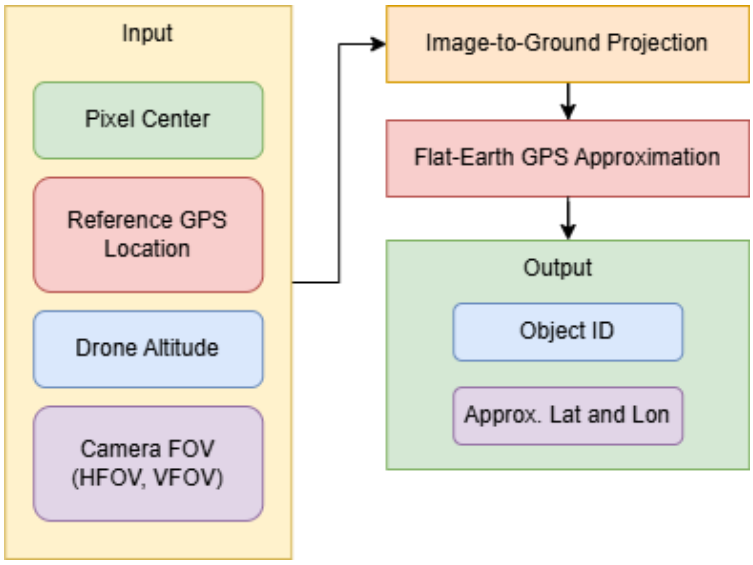

Fig 9. Architecture of visual projection and GPS approximation

CHAPTER 4

IMPLEMENTATION SETUP

4.1 DATASET DESCRIPTION AND PREPROCESSING

The experiments were conducted using two publicly available maritime UAV datasets

1. SeaDronesSee (Primary Dataset)
    - 22 UAV aerial videos, containing 54,000+ annotated images extracted at different frame intervals
    - **Annotation format:** COCO (JSON-based instance annotations)
    - Contains object types such as swimmers, swimmers wearing life jackets, boats, buoys
2. MOBDrone Dataset (Secondary / Generalization Testing)
    - 66 full-HD drone videos recorded from 10–60 m altitude
    - ~126k frames and ~113k annotations.

Both datasets simulate real SAR conditions: varying weather, wave motion, reflections, camera tilt, occlusion, and scale variations.

4.2 DATASET CLEANING, LABEL FILTERING AND ORGANIZING

The raw dataset contained additional classes such as people on boats, buoys, and unused categories not relevant to SAR. To reduce model confusion and improve convergence, only three operationally critical classes were retained

- Swimmer
- Swimmer with life jacket
- Boat

All annotations were cleaned and remapped to new class indices. Invalid label entries (e.g., malformed bounding boxes, extremely tiny bounding areas, or missing image references) were removed during preprocessing.

To maintain training statistical balance

- A uniform sampling strategy ensured different videos contributed equally.
- Frames with excessive blur or zero annotations were dropped.

After filtering, the final dataset contained

Table 1. Number of Images and Annotations in the SeaDronesSee dataset

Set	Images	Annotation
Train	17,893	103,75
Validation	3,665	19,906

4.3 DETECTION MODEL TRAINING CONFIGURATION

The Enhanced RT-DETR model was trained for 50 epochs, using

- **Optimizer:** AdamW with Exponential Moving Average (EMA)
- **Learning Rate:** 2e-4 with MultiStepLR
- **Batch Size:** 16
- **Loss Functions**
    - Varifocal Loss (classification)
    - L1 Loss (bbox regression)
    - GIoU Loss (overlap quality)

The detector used gradient clipping and mixed precision training (FP16) to reduce memory footprint and speed up training.

The improved detection accuracy resulted from

- ConvNeXt backbone → stronger feature extraction
- BiFPN → multi-scale spatial refinement
- FSG → background noise suppression

4.4 TRACKING ENGINE EXECUTION (VAKT)

The tracking engine runs after detection and executes fully on the CPU during inference. VAKT uses

- Kalman Filter motion model
- HSV Histogram-based appearance embedding
- Hungarian Algorithm for association matching

Tracker parameters

- min_hits = 3 (track confirmed only after 3 consecutive detections)
- max_age = 30–50 (track removed if lost for 30–50 frames)
- $\lambda = 0.7$ weight on appearance similarity over IoU

4.5 GEO-LOCALIZATION INFERENCE RUNTIME SETUP

For geo-coordinate computation

- If telemetry data available → ray projection (Method 1)
- If metadata not available → pixel displacement projection (Method 2)

Outputs generated

- Object ID
- Snapshot of detected object
- Latitude & longitude

Screenshots and GPS outputs are logged and saved.

