Xi'an Jiaotong-Liverpool University
西交利物浦大学

Research of 3D Perception Algorithm Based on Multi-Sensor Fusion

3D Perception in USV Target Tracking:
Enhancing Navigation with Multi-Sensor Fusion

By

Yiheng XUE
2254281

Supervised By

Yong LIU, Zhenfeng XUE, Ming XU

A DISSERTATION

Submitted to

Xi'an Jiaotong-Liverpool University

in partial fulfillment of the requirements
for the degree of

MASTER OF RESEARCH

28/12/2023

ABSTRACT

In recent years, the advancement of mobile robots across sectors like intelligent manufacturing and personal transportation has greatly enhanced human life. This study specifically addresses challenges in unmanned surface vehicles (USVs), with a focus on enhancing their operational capabilities through advanced 3D perception algorithms and multi-sensor fusion techniques. Recognizing the importance of edge computing for stability and reliability, our research delves into selecting the most appropriate sensors for diverse tasks and developing robust sensor preprocessing and fusion algorithms.

Central to our study is the innovative application of these methodologies in USV target tracking. We have thoroughly explored the strengths and limitations of various sensors, particularly in dynamic aquatic environments, which are crucial for real-time decision-making and motion planning in USVs. By successfully integrating data from IMUs, GPS, cameras, and LiDAR, our team has constructed dynamic obstacle maps and achieved effective real-time target localization and tracking, even in scenarios lacking direct communication with the target. This research not only presents a practical edge computing solution for USVs but also advances their autonomy and operational efficiency. The development of a unified simulation and real-world USV platform, which combines a multi-sensor system, validates the feasibility, stability, and robustness of our approach. This work significantly contributes to the field of intelligent robotics by providing a scalable and reliable framework for USVs, potentially impacting the development of multi-agent systems and enhancing the capabilities of USV swarms.

Index Terms: 3D perception, multi-sensor fusion, unmanned surface vehicle (USV)

DECLARATION

I hereby certify that this dissertation constitutes my own product, that where the language of others is set forth, quotation marks so indicate, and that appropriate credit is given where I have used the language, ideas, expressions, or writings of another.

I declare that the dissertation describes original work that has not previously been presented for the award of any other degree of any institution.

Signed,

Xue Yiheng
薛毅恒

ACKNOWLEDGEMENTS

In the tapestry of life, academic pursuit is but a single thread, weaving into the fabric of our experiences and perhaps heralding new beginnings. My deepest gratitude goes to my parents, Mr. Weian XUE and Mrs. Ying WANG, whose unwavering support has been a lighthouse in the often tumultuous sea of research. Despite not fully grasping the intricacies of my work, their encouragement has been a constant, nurturing my journey without ever casting their own shadows upon it.

The path of research is often marked by trials and tribulations, and it is here that I must acknowledge the generous support of Huzhou Institute of Zhejiang University (HiZJU), who illuminated my way with a full scholarship. A beacon of guidance, Prof. Yong LIU, and Dr. Zhenfeng XUE, provided pivotal insights that shaped my research voyage. Prof. Ming XU's meticulous mentorship and sagacious counsel have carved indelible marks on my scholarly soul, as have the incisive inquiries and invaluable suggestions from Prof. Myeongsu Seong during my oral presentations. I extend my gratitude to Prof. Shanshan ZHAO from the School of AI and Advanced Computing (AIAC), whose faith in me as a teaching assistant not only enriched my academic life but also honed my English proficiency, lending an additional layer of depth to my university experience.

Research, a realm where positive reinforcement is as rare as it is vital, has been made more approachable thanks to the collaborative efforts of my colleague Doctoral Candidate Tao HUANG in task planning, and the invaluable assistance of Mr. Zheng ZHANG in both the virtual and physical realms of unmanned ship design. To the friends I found in Huzhou, who shared in the simple joys of badminton and football, you have added a hue of normalcy to the often monochromatic life of a researcher.

A special note of appreciation is reserved for my girlfriend, Mrs. Zhenqi CAO, and my feline companion, Switch, a ragdoll cat whose silent presence has been a comforting constant, a soft echo in the quiet moments of contemplation and toil.

And to Mr. Lin WANG, whose artistic acumen brought visual life to my ideas, his contributions are invaluable. My appreciation extends to Doctoral Candidate Chuke CHEN, whose insightful guidance on my thesis's structure was both efficient and illuminating. This journey has been further enriched by the collaborative efforts with Dr. Ningning BAI and Mr. Zhijun LV in specific subtasks of my research, for which I am profoundly grateful.

Contents

LIST OF TABLES

List of Figures

Chapter 1. Introduction

The recent advancements in mobile robotics, particularly in sectors such as intelligent manufacturing and personal transportation, have underscored the importance of computational efficiency in this rapidly evolving field. While acknowledging the significant contributions of large language models (LLMs) and Transformer technology, this study pivots its focus towards the realm of edge computing, especially in the context of unmanned surface vehicles (USVs). The decision to concentrate on edge computing stems from the need to address operational stability and reliability challenges within the dynamic environments where USVs operate, as well as to optimize the balance between computational load and performance efficacy.

Edge computing emerges as a pivotal element in this context, offering a solution to the high computational demands typically associated with advanced robotics algorithms. By facilitating computation closer to the data source, edge computing enables more responsive and efficient processing, a crucial factor for real-time applications in USV operations. This approach dovetails with the study's aim of augmenting USV capabilities through the integration of advanced 3D perception algorithms and multi-sensor fusion techniques.

Central to the methodology of this research is the exploration of sensor selection and optimization for USVs, emphasizing the development of robust algorithms for sensor preprocessing and fusion. These efforts are geared towards harnessing the full potential of edge computing, aligning with the overarching goal of enhancing the operational efficiency of unmanned surface vehicles. This study thereby represents a significant contribution to the field of mobile robotics, focusing on edge-centric solutions tailored to meet the unique challenges and requirements of USVs.

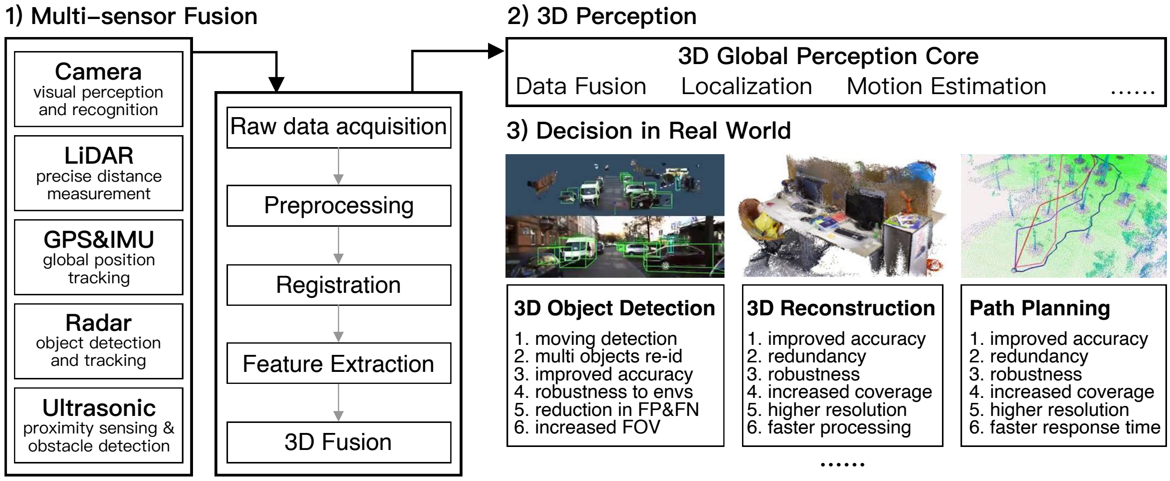

Figure 1.1: 3D Perception Multi-Tasking Enabled by Multi-Sensor Integration

1.1 Scope

The analysis within this thesis is structured around three critical dimensions that collectively provide a comprehensive understanding of USV technology, as illustrated in Figure 1.1. These dimensions are essential to dissecting the challenges and opportunities in advancing USVs, encompassing the sophistication of perception algorithms, the capabilities and interplay of various sensors, and the integration and refinement of multi-sensor fusion algorithms.

1.1.1 Advancements in 3D Perception Algorithms for USVs

The exploration of 3D perception algorithms in USVs is central to enhancing their operational capacities in autonomous navigation, object recognition, and spatial awareness. This includes a deep dive into how these algorithms interpret complex visual and spatial information to create accurate, multi-dimensional environmental models. The focus extends to the challenges inherent in processing and integrating volumetric data from diverse sources, ensuring that USVs can reliably navigate and respond to dynamic maritime environments. Current advancements in this domain are evaluated, emphasizing the development of algorithms that are not only precise but also computationally efficient, catering to the unique demands of USVs operating in edge computing environments.

Figure 1.2: Perception Sensors from 1-D (Left) to 3-D (Right)
Sensors depicted are products of brands Parker, HIKVISION, HESAI, and LIVOX.

1.1.2 Sensor Integration and Functionality in USVs

In this scope, the analysis is concentrated on the array of sensors critical to USVs, encompassing Inertial Measurement Units (IMUs), GPS, cameras, and different types of LiDAR, such as solid-state and rotating LiDAR. The study delves into the distinctive attributes and functionalities of each sensor, elucidating their roles in accurate data gathering and environmental perception. Emphasis is placed on the synergy of these sensors, examining how their collective data contributes to a more nuanced and comprehensive understanding of the USVs' operational landscape. This comparative analysis not only highlights the strengths and limitations of each sensor type but also their interplay in enhancing the overall sensory capability of USVs. An illustrative representation of these sensors, including IMUs, GPS, cameras, rotating LiDAR, and solid-state LiDAR (Livox Mid-360), is provided in Figure 1.2.

1.1.3 Optimizing USV Performance through Multi-Sensor Fusion

This scope addresses the critical role of multi-sensor fusion algorithms in consolidating and interpreting data from various sensors on USVs. The focus here is on the advantages this fusion brings, notably in terms of system stability, robustness, and enhanced decision-making capabilities. The algorithms are explored for their effectiveness in creating an integrated situational view, especially beneficial in Bird's Eye View (BEV) perspectives for complex task execution. The discussion encompasses how this fusion facilitates a holistic approach to USVs' operational tasks, including navigation, obstacle avoidance, and mission-specific objectives. This comprehensive examination underscores the importance of multi-sensor fusion in elevating the operational efficiency, safety, and versatility of USVs in diverse maritime applications.

1.2 Problem Statement

1.2.1 General Question about 3D Perception

The general question investigates the unique advantages of diverse sensors in 3D perception tasks. This inquiry centers on how integrating these sensors can lead to more robust and precise perception systems applicable across various domains. Key focus areas include:

1. **Multi-Sensor Data Fusion:** Exploring strategies for combining data from sensors with different modalities to enhance the accuracy and stability of perception systems.
2. **Adaptive Perception in Dynamic Environments:** Examining how sensor integration contributes to more resilient and adaptable perception in changing conditions, such as varying lighting or weather.
3. **Sensor Synergy for Complex Scenarios:** Assessing the effectiveness of multi-sensor approaches in complex, obstacle-rich settings, where single-sensor systems may fall short.
4. **Perception to Decision-Making:** Investigating the impact of integrated sensor data on the decision-making processes in autonomous systems, emphasizing the transition from raw data to actionable insights.

1.2.2 Specific Question about USV Target Tracking

The specific inquiry focuses on realizing stable and cost-effective 3D target localization and trajectory prediction for USV target tracking on mobile platforms, as depicted in Figure 1.3. The aim is to ensure consistent target pursuit with effective obstacle avoidance. The areas of interest include:

1. **IMU-LIDAR Correction in Turbulent Waters:** This focuses on motion correction for USVs navigating unstable aquatic environments, aiming to refine the integration of IMU

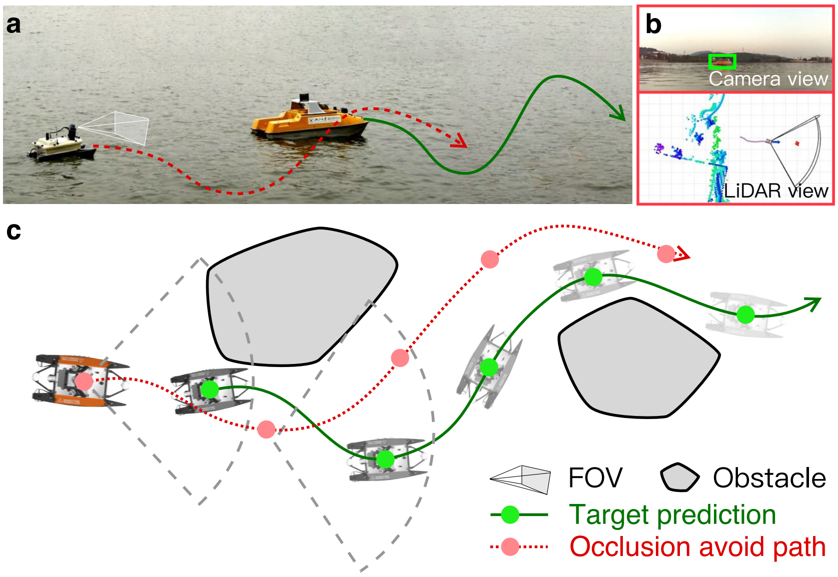

Figure 1.3: USV Tracker Sketch in Obstacle Map

a: Physical experiment in target tracking with a sensor-equipped white boat and a yellow boat as the moving target. b: Target detection from sensor images and obstacle grid-map from point cloud data. c: Target tracking in obstacle map, showing path planning based on predicted target trajectory.

and LIDAR data. The goal is to enhance positional accuracy and stability amidst turbulent conditions.