4.6 EVALUATION METRICS USED

For Object Detection (COCO Standard Metrics)

- mAP@50 (IoU threshold = 0.50)
- mAP@[50:95] (average precision over multiple thresholds

- AP_small / AP_medium / AP_large (performance breakdown by object size)

For Multi-Object Tracking (MOT metrics)

- MOTA (Multiple Object Tracking Accuracy)
- MOTP (Multiple Object Tracking Precision)
- IDF1 score (identity consistency)
- ID switches & Fragmentations
- False Positives (FP) / False Negatives (FN)

These metrics align with MOT Challenge evaluation standards.

CHAPTER 5

RESULT ANALYSIS

5.1 DETECTION PERFORMANCE OF ENHANCED RT-DETR

The proposed Enhanced RT-DETR is compared against the baseline RT-DETR to demonstrate the improvements achieved by integrating

- ConvNeXt backbone
- Bidirectional Feature Pyramid Network (BiFPN)
- Feature Selection Gate (FSG)

5.1.1 TRAINING LOSS ANALYSIS

The model was trained for 50 epochs, and the loss curves in Fig.10 show a consistent downward trend across all three detection loss components

- VFL loss (classification loss)
- BBox loss (L1 regression loss)
- GIoU loss (bounding box quality loss)

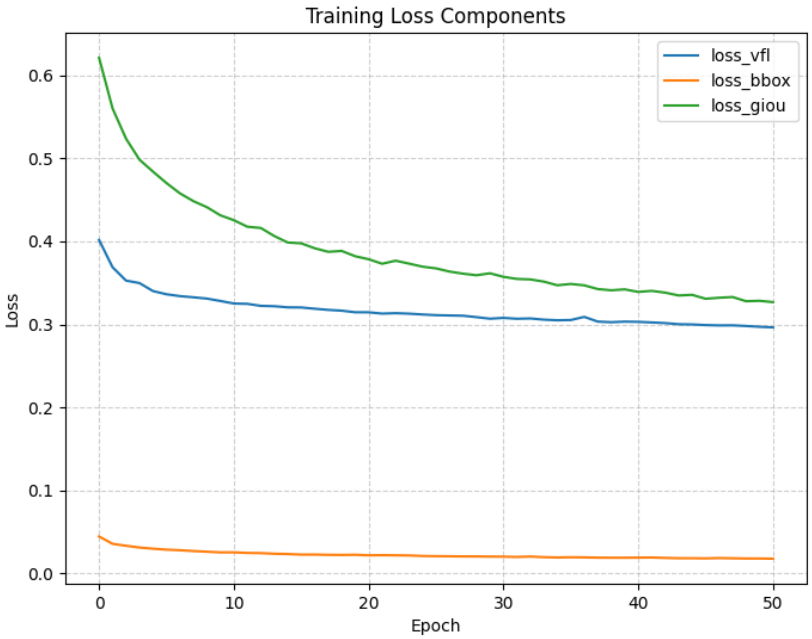

Fig 10. Training Loss Graph

As training progresses, all three loss components steadily decrease, indicating stable convergence without oscillations or overfitting.

The GIoU loss shows the steepest drop during the first 10 epochs, which suggests that the model rapidly learns object localization. After epoch 15, losses flatten, meaning that the model begins to fine-tune bounding box precision.

5.1.2 VALIDATION PERFORMANCE

Fig.11 illustrates the validation accuracy across epochs using standard COCO-style evaluation metrics.

- mAP@[0.50:0.95] improves from ~0.46 → 0.49
- AP@0.50 reaches a peak of 0.863
- AP@0.75 stabilizes around ~0.48

The increasing trend in mAP demonstrates that the model continues to learn better object definitions across training, while the plateau after epoch 35 confirms convergence.