2. **Reducing Aquatic Interference:** The study evaluates dataset augmentation techniques for imaging systems in aquatic settings, concentrating on minimizing the impact of water-specific interferences, such as reflections and refractions, to improve imaging quality.

3. **Sim-to-Real for Experimental Limitations:** This addresses the limitations inherent in real-world testing environments by leveraging simulation platforms. The approach involves using simulated environments to validate perception and planning algorithms, ensuring their efficacy before deployment in actual aquatic scenarios.

4. **Optimizing Localization Range:** The research aims to determine the optimal operational range for effective localization under diverse aquatic conditions. It seeks to strike a balance between range and accuracy, adapting to the dynamic nature of aquatic environments for enhanced USV performance.

1.3 Approach

In the study, four distinct methodologies are employed for separate tasks. The first task, target detection, is addressed using the YOLO algorithm, optimized for efficient and real-time object detection. This method is particularly effective in identifying and classifying objects within images rapidly. The second task involves building an obstacle map, for which the Euclidean Signed Distance Fields (ESDF) method is utilized. This method excels in spatial data representation, providing accurate distance measurements and spatial relationships essential for navigation and obstacle avoidance. For dataset generation, the study employs advanced 3D reconstruction techniques. These techniques are crucial for creating detailed and realistic datasets that closely mimic real-world scenarios, thereby enhancing the robustness of the subsequent analysis. The fourth task, target trajectory prediction, integrates the Extended Kalman Filter (EKF) with neural network methodologies such as Long Short-Term Memory (LSTM) networks. This hybrid approach leverages the strengths of both EKF in handling uncertainties and LSTM in learning from time-series data, resulting in improved prediction accuracy.

The experimental phase of the study was conducted within a sophisticated simulation environment, designed and built in-house. This environment simulates real-world conditions, including dynamic feedback systems, and environmental factors such as wind and wave interactions. It also features a complex background to increase the richness and variability of the image data used in the experiments. The simulator is equipped with multi-sensor capabilities, allowing for the customization and fine-tuning of sensor parameters.

Following the simulation phase, the algorithms were further validated through physical experiments. These experiments were conducted using a specially designed vessel and sensor platform that closely corresponded to the configurations used in the simulation. This approach ensured a consistent and thorough evaluation of the algorithms' effectiveness in both simulated and real-world conditions, thereby enhancing the validity of the research findings.

1.4 Contributions

The outcomes of this thesis are significant contributions to the field of USVs, encompassing the development of simulation platforms, the design of sensory systems, the optimization of multi-sensor fusion algorithms, and providing insightful recommendations for multi-USV systems. These achievements are detailed as follows:

1. **Development of a Gazebo-Based USV Simulation Platform:** A sophisticated simulation platform for USVs was constructed using Gazebo, featuring a realistically modeled catamaran with dynamic feedback, including environmental dynamics like water surface and wind disturbances. This platform integrates various sensors communicating through ROS, providing a comprehensive environment for testing and refining USV functionalities.

2. **Design of a Sensory System for a Physical Catamaran:** Corresponding to the simulation model, a sensory system for a physical twin-hulled boat was designed and implemented. This system, tested in real aquatic environments, effectively fulfills the requirements for target tracking, demonstrating the practical application of the simulated models.

3. **Optimization of Multi-Sensor Fusion Algorithms on Edge Computing Devices:** This research has successfully optimized multi-sensor fusion algorithms for edge computing devices, focusing on advancements in target detection, target depth estimation, target motion prediction, and USVs' own trajectory planning and decision-making within obstacle-rich environments. These algorithms have been seamlessly integrated into the USVs' control systems, enhancing the USVs' ability to navigate through complex scenarios. This integration allows for precise control over the USVs' propulsion, enabling efficient tracking of targets while maintaining a safe distance, thereby significantly improving the autonomous capabilities of the USVs in various operational contexts.

4. **Comparative Analysis of Sensors and Strategic Recommendations for Perception Platforms:** This study conducted a comprehensive analysis of various sensors, integrating them into multi-sensor platforms for edge computing. Strategic recommendations emerged for optimizing sensor fusion, enhancing perception accuracy in USVs. The research utilized a sim-to-real approach, validating sensor platforms through combined simulation and real-world testing, and developed a cost-effective physical USV, advancing practical applications in USV technology.

5. **Advancements in Multi-USV Systems and Swarm Planning:** The research contributed to the development of multi-USV systems and swarm planning, providing a foundational platform for exploring advanced operational concepts. This work is pivotal for future progress in swarm intelligence and coordinated USV operations, opening new avenues for complex maritime missions.

These outcomes collectively represent a significant advancement in USV technology, providing valuable insights and practical solutions that enhance the operational capabilities and efficiency of USVs in diverse maritime environments.

In synthesizing these methodologies and contributions, this thesis not only bridges the gap between theoretical concepts and practical applications in USV technology but also lays a foundational framework for future explorations. The integration of advanced algorithms and multi-sensor fusion within the dynamic and challenging realm of maritime environments marks a significant stride towards autonomous maritime operations. The subsequent chapters will delve deeper into each aspect of our approach, meticulously dissecting the intricacies of our methodologies and the tangible impacts of our contributions. This journey from conceptualization to realization reflects not just a series of technological advancements but also a step forward in our understanding of intelligent systems and their interaction with the complex, ever-changing natural world. Ultimately, this work seeks to pioneer new frontiers in USV capabilities, paving the way for groundbreaking advancements in intelligent robotics and autonomous systems.

Chapter 2. Literature Review

2.1 Comprehensive Overview of 3D Perception: Tasks and Datasets

In the domain of 3D perception tasks, particularly relevant to autonomous systems like USVs and autonomous vehicles, the literature distinguishes between dense and sparse tasks. Dense tasks, such as 3D modeling and semantic segmentation, require dense image data or solid-state LiDAR inputs. These tasks are pivotal for creating detailed environmental models and understanding contextual elements. Sparse tasks, including object detection and tracking, are more versatile in data requirements, allowing for both dense and sparse point cloud inputs. The preprocessing of input data in these tasks often involves a trade-off between data reduction for speed enhancement and the retention of crucial information for accuracy, as depicted in Figure 2.1 and Figure 2.2.

Prominent in the field are public datasets like Waymo (Sun et al., 2019), Kitti (Geiger et al., 2012), nuScenes (Caesar et al., 2020), and Apollo (Huang et al., 2018), which have significantly contributed to advancements in 3D perception. These datasets provide diverse scenarios and challenges, encompassing classic problems like object detection and extending to image semantic segmentation, depth estimation, and object tracking. Crucially, these datasets also include tasks related to environmental perception, such as trajectory and behavior prediction, lane line detection, and SLAM (Simultaneous Localization and Mapping).

The fusion of data from various sensors to accomplish multiple tasks is a recurring theme in the literature. Some methods propose end-to-end multi-tasking within a single network architecture, necessitating substantial computational resources. This presents a challenge in migrating these tasks to edge computing devices, where resource constraints are a significant consid-

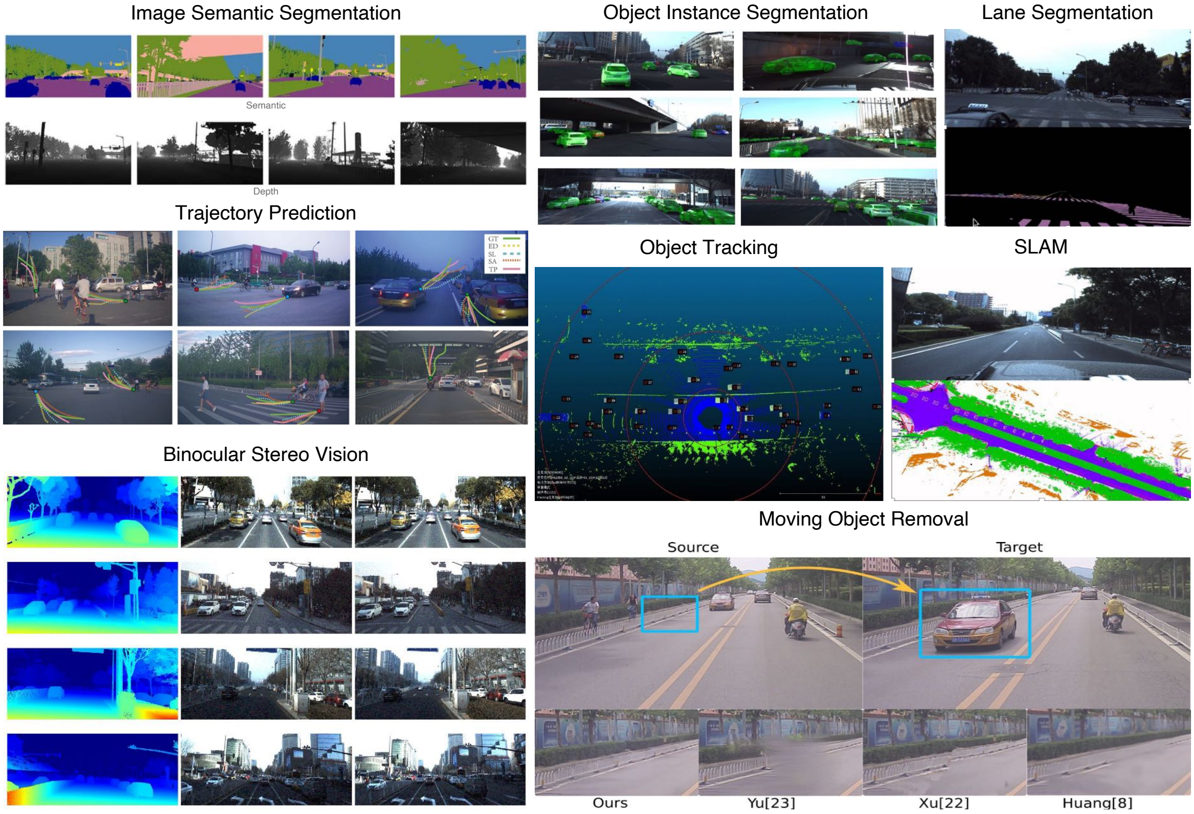

Figure 2.1: Examples of 3D Perception Tasks
Top row: Semantic segmentation images from Waymo (Sun et al., 2019). Middle row: Trajectory prediction using Ma et al. (2019). Bottom row, right: Moving object removal by Liao et al. (2020). All other tasks and data derived from sub-tasks in the Apollo dataset (Huang et al., 2018).

eration. The exploration of efficient algorithmic solutions and hardware optimizations for edge deployment remains an active area of research, aiming to balance computational demands with the real-time processing needs of autonomous systems.

These methods underscores the complexity and diversity of 3D perception tasks and the importance of rich datasets in driving the field forward. It highlights the ongoing efforts to optimize these tasks for practical applications, especially in edge computing environments, reflecting a critical area of development in autonomous systems technology.

2.2 Advancements in Object Detection Technologies

In exploring the field of object detection, this review categorizes advancements into two distinct domains based on data dimensionality. The first domain, *2D Vision-Based Object Detection*,

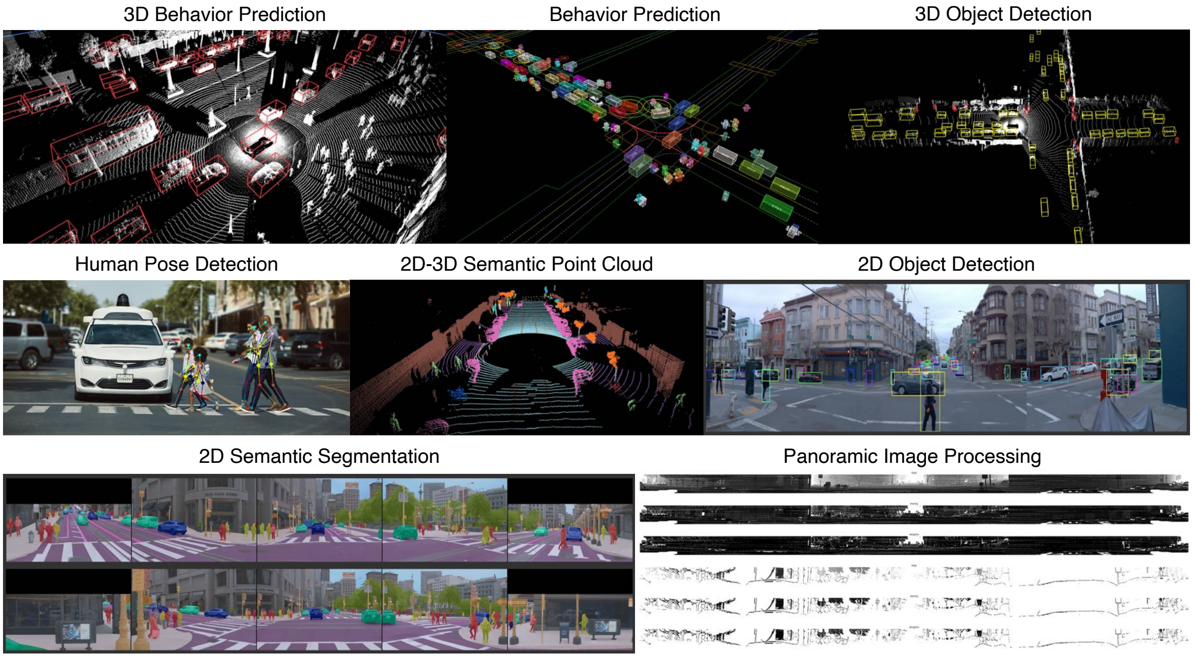

Figure 2.2: 3D Perception Datasets and Applications

primarily focuses on techniques that interpret two-dimensional image data. In contrast, the second domain, *3D Object Detection Methods*, delves into approaches that process and analyze three-dimensional data. This bifurcation provides a structured understanding of how object detection technologies have evolved and adapted to different data formats, each with its unique challenges and applications.

2.2.1 2D Vision Based Object Detection

Faster R-CNN (Ren et al., 2015) represents a pivotal advancement in object detection, integrating Region Proposal Networks (RPNs) with Fast R-CNN (Girshick, 2015). While its RPN layer efficiently identifies object proposals for accurate detection, the computational intensity may pose challenges in edge computing platforms with limited resources. YOLO (Redmon et al., 2016) stands for *You Only Look Once*, is renowned for its exceptional speed and efficiency in object detection. Despite its capability to analyze the entire image in a single evaluation, the high computational demand for processing large images can be a limiting factor in resource-constrained environments. The Detection Transformer (DETR) (Carion et al., 2020) introduces a novel transformer-based architecture to the realm of object detection. While DETR stream-

Table 2.1: Comparison of Algorithms in Different Tasks

Tasks	2D Image	3D Point Cloud	Multi-sensor Fusion
Detection	Faster-RCNN (Ren et al., 2015), Yolo (Redmon et al., 2016), DETR (Carion et al., 2020)	VoteNet (Qi et al., 2019), PV-RCNN (Shi et al., 2020), VoxelNet (Zhou and Tuzel, 2018)	MV3D (Chen, Ma, Wan, Li and Xia, 2017), AVOD (Ku et al., 2018)
Segmentation	U-Net (Ronneberger et al., 2015), DeepLab (Chen, Papandreou, Kokkinos, Murphy and Yuille, 2017)	PointNet (Qi, Su, Mo and Guibas, 2017), PointNet++ (Qi, Yi, Su and Guibas, 2017), 3D-U-Net (Çiçek et al., 2016)	FusionSeg (Jain et al., 2017), PointFusion (Xu et al., 2018)
Tracking	SiamFC (Bertinetto et al., 2016), MDNet (Nam and Han, 2015), DeepSORT (Wojke et al., 2017)	PointTrackNet (Wang et al., 2020), 3D-siamRPN (Fang et al., 2021)	-
BEV	-	-	LSS (Philion and Fidler, 2020), BEVFormer (Li et al., 2022), BEVFusion (Liu et al., 2023)

lines the detection pipeline and showcases the versatility of transformers, its reliance on global reasoning across the entire image might be computationally demanding for edge computing platforms.

U-Net (Ronneberger et al., 2015) designed for biomedical image segmentation, has become widely used in various image segmentation tasks. Its symmetric expanding path enables precise localization, but the complexity can lead to significant computational demands, particularly in real-time analysis scenarios on limited hardware resources. DeepLab (Chen, Papandreou, Kokkinos, Murphy and Yuille, 2017), a state-of-the-art semantic segmentation algorithm, is known for its use of atrous convolution to capture multi-scale context. Although achieving high accuracy, its sophisticated architecture results in substantial computational load, making it challenging to deploy in edge computing scenarios where resources are constrained.

In the realm of tracking, Siamese Fully Convolutional (SiamFC) Bertinetto et al. (2016) networks offer a novel approach using a fully convolutional siamese network. While noted for its simplicity and speed, its potential struggle with significant appearance changes and occlusions may be a concern in edge computing environments. Multi-Domain Network (MDNet) (Nam and Han, 2015) employs a novel domain-specific layer approach, allowing robust tracking across various scenarios. However, its computational intensity can limit real-time application in resource-constrained environments. DeepSORT (Wojke et al., 2017), an extension of the Simple Online and RealtimeTracking (SORT) algorithm, incorporates deep learning features for improved accuracy. Despite its enhanced performance, the increased complexity of DeepSORT can result in higher computational demands, making it less suitable for real-time applications in edge computing platforms.

These algorithms, each contributing significantly to their respective fields, demonstrate a balance between accuracy and efficiency. However, their computational requirements present challenges in edge computing environments, where processing power and memory are often limited, affecting real-time processing and scalability in demanding applications.

2.2.2 3D Object Detection Methods

In 3D Point Cloud Algorithms, several methods have shown promise, yet they pose substantial computational challenges in edge computing environments.

For detection, VoteNet (Qi et al., 2019) introduces deep learning for 3D object detection in point clouds. While innovative, its complex voting scheme and 3D convolution operations are computationally intensive, making it less feasible on limited-resource edge platforms. PV-RCNN (Shi et al., 2020) combines voxel and point-based networks for enhanced detection accuracy. However, this combination leads to increased computational load, challenging its deployment on edge devices with constrained processing capabilities. Similarly, VoxelNet (Zhou and Tuzel, 2018) utilizes 3D convolutions for voxel feature encoding, which demands high computational power, posing difficulties in resource-limited edge computing scenarios.

In the segmentation domain, PointNet (Qi, Su, Mo and Guibas, 2017) and PointNet++ (Qi, Yi, Su and Guibas, 2017) directly process point clouds, which can be computationally demanding due to the need for processing large numbers of points and complex feature extraction. 3D-U-Net (Çiçek et al., 2016), adapted for volumetric segmentation, faces similar challenges. The 3D convolutions and large model size inherent in 3D-U-Net require substantial computational resources, hindering its practicality on edge devices.

Regarding tracking, PointTrackNet (Wang et al., 2020) and 3D-siamRPN (Fang et al., 2021) offer advanced tracking capabilities. However, their sophisticated architectures, including point cloud processing and Siamese networks, result in high computational overhead. This makes them challenging to deploy on CPUs of edge computing platforms, where processing power and memory are limited.

Overall, while these methods have advanced 3D point cloud processing, their computational requirements pose significant challenges in edge computing environments. The high processing demands and memory requirements limit their practical application on edge platforms, particularly when only CPU-based processing is available. This highlights the need for more computationally efficient algorithms in scenarios where hardware resources are constrained.

2.3 Innovations in Multi-Sensor Fusion Techniques

In Multi-sensor Fusion, various methods have been developed to enhance detection, segmentation, and Bird's Eye View (BEV) representation, crucial for applications like autonomous driving and robotic navigation.

Detection in multi-sensor fusion employs methods like MV3D (Chen, Ma, Wan, Li and Xia, 2017) and Aggregate View Object Detection (AVOD) (Ku et al., 2018). MV3D combines LiDAR point cloud and image data to generate accurate 3D object detections, but its early fusion approach, integrating raw data from different sensors, can be computationally expensive and sensitive to sensor misalignments. AVOD, on the other hand, focuses on a later stage fusion by aggregating region proposals from both camera and LiDAR features, which, while more efficient in handling sensor discrepancies, still faces the challenge of high computational load for real-time processing on edge devices.

Segmentation techniques like FusionSeg (Jain et al., 2017) and PointFusion (Xu et al., 2018) illustrate the diversity in fusion strategies. FusionSeg applies early fusion at the data level, combining visual and spatial information, which can lead to improved segmentation accuracy but at the cost of increased computational complexity. PointFusion, employing late fusion at the feature level, fuses deep features from images and point clouds, offering a balance between computational efficiency and accuracy, yet may not fully exploit the complementary nature of different sensor modalities.

For BEV methods such as LSS (Philion and Fidler, 2020), BEVFormer (Li et al., 2022), and BEVFusion (Liu et al., 2023) demonstrate advanced techniques. LSS (Lift, Splat, Shoot) lifts 2D semantics into a 3D space, providing a comprehensive BEV representation, but its reliance on high-dimensional feature lifting can be computationally demanding. BEVFormer and BEV-Fusion, leveraging transformer architectures and deep fusion strategies, offer improved BEV representations. These methods exploit the depth accuracy of point clouds to enhance the BEV perspective, which is particularly beneficial for robotic planning and navigation. However, their complex architectures and intensive computational requirements pose significant challenges for deployment on resource-constrained edge computing platforms.

In summary, while multi-sensor fusion methods offer enhanced detection, segmentation, and BEV representation, they often come with high computational demands, especially when early fusion strategies are employed. For BEV applications, the depth precision from point clouds is valuable, yet the computational challenges must be carefully managed, particularly for real-time applications in edge computing scenarios.

Chapter 3. Research Methodology

3.1 Experiment Statement

In dynamic motion environments, especially in unstructured outdoor scenarios, the strategic selection and combination of sensor technologies is paramount. This study underscores the importance of GPS and IMU in estimating the pose and position of mobile robots within the world coordinate system, essential for motion correction in perception sensors. Central to this research is the *Information Conservation Concept*, which posits a correlation between the dimensionality of sensors, their operational frequency, and the resultant information loss.

This concept is substantiated by examining various sensors: 1D pressure sensors offer high precision but are limited in capturing 3D details. Conversely, 2D image sensors, like cameras, provide rich features suitable for dense 3D reconstruction. However, 3D sensors such as LiDAR, while providing spatial accuracy, face limitations in dynamic settings due to lower operational frequencies, leading to information loss.

This interplay between sensor dimensionality, operational frequency, and information loss is illustrated in Figure 3.1. The diagram demonstrates how changes in sensor dimensions and frequencies align with the Information Conservation Concept. It elucidates the inherent trade-offs in designing sensor systems for dynamic robotic applications, emphasizing the necessity of integrating 2D and 3D data. This integration mitigates individual sensor limitations and enhances the overall perception system. Consequently, the research methodology focuses on developing multi-sensor fusion techniques, designed to amalgamate diverse data sources effectively, thus augmenting mobile robots' performance in unstructured outdoor environments.

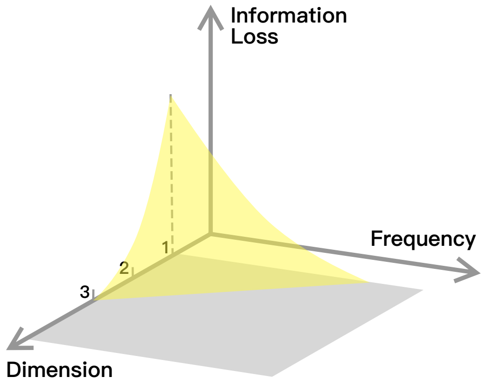

Figure 3.1: Hypothesis of Information Loss with Different Dimension

Table 3.1: Sensors Comparison

Sensor	**Dimension**	**Frequency**	**Accuracy**	**Spatial Information**
IMU	1D, 6/9DoF	100-1000Hz	B, A w/ loop	Full-body motion
GPS	1D, 3D pos	1-10Hz	B, A w/ DGPS	Outdoor position
Bio-sensor	1D	0.02-400Hz	A+	3D contact pressure
Camera	2D	10-150Hz	A	Dense 2D pixels
LiDAR	3D	$\sim$10Hz rotary	B	Sparse 3D point cloud

Accuracy is categorized by levels, indicating precision ranges. *LiDAR frequency* refers to the rotational speed of mechanical LiDAR. *Loop* denotes IMU loop closure correction in SLAM. *DGPS* signifies Differential Global Positioning System, enhancing GPS accuracy.

3.1.1 Optimizing Multi-Sensor Selection for Enhanced Perception

As demonstrated in Figure 3.2, high-frequency sensors are utilized for the detection and classification of 3D micro-features (Bai et al., 2023). This setup employs a combination of 1D high-frequency sensors and IMUs, enabling the recovery of posture in dynamically complex motions and providing real-time feedback. Figure 3.3 illustrates a mixed-reality scenario in autonomous driving for 3D reconstruction tasks, where 2D image sensors rich in features are integrated with IMU and GPS. This combination facilitates dense 3D reconstruction in environments characterized by rich features but limited viewing angles (limited-view specifically refers to camera orientations fixed in alignment with the motion direction of the mobile robot).

These two tasks test sensor combinations under challenges analogous to those faced in USV target tracking: sparse visual features, dynamic environments, complex motions, systems requiring real-time feedback, and limited viewing perspectives. The objective is to design an efficient, real-time, and precise multi-sensor fusion platform for 3D target perception in dy-

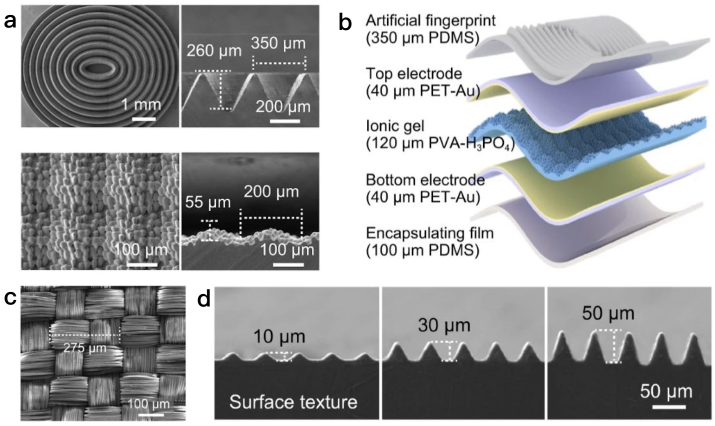

Figure 3.2: Combination I: High-Frequency Sensing for Dynamic Signal Classification

namic settings. Algorithm optimization is conducted in a custom-built simulation environment, while appropriate sensors are employed in physical experiments for algorithm validation.