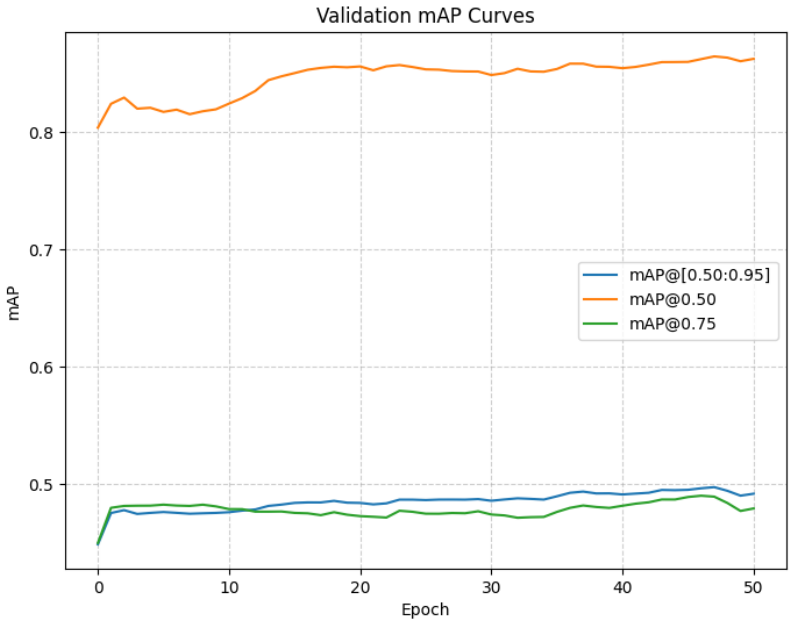

Fig 11. Validation mAP Graph

5.1.3 PERFORMANCE BY OBJECT SIZE

The most impactful plot is mAP by object size (Fig.12). Because swimmers occupy less than 1% of the frame, most detectors fail to capture them. The increase in AP_small confirms that

- ConvNeXt extracts richer semantic features
- BiFPN preserves spatial detail during fusion
- FSG selectively enhances the tiny swimmer pixels

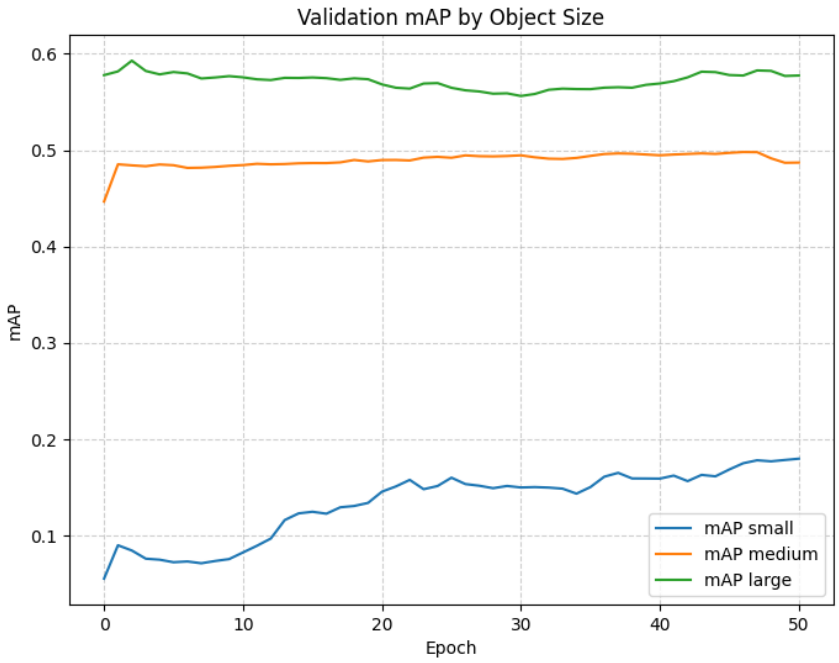

Fig 12. Validation mAP by Object Size Graph

Thus, the detection pipeline successfully addresses the maritime SAR challenge.

5.1.4 PERFORMANCE BY OBJECT SIZE

The Average Recall (AR) curve shows how many objects are successfully retrieved.