The first sensor combination confirms that low-dimensional sensors can capture 3D spatial information, with 100Hz IMU data being effective for motion distortion correction and real-time posture estimation. The second combination clarifies that while target detection with 2D image sensors is stable, depth estimation remains unreliable, and a 30Hz image frame rate suffices for stable target detection in moving environments.

Our experimental platform consists of a combination of 3D LiDAR, 2D Camera, IMU, and GPS. This assembly forms a robust and stable system, addressing the unique requirements of dynamic scene perception and targeted operation in USV applications.

3.1.2 From Simulation Validation to Real-World Implementation

The use of popular simulation engines like Unity[1] and Unreal Engine 5[2] (UE5), and their application in the field of autonomous driving through platforms such as Carla, has significantly facilitated research, especially in areas like reinforcement learning. However, these simulators often require complex development environments and substantial computational resources.

[1] Unity website: https://unity.com/

[2] Unreal Engine website: https://www.unrealengine.com/en-US

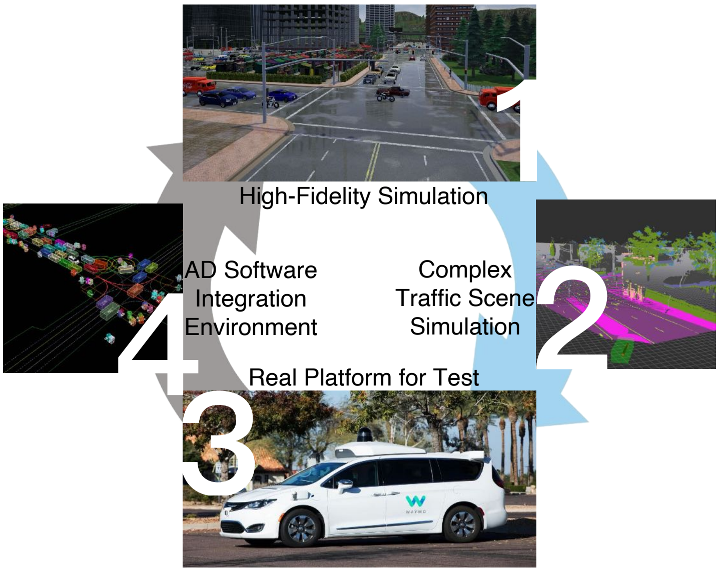
1
High-Fidelity Simulation
2
Complex Traffic Scene Simulation
3
Real Platform for Test
4
AD Software Integration Environment

Figure 3.3: Combination II: Integrated Sim-to-Real Pipeline for Enhanced 3D Reconstruction

In the experimental phase described as Combination II, experience has been gained in transitioning from the CARLA[3] simulator to real-world datasets such as KITTI[4]. This transition does not dwell on the issues of transfer learning across domains but focuses on the specific needs of USVs, particularly in dynamics feedback and control. In the simulation environment, the objective is to render images optically with limited computational resources, incorporating diverse rigid body models to mimic real-world scenarios. This approach aims to ensure that the dynamics feedback in the simulator closely mirrors that of the real world.

A key aspect of this simulation-to-reality approach is maintaining consistency in certain elements, such as dynamics feedback and raw sensor parameters. By ensuring these elements remain uniform across both simulated and real-world environments, the research reduces experimental workload. The parameters optimized in the simulator can be transferred to the real environment, where fine-tuning through minimal adjustments leads to effective validation. This methodology not only streamlines the research process but also provides a robust framework for testing and validating algorithms in a controlled yet realistic setting, thereby enhancing the overall development and application of USV technology.

[3] CARLA homepage: https://carla.org/

[4] KITTI homepage: https://www.cvlibs.net/datasets/kitti/

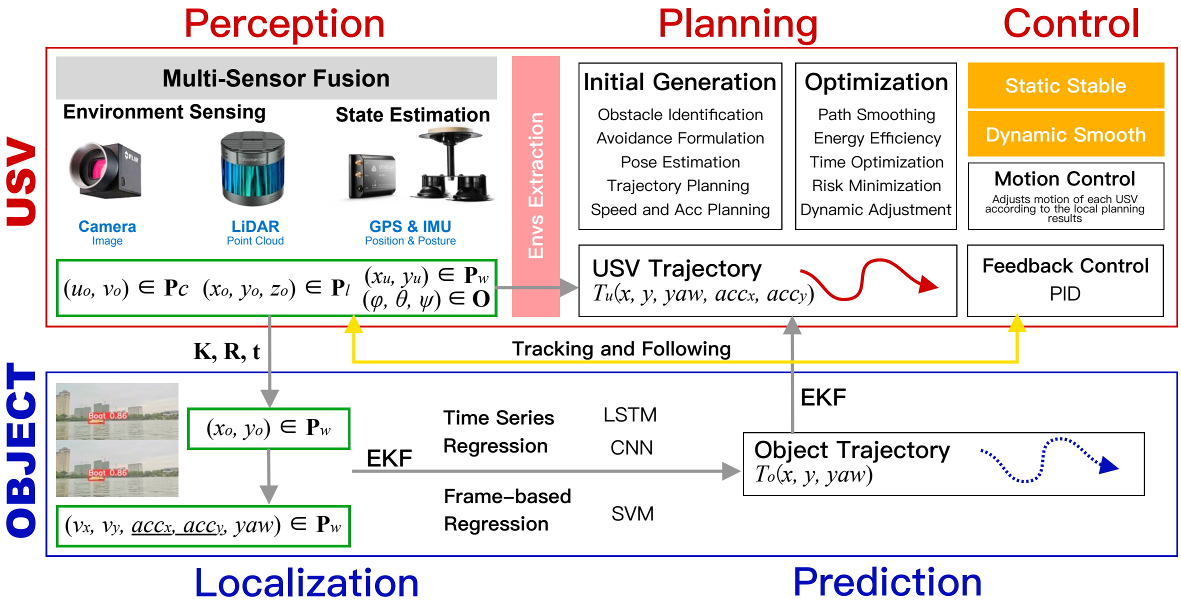

Figure 3.4: Diagram of the USV Target Tracking System

Subsequent sections will delve into the specifics of both the simulation tools and the actual physical products, providing a detailed account of how each contributes to the holistic development and validation process in USV research.

3.2 Target Detection and Trajectory Prediction Based on Multi-Sensor Fusion

The USV Tracking system, delineated in this research, is segmented into three integral components: perception, planning, and control. The system's architecture, as depicted in Figure 3.4, integrates multi-sensor data into the USV's perception layer, fulfilling two key functionalities:

1. **Construction of Dynamic Obstacle Map:** Utilizing sparse grid-map formatting with a voxel resolution of 0.2 meters, the system adeptly constructs dynamic obstacle maps. This capability is crucial for accurately depicting and navigating the USV's immediate operational environment.

2. **Detection of Target 3D Coordinates:** The system demonstrates remarkable precision in detecting the three-dimensional coordinates of targets, maintaining an average error margin of approximately 0.16 meters. Such accuracy is vital for effective target engagement and maneuvering.

Coordinate transformation, a pivotal process in the system, is approached separately for the USV and the target. This separation facilitates a nuanced understanding and prediction of the target's motion. Within the perception platform of the USV, a rigorous calibration of the multi-sensor array is undertaken. In the simulated environment, the Rodrigues' formula as indicated in Equation 3.2 is employed, obviating the need for traditional calibration boards due to the lack of distortion and pre-determined FOV parameters. The resulting transformation matrix, linking the camera and LiDAR, is computed within the SE(3) space.

The global position of the target is ascertained through the camera's intrinsic parameters, alongside the meticulously calibrated rotation matrix and translation vector. The target's image-based position is extracted using the YOLO image detection algorithm, while its 3D depth information is derived by correlating the clustered LiDAR point cloud data with the image detection algorithm's two-dimensional coordinates.

An EKF is employed to model the target's position over time, facilitating an analysis that yields insights into the object's velocity and acceleration. This information is instrumental in forecasting the target's trajectory and directional movement. The forecasted trajectory is then integrated into the planning algorithm as a constraining factor, optimizing the USV's own trajectory. Such an optimized approach ensures the USV's movement remains stable and responsive in dynamic environments, thereby enhancing the efficacy of target tracking.

In the planning layer of the USV Tracking system, an additional constraint is the orientation of the USV's viewing angle, specifically to accommodate the limited-view requirements of the perception system. This design ensures that the target is positioned as centrally as possible in the forward trajectory of the USV, maintaining stability and focus within the USV's field of vision.

This multi-sensor fusion system exhibits several distinct advantages:

1. **Extended Detection Range Compared to LiDAR-Only Systems:** Unlike systems reliant solely on LiDAR, where point cloud density decreases with distance, this integrated

system can detect targets at greater distances with reduced computational load. This is achieved by augmenting sparser LiDAR data with additional sensor inputs.

2. **Enhanced Depth Accuracy Over Vision-Only Systems:** Compared to purely vision-based systems, this multi-sensor approach yields more precise depth measurements, leading to more accurate predictions of target behavior and trajectory. The real-time obstacle map constructed by the system carries higher confidence levels, thanks to the integration of depth information from multiple sources.

The system's ability to predict target trajectories plays a crucial role in navigation, especially in obstacle-dense environments. Positioning the target in the central field of view, while maintaining a stable tracking distance, typically results in an arced travel path. Even in situations where obstacles obscure the line of sight between the USV and the target, the predicted trajectory allows for continued tracking planning for a certain duration until the target is visually reacquired. This capability ensures the USV's tracking path remains unaffected by temporary visual obstructions, showcasing the robustness and adaptability of the system in complex maritime scenarios.

Transitioning from the conceptual overview, the subsequent content explicates the transformation matrix formulas and their parameters, as referenced in Figure 3.4. These formulas are essential for the coordinate transformations crucial in multi-sensor data integration and accurate spatial analysis within the USV system.

$$f = \frac{I_w}{2\tan\left(\frac{\text{FOV}}{2}\right)} \tag{3.1}$$

$$R = I + \sin(\theta)K + (1 - \cos(\theta))K^2 \tag{3.2}$$

$$\mathbf{p}_{\text{norm}} = K^{-1}\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \tag{3.3}$$

$$\mathbf{p}_{\text{usv}} = \begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix}\mathbf{p}_{\text{norm}} \tag{3.4}$$

$$\mathbf{p}_{\text{point\_cloud}} = \text{SE}(3)\mathbf{p}_{\text{usv}} \tag{3.5}$$

$$\mathbf{p}_{\text{target\_global}} = R_\theta\mathbf{p}_{\text{usv}} + \mathbf{p}_{\text{usv\_global}} \tag{3.6}$$

$$R_\theta = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3.7}$$

Table 3.2: Parameters Utilized in Coordinate Transformation

Symbol	Description
f	Camera focal length
I_w	Width of the image
FOV	Field of view of the camera, in radians
R	Rotation matrix, computed using Rodrigues' rotation formula in simulator
θ	Magnitude of the rotation vector
K	Skew-symmetric matrix derived from the unit rotation vector
$\mathbf{p}_{\text{norm}}$	Normalized coordinates of image points in 3D space
$\mathbf{p}_{\text{usv}}$	Coordinates transformed to the USV
$\mathbf{p}_{\text{point\_cloud}}$	USV coordinates mapped to the LiDAR space
$\mathbf{p}_{\text{target\_global}}$	Target's global coordinates, computed using the observer's orientation and global position of the USV
R_θ	Rotation matrix representing the observer's orientation

3.3 Experiment Platform

The study presents a cohesive platform integrating simulation with physical experimentation. This integrated approach is crucial for achieving a seamless transition from simulated to real-world environments. The simulation, conducted on Gazebo, meticulously replicates the physical USV and its sensor setup, ensuring consistency across both domains. The physical USV, crafted through 3D printing, mirrors its simulated counterpart in design and functionality. Sensor configurations on the physical model are directly informed by the findings from the simulation, ensuring each component aligns precisely with the validated simulation parameters. This harmonized system facilitates the efficient and rational validation of perception and planning algorithms, embodying a comprehensive and effective approach to USV system development and testing.

3.3.1 Simulator Setup

In setting up the simulator, the focus is on utilizing Gazebo, an advanced simulation environment. This setup incorporates an open-source catamaran model, enhanced with realistic water surface rendering and environmental elements like trees and grass. Sensor parameters within Gazebo are meticulously configured to mirror real-world counterparts. The primary aim of this simulated environment is to emulate authentic maritime conditions, facilitating the development and validation of perception and planning algorithms for the USV.

3.3.1.1 Platform Comparisons

In the realm of autonomous systems research, the selection of an appropriate simulator is pivotal for the validation and testing of algorithms. Among the prominent simulators, Unity and Unreal Engine are distinguished for their exceptional capabilities in high-fidelity image rendering. These engines, primarily designed for the creation of visually immersive environments, are advantageous in contexts where photorealism is of paramount importance. However, their substantial computational requirements, a byproduct of their focus on visual realism, render them less suitable for projects where the emphasis lies on kinematic and dynamic feedback rather than on visual detail.

In contrast, the Robot Operating System (ROS), in conjunction with the Gazebo simulator, offers a more balanced approach, particularly beneficial for robotics applications. Gazebo's integration with ROS facilitates efficient inter-process communication and provides a comprehensive range of virtual sensors, including cameras, LiDAR, and position and orientation data simulation, analogous to GPS/IMU systems. This integration is instrumental in the field of robotics, where the simulation of sensor inputs and the accuracy of physical interactions are of greater significance than the intricacy of visual rendering.