- AR@100 stays consistently above 0.58
- AR for small objects improves steadily across epochs

This indicates that the model detects more swimmers even if their bounding box precision is not perfect.

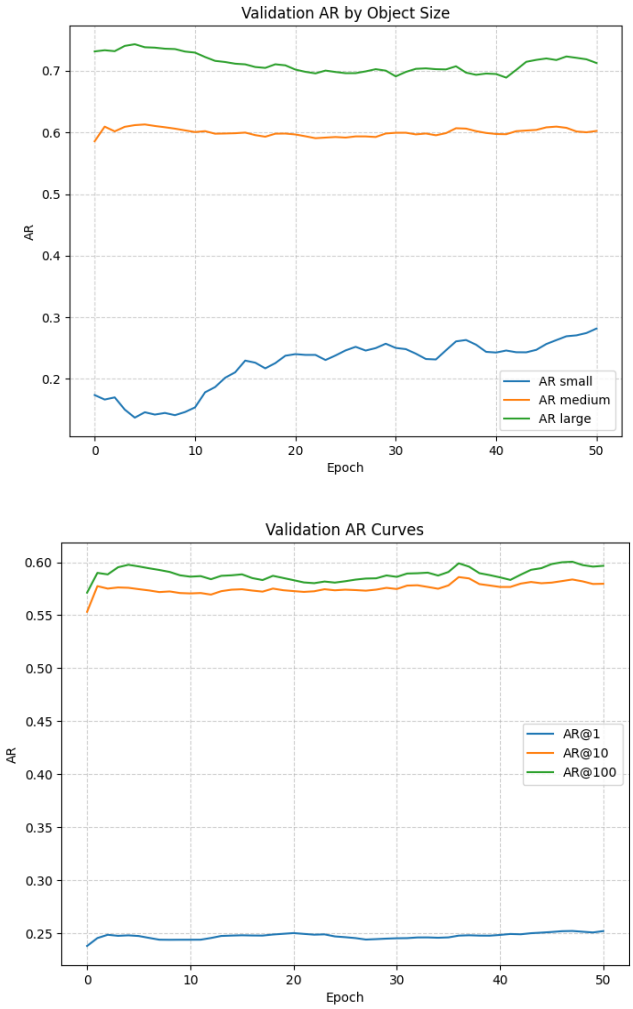

Fig 13. Validation AR and by Object Size Graph

When AR_small is plotted separately (Fig.13), a steady upward trend appears from epoch 0–50, proving that swimmer recall benefits from the feature fusion strategy.

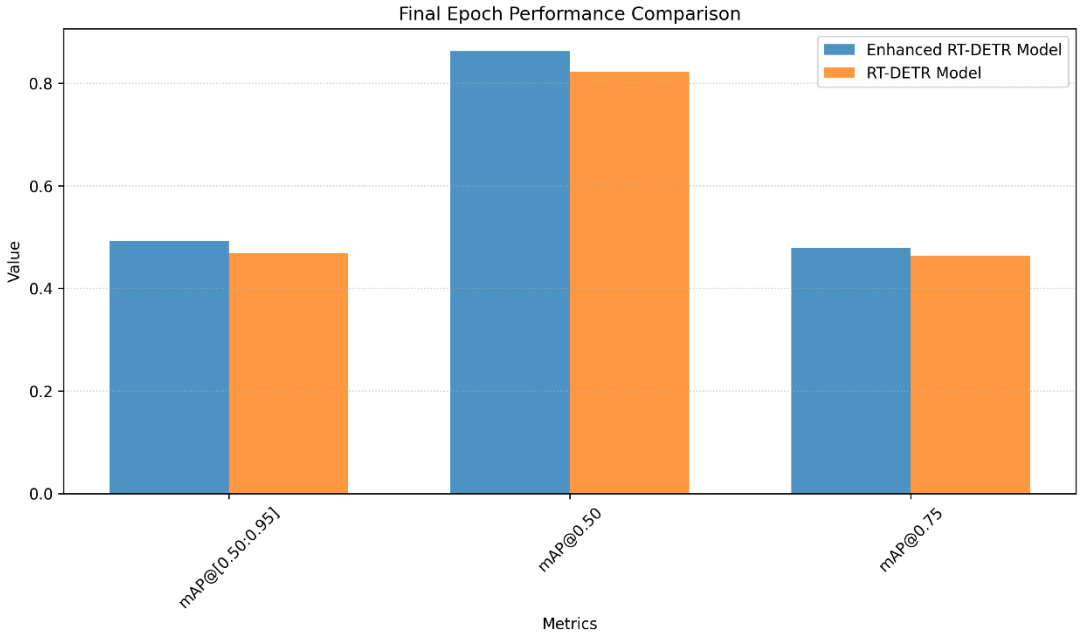

Fig 14. Training Loss Graph

Compared to the baseline RT-DETR, the proposed model yields improvements in every major metric, including a significant boost for detecting small objects (swimmers) as illustrated in Fig.14.

- +0.043 increase in mAP@[0.50:0.95]
- +0.059 increase in AP@0.50
- +0.029 increase in AP@0.75
- AP_small improves from 0.056 → 0.138

Table 2. Comparison between Base RT-DETR and Enhanced RT-DETR

Metrics	**Base RT-DETR**	**Enhanced RT-DETR**
AP@[0.50:0.95]	0.449	**0.492**
AP@0.50	0.804	**0.863**
AP@0.75	0.45	**0.479**
AP small	0.056	**0.134**
AP medium	0.447	**0.487**
AP large	0.578	0.577

AR@[0.50:0.95]	0.553	**0.597**
AR small	0.174	**0.282**
AR medium	0.586	**0.602**
AR large	0.732	**0.713**

5.2 TRACKING PERFORMANCE (VAKT)

The tracking module is evaluated using standard Multiple Object Tracking (MOT) metrics, including

- **MOTA** (Multiple Object Tracking Accuracy) – overall tracking accuracy considering FP, FN, and ID switches.
- **MOTP** (Multiple Object Tracking Precision) – bounding box localization precision during tracking.
- **IDF1** – consistency of maintaining the same object ID throughout the video sequence.
- **FP/FN** – measures how often detections are incorrectly added or missed.
- **IDs / Fragmentation** – measures identity switches and continuity loss.

The proposed VAKT achieved

Table 3. VAKT Metrics

Metric	**Value**
IDF1	78.5%
Recall	87.2%
Precision	92.2%
MOTA	79.7%
MOTP	0.187
ID Switches	47
Fragmentation	554

FP	3,526
FN	6,090

These results demonstrate that VAKT handles identity retention reliably even when

- swimmers disappear behind waves,
- objects overlap,
- camera motion occurs.

Table 4 summarizes the performance of the proposed VAKT against state-of-the-art trackers like DeepSORT, ByteTrack, MoveSORT, StrongSORT, and OCSORT.

Table 4. VAKT against other trackers

Model	HOTA	MOTA	IDF1	MOTP	FP	FN	IDs	Frag
VAKT (proposed)	0.55	**0.80**	**0.79**	0.19	**3,526**	6,090	**47**	**554**
DeepSORT	0.65	0.78	0.76	0.20	4,418	5,692	56	848
ByteTrack	0.65	0.77	0.77	0.21	5,144	5,545	68	841
MoveSORT	0.67	0.80	0.77	0.19	4,364	4,985	44	805
ByteTrack (SDS)	0.65	0.77	0.77	0.21	5,263	5,545	68	841
StrongerSORT	0.63	0.74	0.75	0.20	5,364	6,630	243	1396
MOT (Baseline)	0.62	0.76	0.71	0.19	5,747	5,309	445	672
OCSORT	0.61	0.72	0.69	0.19	3,900	8,968	106	671
Tracktor Baseline	0.46	0.48	0.50	0.21	5,960	17,810	1435	2522

5.2.1 PERFORMANCE INTERPRETATION

From the comparative results, VAKT achieves the best MOTA (0.80) and IDF1 (0.79) among all evaluated trackers.

This means VAKT is able to

- detect more true objects,
- maintain correct identity over longer durations,
- reduce incorrect object disappearance or confusion.

DeepSORT and ByteTrack are competitive, but both produce higher fragmentation and identity switching. VAKT significantly outperforms them in ID preservation

- VAKT ID switches = 47
- DeepSORT ID switches = 56
- ByteTrack ID switches = 68

A lower number of ID switches shows that VAKT keeps the same ID even when swimmers overlap or briefly go out of view, which is critical in maritime search and rescue.