Focusing on the domain of 3D perception algorithms for multi-sensor fusion, especially in the development of USV tracking systems, the necessity of a simulator that prioritizes dynamic feedback over visual fidelity becomes apparent. The bespoke simulator, developed on the Gazebo platform, addresses this specific requirement. It facilitates the effective testing of USV target tracking algorithms by providing essential sensory simulations without incurring the computational overhead associated with high-fidelity visual rendering. This strategic focus on kinematic modeling accuracy and sensor data integration, as opposed to detailed visual representation, renders the simulator not only efficient but also acutely relevant to the specific needs of USV tracking systems.

3.3.1.2 Gazebo Simulator Insights

The simulation environment is meticulously designed to facilitate the validation of perception and planning algorithms for USV. This platform aims to mirror real-world conditions, encompassing environmental factors, dynamic feedback, the USV's propulsion system, and an array of onboard sensors, all modeled with fidelity to their actual counterparts. Detailed aspects of this simulation are as follows:

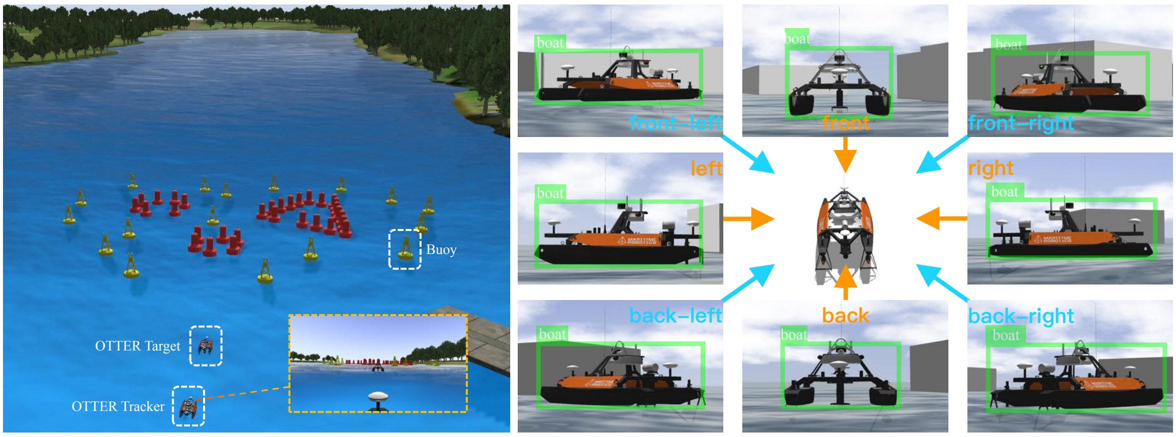

Figure 3.5: USV Simulator and Image-Based Orientation Prediction
Left: USV Simulator in Gazebo with Obstacle Targets and Camera Captured Images. Right: Orientation Prediction Through 8 Viewpoints.

1. **Dynamic Environmental Factors**. In the simulator designed for USV tracking, incorporating stochastic elements to model water wave disturbances and atmospheric wind perturbations is critical for achieving realism. For water surface waves, a stochastic function is utilized, where key adjustable parameters include amplitude, wavelength, phase speed, and directionality. These parameters collectively define the wave's height, frequency, and movement, allowing for a spectrum of sea conditions to be simulated. For atmospheric wind perturbations, a separate stochastic function is employed, characterized by parameters such as mean wind speed, turbulence intensity, gust factor, and wind direction variability. These parameters enable the simulation of wind's random and dynamic behavior, ranging from steady breezes to erratic gusts. Both functions' randomness is crucial for mimicking the unpredictability of real-world environmental conditions. By manipulating these parameters, the simulator can accurately reproduce a diverse range of challenging scenarios, thereby providing a robust platform for testing and optimizing USV tracking algorithms under realistic conditions.

2. **Catamaran Model**. The simulated vessel in the USV tracking system is an advanced open-source catamaran model, specifically selected for its superior dynamic stability. This stability is crucial in maintaining sensor alignment and reducing data distortion, particularly in rough sea conditions. The catamaran's structure allows for the strategic placement of multiple sensors, which are rigidly fixed to ensure consistent and accurate

data acquisition. These sensors, essential for navigation and environment perception, include cameras, LiDAR, and GPS/IMU systems, each meticulously integrated to replicate real-world USV sensor setups. Interactive control of the catamaran is enabled through keyboard inputs, allowing for precise maneuvering and facilitating the testing of various control algorithms under different scenarios. This feature enhances the versatility of the simulation, providing researchers with a hands-on approach to evaluate the USV's performance. Physically, the catamaran adheres to realistic environmental interactions. It is subjected to gravitational forces and buoyancy, closely mimicking the floating dynamics of a real vessel. The simulation also incorporates environmental factors such as wind and water waves, providing a comprehensive testbed for assessing the USV's stability and sensor effectiveness under varying conditions. This detailed and realistic simulation of the catamaran, with its dynamic stability, sensor integration, interactive control, and physical realism, forms a sophisticated platform for the development and validation of USV tracking algorithms, ensuring a high degree of fidelity and applicability to real-world maritime scenarios.

3. **Sensor Configuration**. Sensor configuration utilizes open-source methodologies to replicate real-world data acquisition. The camera sensor, designed to capture the 3D environment, operates by mapping all rigid bodies within the simulator to pixel values based on preset parameters. This process involves calculating the projection from each object to the camera's focal point, ensuring an accurate representation of the simulated environment. The camera's resolution is set to 1242×376 pixels in RGB format, mirroring the specifications of the widely recognized KITTI dataset. This choice facilitates comparability and validation against a standard benchmark in autonomous vehicle research. The camera's FOV is configured to 90° degrees by 60°, providing a comprehensive visual coverage. In parallel, a LiDAR sensor is implemented to generate point clouds, emulating the data acquisition process of real-world LiDAR systems. The LiDAR sensor in the simulator captures the spatial arrangement of all rigid bodies within its range, converting these into a point cloud format. This sensor is characterized by a 32-line array with a vertical FOV of 30°. The scanning mechanism of the simulated LiDAR replicates

the rotational scanning method used in actual LiDAR systems, ensuring realism in data collection. Random noise is integrated into the LiDAR data to mimic the inherent noise present in real sensor outputs. These sensor configurations within the simulator play a pivotal role in creating a realistic and robust environment for testing and validating USV tracking algorithms, ensuring that the simulated data closely approximates real-world sensor inputs.

Table 3.3: Detailed Specifications of the Simulation Platform

Component	**Detail**
Simulation Engine	Gazebo
GPU	NVIDIA GTX 3080ti
CPU	Intel i7-9600k
Obstacle Dimensions	$\sim$ 2m × 3m
USV Dimensions	$\sim$ 2m × 1m
LiDAR Type	32-line LiDAR Sensor
Camera Resolution	1241×376, RGB Format
Auxiliary Sensors	P3D (Incorporating IMU and GPS)
Operational Field Size	$\sim$ 200m × 100m

3.3.2 Physical USV Configuration

The physical USV is meticulously designed based on the dynamic model of the catamaran used in the simulation environment. The sensor platform of the physical USV closely mirrors that of the simulator, ensuring consistency in data and performance.

A Livox-mid360 LiDAR is selected for its wide FOV ranging from -7° to 52° and a high sampling rate of 200k points per second. As shown in Figure 3.6, the LiDAR is mounted with a slight downward tilt, while the camera is oriented slightly upward. This arrangement facilitates optimal data capture. Calibration between these two components is achieved using a checkerboard pattern, allowing for the precise determination of the SE(3) transformation matrix between their coordinate systems.

The hull of the USV is constructed through 3D printing technology, incorporating specialized sealing mechanisms to prevent water intrusion and protect the internal components. For en-

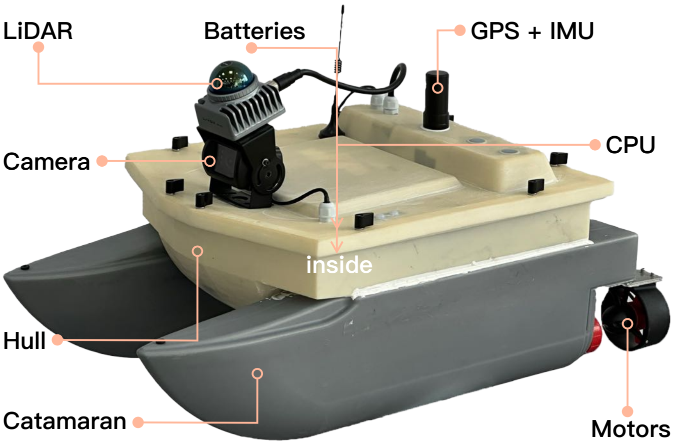

Figure 3.6: USV Hardware and Sensor Layout

Table 3.4: Detailed Specifications of the Physical USV

Component	**Detail**
CPU	Intel i7-1165G7@4.7GHz
Camera	USB camera, waterproof, night vision, 1920×1080, 30fps, FOV $80^\circ \times 60^\circ$
LiDAR	Livox-mid360, waterproof, FOV $360^\circ \times -7^\circ \times 52^\circ$, 200k pts/s
Motors	2 brushless motors, 180W
GPS	Ublox-zedf9p, rtk
IMU	Witmotion-hwt905
Autopilot	PX4
Max Speed	2.7m/s
Weight	$\sim$ 5kg
Duration	$\sim$ 35mins
Distance	$\leq$ 500m
Tracking Range	$\sim$ 7m

hanced positional accuracy, the GPS system is augmented with Real-Time Kinematic (RTK) technology. This careful design and integration of components ensure that the physical USV replicates the dynamics of its simulated counterpart, providing a reliable platform for real-world testing and validation of algorithms developed in the simulated environment.

3.3.3 3D Reconstruction for Dataset Creation

A comprehensive closed-loop data annotation pipeline is established to create a labeled dataset, essential for image detection tasks. This pipeline is designed for flexibility and efficiency in

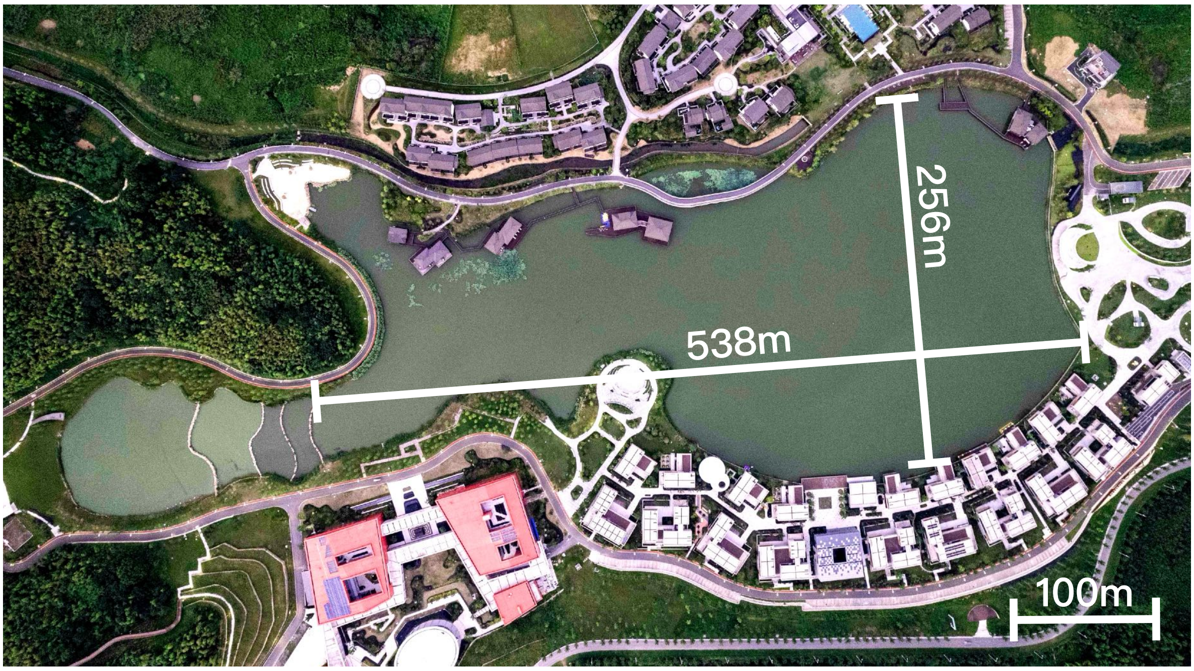

Figure 3.7: Overhead View of Experimental Site, Huzhou, Zhejiang

handling various targets, streamlining the process of generating a robust dataset for training purposes.

The process begins with the real-time collection of data for a specific target. Initial labeling is conducted using online large-scale image detection tools, providing an automated first pass of annotation. This is followed by a meticulous review and manual correction of labels using the `labelImg`[5] tool, ensuring accuracy and consistency across the dataset.

As depicted in Figure 3.8, the target model is then refined through dense 3D modeling techniques. The model is imported into `meshLab`[6], where manual adjustments and completions are made to enhance its fidelity. This refined model is subsequently integrated into the simulation environment, where simulated lighting conditions are applied to generate a dataset that closely mimics real-world scenarios.