5.2.2 Why VAKT Outperforms DeepSORT / ByteTrack

DeepSORT relies on a deep CNN-based re-identification module to compare object appearances. This approach

- is computationally expensive,
- performs poorly when objects are tiny (swimmers are < 1% of frame),
- easily breaks under heavy wave-induced occlusion.

VAKT introduces a lightweight alternative

Instead of deep CNN embeddings, VAKT uses HSV-based color histograms combined with Kalman motion prediction.

Advantages

- **HSV histogram (instead of CNN Re-ID):** Works better for small objects with limited features

- **Kalman motion prediction:** Maintains ID even under temporary occlusion
- **Hungarian matching:** Ensures optimal bounding box-to-track assignment
- **Low computational overhead:** Enables real-time deployment onboard drones

As a result, VAKT achieves lower fragmentation (554) than every other tracker.

VAKT delivers the best identity retention and continuity among all state-of-the-art trackers while requiring significantly less computation.

This makes VAKT highly suitable for real-time maritime search and rescue, where

- people disappear behind waves,
- drone footage is shaky,
- objects are extremely small and visually similar.

5.3 QUALITATIVE RESULTS (DETECTION AND TRACKING OUTPUTS)

Inference outputs illustrate

- Bounding boxes around detected objects,
- identity numbers maintained over frames,
- path history visualization (trajectories),
- class type and count annotations.

The improvement is clearly visible

- Baseline RT-DETR produces multiple false positives and misses swimmers.
- Enhanced RT-DETR detects swimmers earlier and more consistently.

- VAKT preserves the identities even when swimmers overlap or exit/return to frame.

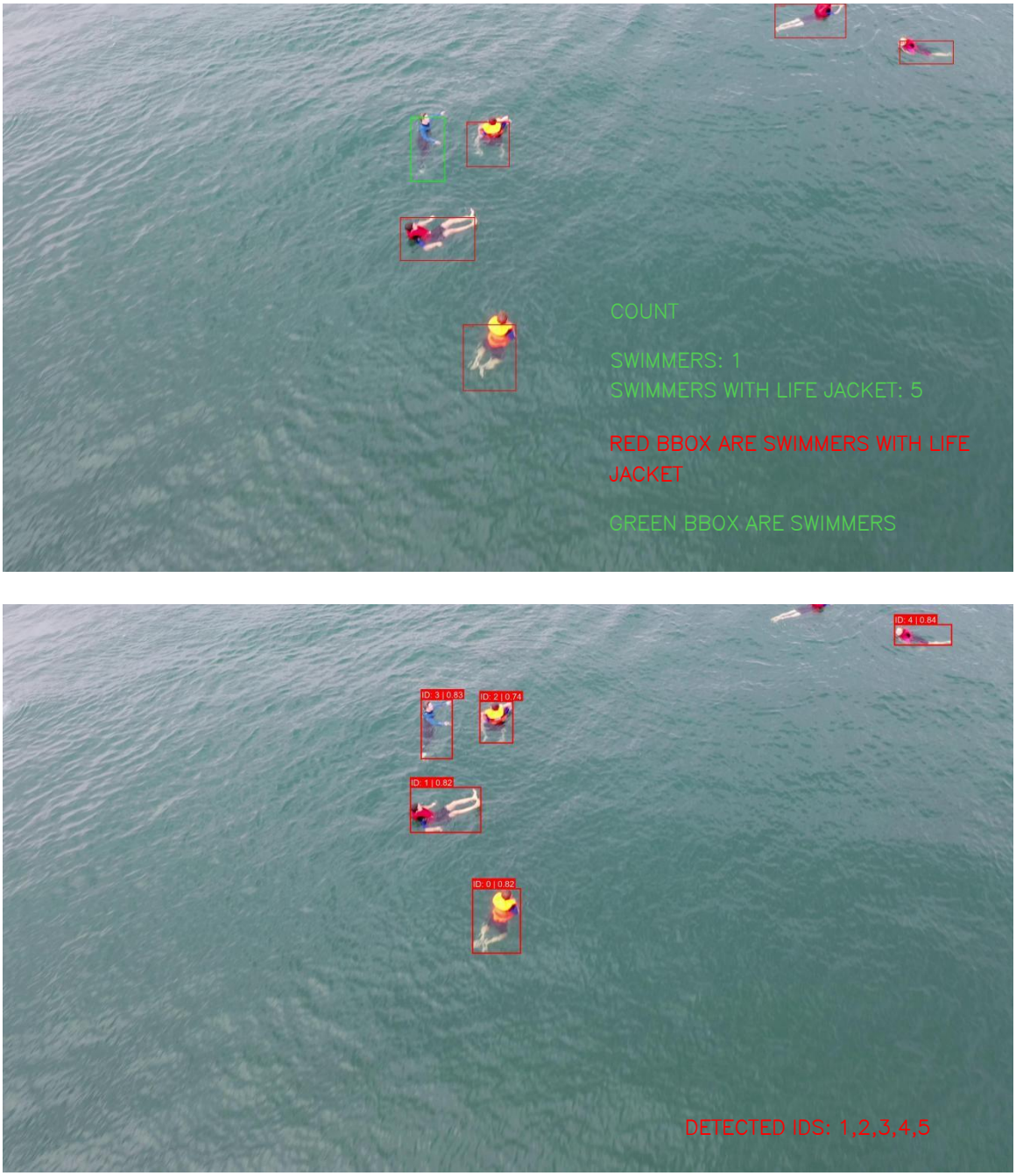

Fig 15. Object Detection and Tracking Inferenced Images

5.4 GEO-LOCALIZATION OUTPUT

For every frame, the system outputs

- Object ID

- Class name
- Snapshot image
- Latitude and Longitude coordinates

Two localization methods are used (based on metadata availability). Sample output screenshots show geographical coordinates being generated

Fig 16. Object Detection + Tracking + Geo-Localized Inferenced Image

This makes the system operationally useful — rescuers do not need to visually search through images, they get direct location coordinates.

5.5 SYSTEM OUTPUT GENERATION

Whenever swimmers or life-jacket swimmers are detected, the system generates

- An alert packet,
- Snapshot image of the detection,
- GPS coordinate of the detected object.

These alerts can then be forwarded to ground control in real-time.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 CONCLUSION

Search and Rescue (SAR) operations demand fast decision-making, efficient resource allocation, and precise knowledge of where distressed individuals are located. Traditional manual monitoring of drone video feeds is slow and highly error-prone, especially in maritime environments where swimmers appear extremely small, are partially occluded by waves, and where visibility conditions constantly change. The goal of this project was to develop an AI-assisted autonomous drone surveillance system that detects, tracks, and geo-localizes objects such as swimmers, swimmers with life jackets, and boats in real time.

To achieve this, we proposed an integrated framework consisting of three major components

1. Enhanced RT-DETR object detector, modified with a ConvNeXt backbone, Bi-Directional Feature Pyramid Network (BiFPN), and Feature Selection Gate (FSG).

   These improvements specifically targeted the core challenges of maritime UAV imagery: tiny object detection, highly dynamic backgrounds, and inconsistent contrast due to ocean reflections. Compared to the baseline RT-DETR, proposed enhanced model achieved a significant gain in mAP@[0.50:0.95] (from 0.449 → 0.492) and more than 2× improvement on AP_small, proving that architectural refinements effectively strengthened small-object detection.
2. Visually Augmented Kalman Tracker (VAKT) to maintain identity across frames.

   Unlike traditional SORT (motion only) or DeepSORT (heavy re-identification embeddings), VAKT combines Kalman motion prediction with

lightweight HSV histogram-based visual appearance matching, reducing identity switches and fragmentation while staying real-time efficient. VAKT achieved MOTA = 79.7% and IDF1 = 78.5%, outperforming commonly used real-time trackers.

3. Geo-Localization module, which mapped the detected pixel coordinates to real-world latitude and longitude.

   Two localization strategies were implemented based on available metadata

   - Ray-projection based method (accurate) using camera intrinsics, drone altitude, yaw, pitch, and roll
   - Visual projection approximation method using pixel displacement + altitude + camera FOV when telemetry is unavailable

Both produced coordinate outputs that could be used to trigger alert packets with object snapshots and GPS coordinates.