The data generated through these two methods are combined, with a ratio of 8:2, to form the final dataset for training. This approach offers significant advantages in terms of adaptability; when switching to a different target, the pipeline allows for easy integration of new data. If

[5] The source code and additional documentation can be found in the GitHub repository at https://github.com/HumanSignal/labelImg

[6] For further information, please visit the MeshLab website at https://www.meshlab.net/

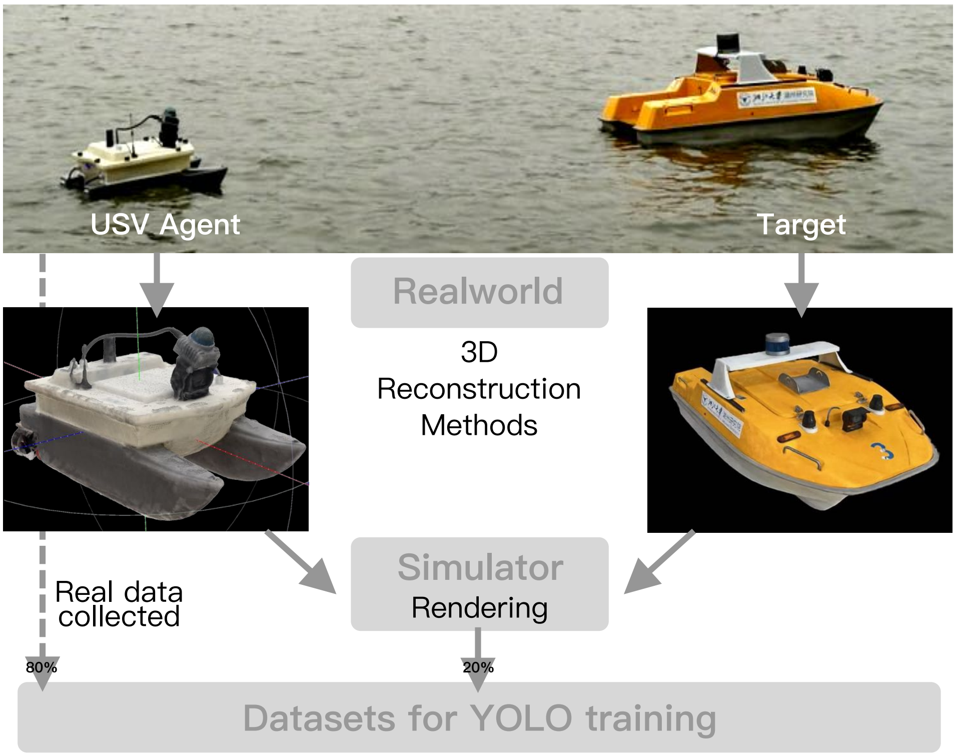

Figure 3.8: Dataset Creation in Simulator using 3D Reconstructed Models

dense 3D reconstruction of the new target is feasible, it can be seamlessly incorporated into the dataset using the same dual-method approach. Alternatively, short-term tracking data of the target can be annotated and added to the dataset. This closed-loop data annotation pipeline thus provides a highly versatile and efficient method for generating datasets, facilitating the easy interchange of targets within the USV tracking system.

CHAPTER 4. EXPERIMENT RESULTS

4.1 3D Reconstruction on Modeling and Building Obstacle Map

As depicted in Figure 4.1, the left panel showcases an indoor reconstruction result utilizing the `ARKit`[1] development tool integrated within an iPhone, combined with the NeuralRecon method for reconstruction. NeuralRecon[2] (Sun et al., 2021) is noted for its efficiency and speed in generating 3D models. In this instance, a rudimentary model of a USV is reconstructed. While the model is recognizable in form, it falls short in meeting the requirements for detailed image rendering, highlighting an area for enhancement in high-fidelity visual representation.

On the right, the FIEST[3] (Han et al., 2019) method is employed to construct a grid-map obstacle map, formatted in ESDF. This method proves effective in spatial planning within three-dimensional environments, fulfilling the specific requirements of dynamic obstacle mapping in USV navigation. The FIEST approach, known for its precision in distance measurement and efficient representation of space, contributes significantly to the accuracy of the obstacle map.

This experimental result primarily serves as a qualitative analysis of feasibility. The construction of the dynamic obstacle map meets the established criteria, demonstrating the practicality of the approach. However, the dense 3D reconstruction aspect, while functional, requires further refinement to achieve the desired level of detail and accuracy. Enhancing this component will significantly improve the overall quality and utility of the 3D models in simulating and planning for real-world navigational scenarios.

[1] ARKit website: https://developer.apple.com/augmented-reality/arkit/

[2] NeuralRecon project page: https://zju3dv.github.io/neuralrecon/

[3] FIEST code repository: https://github.com/HKUST-Aerial-Robotics/FIESTA

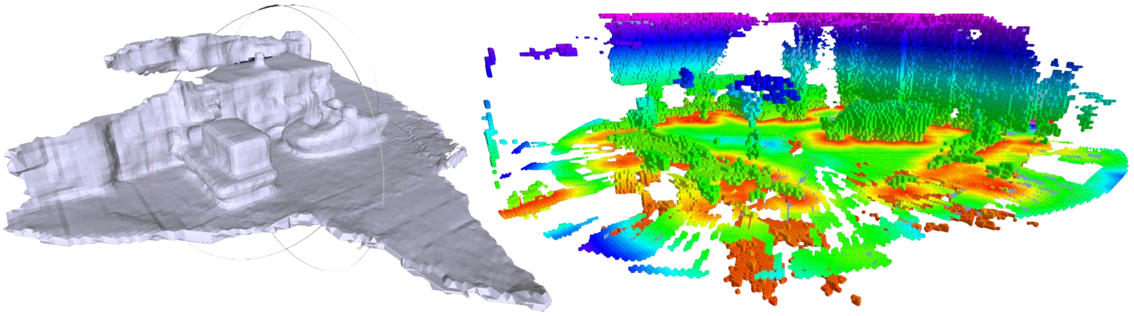

Figure 4.1: 3D Object Modeling (Left) and Obstacle Map Construction (Right)
Integration of 3D point clouds with IMU for accurate object modeling and obstacle mapping.

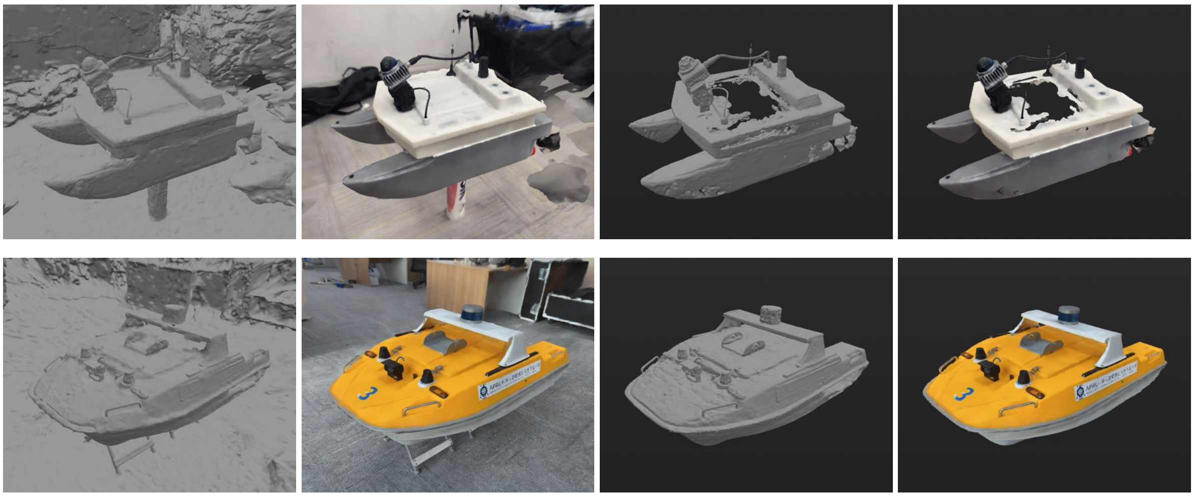

Figure 4.2: 3D USV Model (Upper row) and Target Model (Bottom row)

To augment the fidelity of dense 3D reconstruction, this study has utilized the advanced capabilities of the `Luma AI`[4] tool, grounded in the Neural Radiance Fields (NeRF) methodology. `Luma AI` excels in model reconstruction from an array of multi-view images, coupled with precise camera pose data. This potent combination enables the isolation of models from their backgrounds and facilitates the generation of detailed mesh files in `PLY` format, perfectly suited for simulator integration. As illustrated in Figure 4.2, the reconstructed models are presented in four distinct variations: both with and without background elements, and each either incorporating or omitting texture features. These variations collectively satisfy the stringent requisites for dense 3D reconstruction, thereby significantly enhancing the caliber of rendered image datasets, which are integral to the refinement of image detection tasks.

[4] Luma AI website: https://lumalabs.ai/

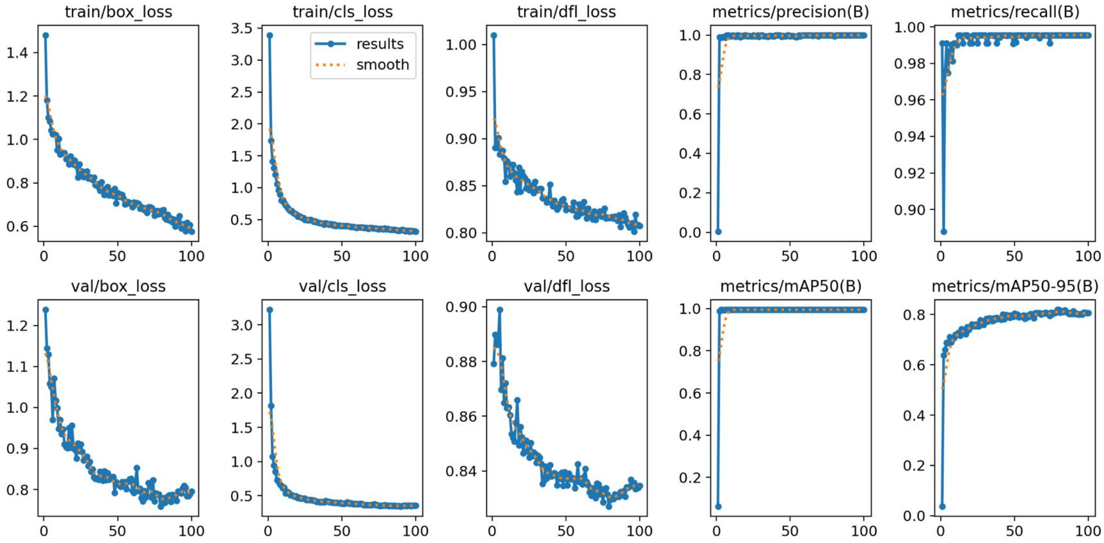

Figure 4.3: Comprehensive Training Metrics of YOLO Model

4.2 Target Detection Based on Multi-Sensor System

4.2.1 Visual Target Detection in 2D

Our YOLO model was trained on a custom dataset with the following parameters in Table 4.1.

Table 4.1: Parameters for YOLO Training

Parameter	**Value**
Epochs	100
Batch Size	16
Learning Rate	0.001, reduced by 0.1 every 30 epochs
Optimizer	Adam
Loss Function Weights	Classification: 1.0, Localization: 5.0, Confidence: 0.5
Data Augmentation	Random rotations, horizontal flips, scaling
Anchor Boxes	Customized for dataset object dimensions
Regularization	L2 ($\lambda = 0.0005$), Dropout (rate = 0.4)
Input Image Size	1241×376 pixels

The hardware setup for this training shown in Table 4.2.

This combination of training parameters and robust hardware facilitated effective and efficient training of the YOLO model, as evidenced by the model’s convergence depicted in Figure 4.3 of our study.

Table 4.2: Hardware Setup for YOLO Training

Component	Specification
GPUs	2 NVIDIA GTX 3080ti, 12 GB GDDR6X each
CPU	Intel i7-9600k @ 4.5GHz
Memory	64 GB RAM

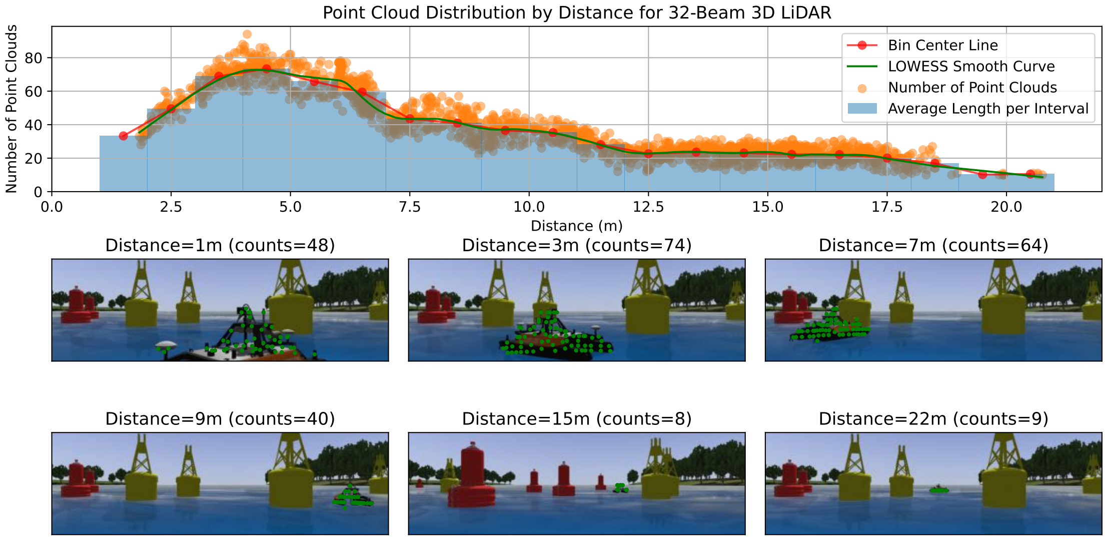

Figure 4.4: Numbers of Point Cloud with Different Depth

4.2.2 Target Localization in 3D

The sparsity of LiDAR data increases with distance, becoming more pronounced for rotating LiDAR systems where point cloud data on distant targets may appear random. In an experiment conducted in the simulation environment, as shown in Figure 4.4, the perception accuracy of a 32-line LiDAR for a 2m × 1m target was assessed, using the number of points on the target as the metric. It was observed that in BEV, objects with dense point clouds appeared rectangular (like the target USV), whereas sparsity in point clouds created ambiguity, making it challenging to distinguish between point obstacles and the target USV. Based on the trend of point cloud density with distance, a tracking distance of 7m was established as an optimal constraint for the planning algorithm. Maintaining the tracking within ±1m of this distance enhances perception accuracy and robustness in target tracking tasks.

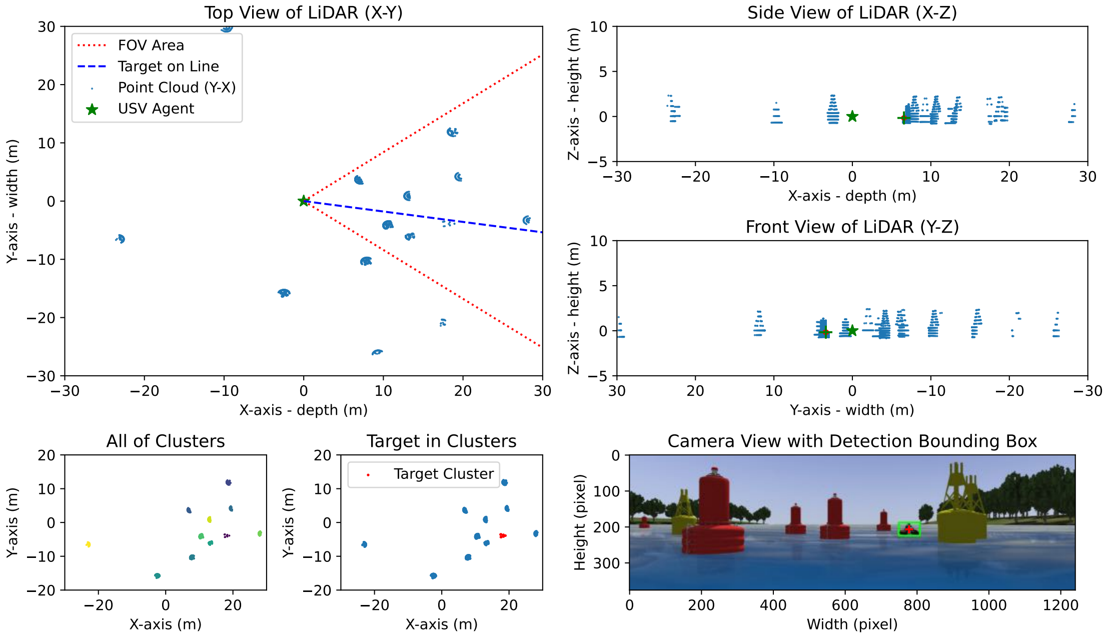

Figure 4.5: Pipeline of Point Cloud Clustering in BEV

LiDAR is renowned for its precise depth information, a feature that becomes particularly valuable in BEV. In this perspective, clustering techniques enable the separation of targets from obstacles, with accuracy bolstered through shape estimation. Known target shapes, discernible within a 7m range, contrast with typically circular, unknown obstacle shapes. Image-based YOLO detection provides 2D pixel coordinates of the target's center, corresponding to a 3D ray in space, illustrated by a blue dashed line in the Figure 4.5. This ray, unlikely to intersect a specific point cloud point, is generally matched by locating the nearest point. In BEV, after clustering the point clouds, the ray from target detection intersects a specific cluster, identifying the actual target. By extracting the point cloud of this cluster and estimating depth, the target's coordinates in the global coordinate system are obtained. Running this pipeline over a time series, and storing and analyzing both the USV's self-localization and target localization data, enhances the understanding of target behavior and USV navigation.

4.3 Trajectory Prediction

As illustrated in Figure 4.6, the average error in continuous target perception and localization was found to be 16 cm, significantly smaller than the size of the target itself, thus meeting

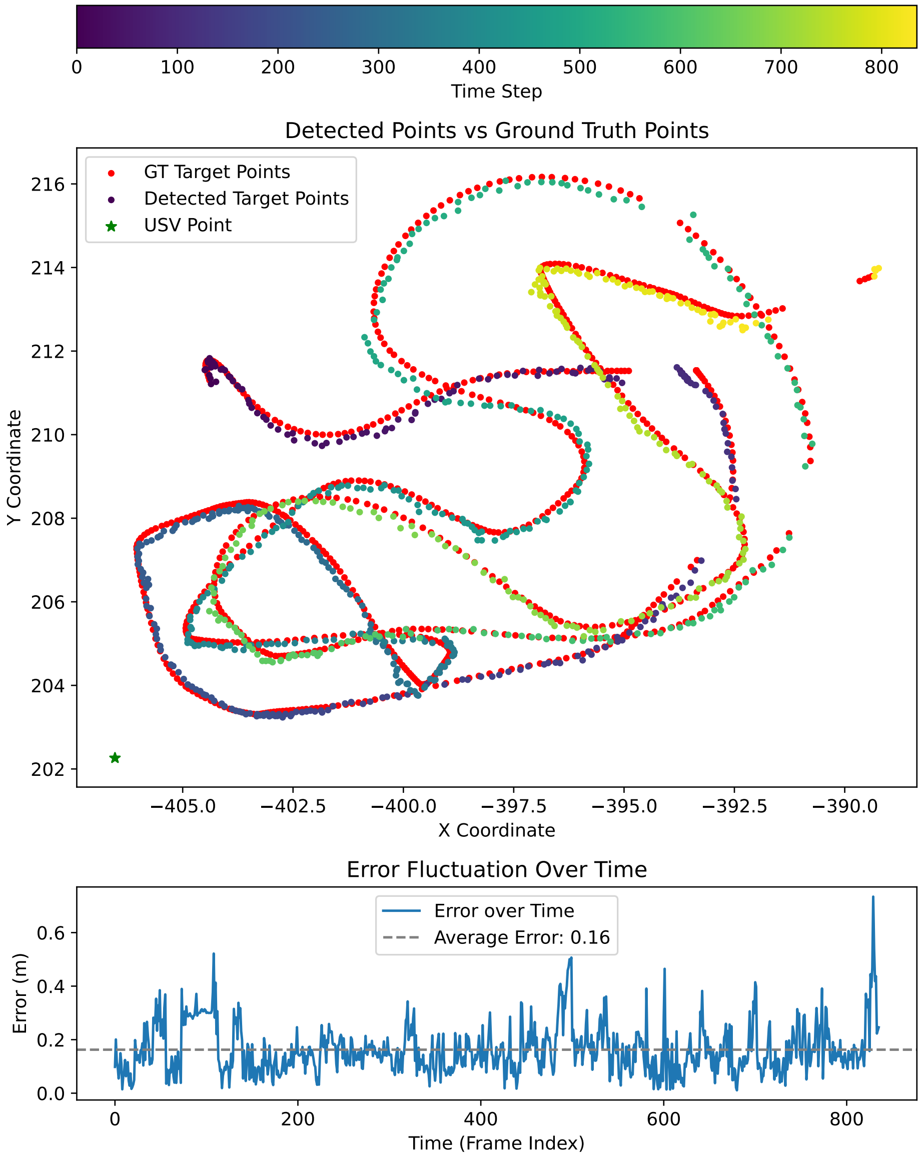

Figure 4.6: Target Detection in Continuous Time Series

the precision requirements for perception and localization. The study employed EKF, Linear Regression, and LSTM models to predict the time-series trajectory of the target. EKF, while effective, showed over 20 cm of localization error during target's rotational movements, leading to substantial trajectory prediction variations. Linear Regression, predicting the trajectory as the tangent of the current path, offered more accurate speed estimations. LSTM, trained on existing trajectory data, demonstrated superior capability in fitting the trajectory curve, as depicted in Figure 4.7. This comparative analysis of predictive models highlights the trade-offs between

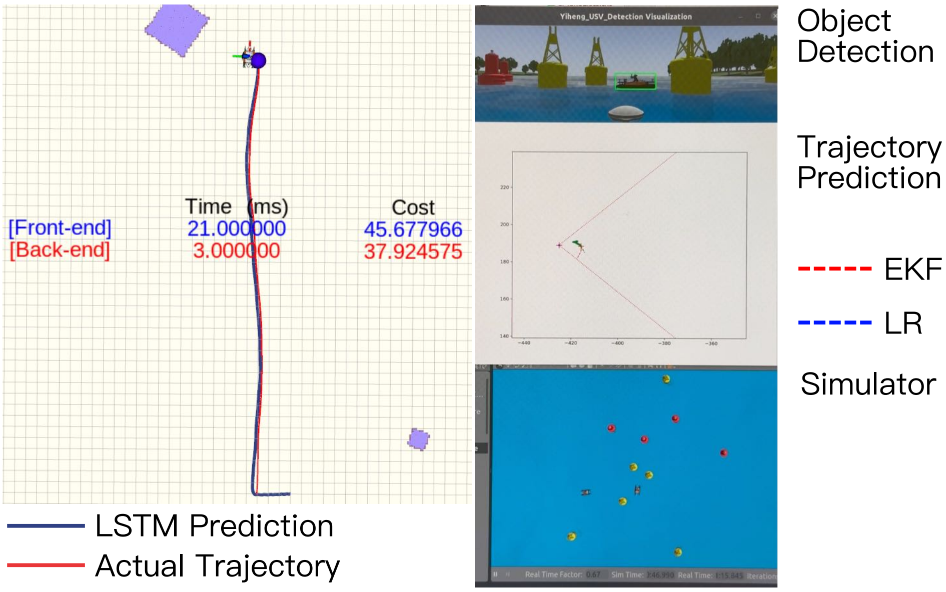

Figure 4.7: Trajectory Prediction Results

accuracy and responsiveness in dynamic target tracking, underscoring the need for selecting appropriate predictive methodologies based on the specific requirements of the tracking task.

4.4 USV Tracking with Planning Algorithms

In the final experimental stage, a stable target tracking system was developed by fusing visual target detection with 3D sensor data. This integration significantly enriched the planning process, which was the work of Tao HUANG[5], providing valuable 3D perception insights. The algorithm's feasibility and robustness in planning were initially verified through simulations and subsequently confirmed in real-world applications using a physical USV. As depicted in Figure 4.8, our system, in contrast to the baseline, incorporates a FOV constraint and maintains an optimal tracking distance of 7 meters, ensuring smooth and accurate execution of visual perception tasks. Even in instances where the target is temporarily obscured by obstacles, the trajectory information provided enables the USV to maintain consistent planning and control, a testament to Huang's planning algorithm. Compared to the Elastic Tracker (Ji et al., 2022), our method demonstrates notable improvements in stability. Importantly, this approach allows the USV Agent to independently execute target tracking tasks without the need for communi-

[5] Tao HUANG is PhD student at Zhejiang University, specializes in control algorithms for automation and mobile robotics, as well as in reinforcement learning. His contact information is 12332005@zju.edu.cn

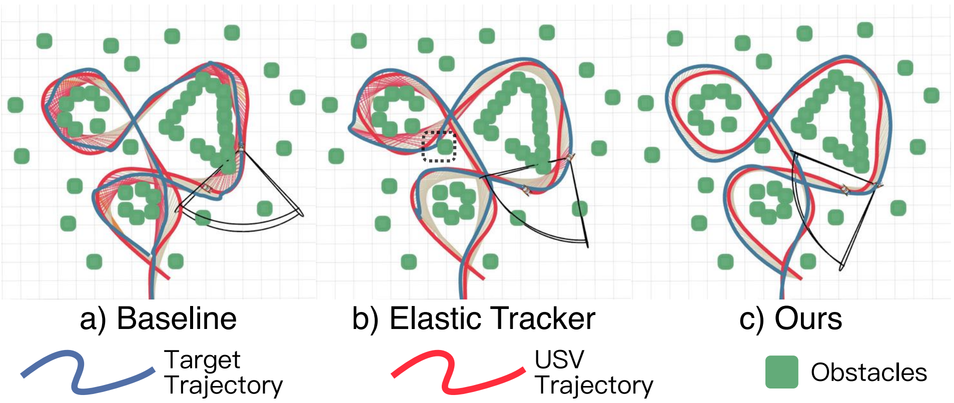

Figure 4.8: Comparative Experimental Results in Complex Aquatic Environment

cation or external information exchange with the target, showcasing stability and robustness in environments with dense obstacles.

4.5 Target Tracking in Real-World

The primary objective of the real-world experiment was to validate the feasibility of the proposed solution, particularly emphasizing the interconnection between perception tasks and planning/control tasks within the experimental framework. The planning aspect of the experiment was conducted in collaboration with Tao HUANG. The results of the perception and trajectory prediction played a pivotal role in the overall success of the task. Notably, this experiment was carried out in challenging conditions characterized by wind and wave disturbances, underscoring the robustness and adaptability of the approach in dynamic real-world environments.