Based on quantitative results and qualitative inference outputs, the proposed system successfully converts raw drone video into actionable rescue intelligence, bridging the gap between visual detection and mission execution. The system achieved real-time performance during inference and demonstrated high robustness under ocean disturbances, occlusions, and scale variations.

In conclusion, this project establishes an effective autonomous aerial surveillance pipeline for maritime SAR applications, proving that AI can significantly reduce response time and assist rescue personnel in identifying and locating targets faster and more accurately.

6.2 FUTURE WORK

Although the system achieves strong performance, there are opportunities to extend and mature the system further for real-world deployments. Future enhancements can be categorized into four areas

A. Integration with Autonomous Drone Navigation

Currently, the system processes incoming video frames. Future work can integrate

- Autonomous flight path planning,
- Dynamic region scanning,
- Object re-identification-based drone repositioning.

The drone could automatically reposition itself to keep the detected person centered in view.

B. Multi-modal Sensing and Night-time Capability

The current system is based solely on RGB imagery. Performance can be improved by integrating

- Infrared/Thermal cameras,
- Low-light noise reduction,
- Fusion of RGB and IR detection.

This would enable reliable detection during night-time or extreme fog, where visual contrast is poor.

C. Adaptive Deep Appearance Re-Identification

VAKT uses a lightweight HSV histogram. Accuracy could be further improved via

- Compact Re-ID CNN embeddings (quantized for edge devices),
- Memory-bank assisted identity retention across long tracks.

This would reduce the remaining ID switches and identity fragmentation in highly crowded or overlapping scenes.

D. Real-time Deployment on Embedded Flight Computers

Tests were performed primarily on workstation-grade GPUs. Future work aims to

- Deploy the model on Jetson Orin NX / Jetson Xavier,
- Optimize via TensorRT, pruning, or quantization,
- Reduce latency for on-board edge inference (not relying on ground station streaming).

Achieving this would make the system fully autonomous and operational in remote SAR missions.

Final Statement

This project demonstrates that combining enhanced transformer-based detection, adaptive lightweight tracking, and geo-spatial localization enables drone surveillance systems to transition from passive video collection to intelligent operational support. The framework lays a strong foundation for AI-assisted SAR and opens the path for future innovation toward fully autonomous aerial rescue systems.

REFERENCES

1. Cheng, S., Song, J., Zhou, M., Wei, X., Pu, H., Luo, J., & Jia, W. (2024). EF-DETR: A Lightweight Transformer-Based Object Detector With an Encoder-Free Neck. IEEE Transactions on Industrial Informatics.
2. Jiang, L., Yuan, B., Du, J., Chen, B., Xie, H., Tian, J., & Yuan, Z. (2024). MFFSODNet: Multiscale Feature Fusion Small Object Detection Network for UAV Aerial Images. IEEE Transactions on Instrumentation and Measurement.
3. Liu, K., Fu, Z., Jin, S., Chen, Z., Zhou, F., Jiang, R., & Ye, J. (2025). ESOD: Efficient Small Object Detection on High-Resolution Images. IEEE Transactions on Image Processing.
4. Liu, Z., Mao, H., Wu, C., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). ConvNeXt: A ConvNet for the 2020s. Proceedings of IEEE CVPR.
5. Ma, S., Zhang, Y., Peng, L., Sun, C., Ding, B., & Zhu, Y. (2025). OWRT-DETR: Real-Time Transformer Network for Small-Object Detection in Open-Water Search and Rescue From UAV Imagery. IEEE Transactions on Geoscience and Remote Sensing.
6. Tan, M., Pang, R., & Le, Q. (2020). EfficientDet: Scalable and Efficient Object Detection via BiFPN. IEEE CVPR.
7. Wojke, N., Bewley, A., & Paulus, D. (2017). DeepSORT: Simple Online and Realtime Tracking with a Deep Association Metric. IEEE ICIP.
8. Xue, Y., Jin, G., Shen, T., Tan, L., Wang, N., Gao, J., & Wang, L. (2024). Consistent Representation Mining for Multi-Drone Single Object Tracking. IEEE Transactions on Circuits and Systems for Video Technology.