Figure 4.9 presents a sequential depiction of a target tracking task conducted in a real-world setting, arranged chronologically from left to right across five columns. The topmost row provides an aerial perspective of the experimental setup, with each image incorporating a small inset in the lower-left corner. This inset displays the visual data captured from the camera mounted on a white unmanned vessel, which is employed for intricate perception and localization tasks. The middle row offers a lateral view of the experimental arena, captured through aerial drone photography, showcasing a more granular understanding of the spatial

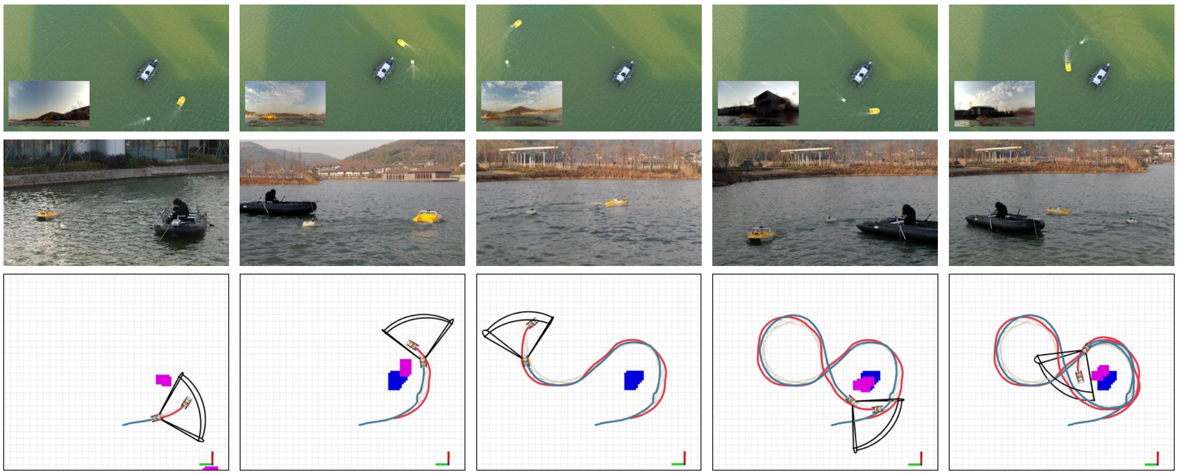

Figure 4.9: Target Tracking in Real-World

arrangement and positional dynamics within the scene. The final row at the bottom portrays the construction of a dynamic obstacle map, featuring a grid resolution of 1 meter for enhanced precision. Within the scope of this experiment, a yellow boat is designated as the target for tracking purposes, while a black kayak, with dimensions of 4 meters in length and 1.5 meters in width, is strategically positioned to serve as a dynamic obstacle, adding complexity to the tracking task.

Chapter 5. Analysis

The experiments conducted in this study affirm the significance of 1D sensors, particularly IMU and GPS. The IMU, operating at a frequency of 100Hz (up to 200Hz in Livox-mid360), is crucial for recording the motion posture of mobile robots, providing essential data for correcting motion distortions in other sensors. The study also highlights the richness of algorithmic choices available for 2D image sensors. Compared to point cloud processing, these algorithms require less computational power, handle various complex environmental challenges, and offer richer target features in images for accurate identification, making visual detection essential in dynamic environments.

Furthermore, the reliability of depth measurements from 3D LiDAR stands out. Unlike depth estimations from vision, the LiDAR's depth data, obtained through direct measurements, offers higher accuracy and credibility. This aligns with the proposed Information Conservation concept, which treats sensor operating frequency and information retention as a trade-off. Thus, our system integrates 3D LiDAR, 2D cameras, IMU, and GPS for comprehensive perception.

The simulation environment developed mirrors the physical feedback characteristics of real-world settings, facilitating realistic emulation. This environment was instrumental in validating the multi-sensor platform's perception performance, which matched the physical counterpart in precision. The integration of this perception method with planning algorithms in the simulator significantly reduced experimental workload and enhanced algorithm development efficiency, demonstrating the effectiveness of our unified simulation and physical USV platform in streamlining the research process. Our perception platform successfully maintains stable target tracking within a 7-meter range, as evidenced in Figure 5.1, adeptly keeping the target centered in the field of view with consistent tracking distance. This exemplary performance is attributed

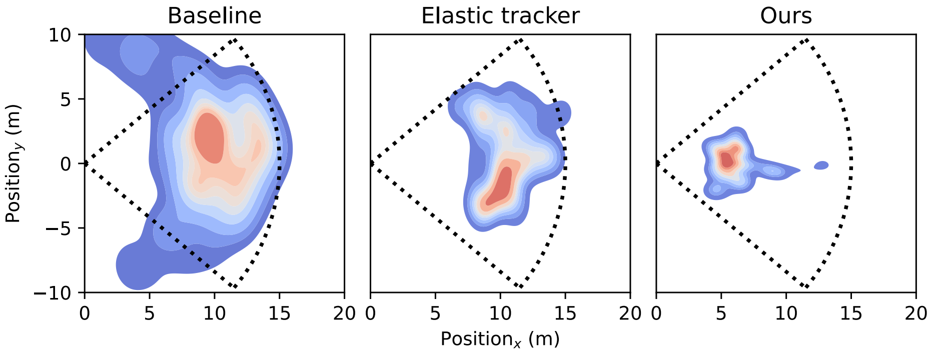

Figure 5.1: The Distribution of the Target Positions Relative to the USV on X-Y Plane

not only to the high accuracy and robustness of the perception platform but also to precise trajectory predictions, a vital input for the planning algorithm. Predicting the target's motion, direction, speed, and trajectory enables effective navigational planning in obstacle-dense environments, ensuring continued motion planning even during brief target loss, until the target's 3D coordinates are reacquired.

In real-world applications, the platform demonstrates feasibility by effectively preprocessing point cloud data to eliminate surface noise, creating accurate dynamic obstacle maps. This robust and stable perception platform in physical settings can be adapted to incorporate more precise and high-frequency visual detection devices, such as infrared cameras for night vision and multi-focal length camera arrays for long-distance tracking. These devices can be integrated with LiDAR depth information in BEV to precisely and continuously track the target's 3D location.

Chapter 6. Discussion

This research critically examines the characteristics and application scenarios of various sensors in USV technology, addressing both general and specific challenges in 3D perception and multi-sensor fusion. A key finding is the crucial role of high-frequency sensors, especially IMUs, in dynamic movements for posture correction in USVs. Given the limitations of edge computing, this study innovatively adapted target detection techniques to feature-rich images, circumventing the challenges posed by sparse point cloud data and the complexity of aquatic surface features.

In addressing the precise question of USV target tracking, our multi-sensor system demonstrated its capability to efficiently track targets in environments dense with obstacles, without necessitating direct communication with the target. This breakthrough holds immense potential for both military and commercial applications, such as in unmanned maritime shows and aerial marine photography. Importantly, this research lays a robust foundation for future explorations in USV swarm formation, a critical step towards solving collective operation challenges in autonomous maritime systems.

Looking ahead, several avenues for advancement have been identified:

1. **Point Cloud Algorithm Enhancement:** Future efforts will be directed towards refining preprocessing of point cloud data. By leveraging time-series information, the aim is to filter out water surface noise more effectively, thus enhancing the accuracy of the data. This advancement is crucial in developing more sophisticated and accurate models for dynamic maritime environments.

2. **Image Detection Algorithm Optimization:** While current algorithms leverage time-series data for accuracy and computational efficiency, there is a growing need to explore more lightweight algorithms, such as channel-separated MobileNet and ShuffleNet series. Single-channel image detection methodologies will be a focal point, aiming to enhance semantic segmentation of overexposed pixels and water stain removal. These developments are essential for robust target detection, particularly under extreme environmental conditions.

3. **Physical USV Redesign:** To enhance stability, there is a plan to redesign the physical USV, potentially enlarging its size. The sensor suite may be expanded to include higher-resolution infrared cameras, adapting to diverse environments and ensuring effective target tracking.

4. **Engineering Challenges in Network Integration:** A key challenge lies in the integration of various tasks into a cohesive end-to-end network. Current limitations, such as the independence of modules in the pipeline and the reliability requirements for control-related tasks, will be addressed. The exploration of neural networks, particularly CNNs and Transformers, will be pivotal in balancing computational efficiency and accuracy for complex multi-task operations.

5. **Advanced Sensor Research:** The study will investigate the potential of highly sensitive flexible pressure sensors to assess water flow conditions beneath the USV. This innovative approach could revolutionize the vessel's control system, enhancing navigation precision through advanced PID methods. Additionally, this data could provide invaluable environmental feedback for reinforcement learning algorithm research, contributing significantly to the field's advancement.

These future directions not only promise to enhance the current state of USV technology but also align with the forefront of research in intelligent robotics and autonomous navigation systems. The integration of these advancements will undoubtedly contribute to a new era of maritime operations, driven by sophisticated autonomous technologies.

Chapter 7. Conclusions

This thesis represents a comprehensive exploration into the realm of 3D perception and multi-sensor fusion within the context of USVs. The journey embarked upon in these pages has led to significant strides in the operational capabilities and autonomy of USVs, marking a notable contribution to the field of intelligent robotics and autonomous navigation.

Central to this exploration was the development and implementation of advanced perception algorithms and the integration of a diverse array of sensors. The successful application of the YOLO algorithm, ESDF method, 3D reconstruction techniques, and a hybrid approach combining EKF with LSTM networks has culminated in a robust, efficient, and accurate system for target detection, obstacle mapping, and trajectory prediction. This system represents a leap forward in USV technology, enabling these vessels to operate more autonomously and effectively in dynamic and challenging maritime environments.

The creation of a sophisticated Gazebo-based USV simulation platform and the corresponding physical sensory system for a catamaran have been pivotal in validating the effectiveness of these methodologies. By seamlessly integrating these systems within the framework of edge computing, the research has optimized multi-sensor fusion algorithms, enhancing the real-time decision-making capabilities of USVs. This has not only improved the precision of target tracking and obstacle avoidance but also ensured the stability and reliability of USV operations, even in complex scenarios.

The comparative analysis of sensors and the strategic recommendations for perception platforms have provided valuable insights into optimizing sensor fusion, thereby refining the overall perception accuracy in USVs. The sim-to-real approach adopted in this thesis has been instrumental in bridging the gap between simulation and real-world application, ensuring the scalability and practicality of the proposed solutions.

The advancements made in multi-USV systems and swarm planning underscore the potential for these technologies in coordinated maritime missions. This thesis lays the groundwork for future research in swarm intelligence, expanding the operational scope of USVs and opening new avenues for complex, multi-agent maritime operations.

In conclusion, this thesis not only addresses the specific challenges in USV target tracking but also contributes a versatile and scalable framework for 3D perception and multi-sensor fusion in intelligent robotics. The findings and methodologies presented here not only enhance the current state of USV technology but also offer a blueprint for future innovations in autonomous systems, paving the way for a new era in intelligent maritime operations.

References Cited

Bai, N., Xue, Y., Chen, S., Shi, L., Shi, J., Zhang, Y., Hou, X., Cheng, Y., Huang, K., Wang, W. et al. (2023), 'A robotic sensory system with high spatiotemporal resolution for texture recognition', *Nature Communications* **14**(1), 7121.

Bertinetto, L., Valmadre, J., Henriques, J. F., Vedaldi, A. and Torr, P. H. S. (2016), 'Fully-convolutional siamese networks for object tracking', *arXiv preprint arXiv:1606.09549* .

Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G. and Beijbom, O. (2020), nuscenes: A multimodal dataset for autonomous driving, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)', IEEE, pp. 11621–11631.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A. and Zagoruyko, S. (2020), End-to-end object detection with transformers, *in* 'European Conference on Computer Vision (ECCV)', Springer, pp. 213–229.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. and Yuille, A. L. (2017), 'Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs', *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* **40**(4), 834–848.

Chen, X., Ma, H., Wan, J., Li, B. and Xia, T. (2017), Multi-view 3d object detection network for autonomous driving, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)', IEEE, pp. 1907–1915.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T. and Ronneberger, O. (2016), '3d u-net: Learning dense volumetric segmentation from sparse annotation', pp. 424–432.

Fang, Z., Zhou, S., Cui, Y. and Scherer, S. (2021), '3d-siamrpn: An end-to-end learning method for real-time 3d single object tracking using raw point cloud', *arXiv preprint arXiv:2108.05630* .

Geiger, A., Lenz, P. and Urtasun, R. (2012), Are we ready for autonomous driving? the kitti vision benchmark suite, *in* 'Proceedings of the IEEE International Conference on Computer Vision (ICCV)', IEEE, pp. 3354–3361.

Girshick, R. (2015), Fast r-cnn, *in* 'Proceedings of the IEEE International Conference on Computer Vision (ICCV)', IEEE, pp. 1440–1448.

Han, L., Gao, F., Zhou, B. and Shen, S. (2019), Fiesta: Fast incremental euclidean distance fields for online motion planning of aerial robots, *in* '2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)', IEEE, pp. 4423–4430.

Huang, X., Wang, P., Cheng, X., Zhou, D., Geng, Q. and Yang, R. (2018), 'The apolloscape open dataset for autonomous driving and its application', *arXiv preprint arXiv:1803.06184* .

Jain, S., Xiong, B. and Grauman, K. (2017), Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)', IEEE.

Ji, J., Pan, N., Xu, C. and Gao, F. (2022), Elastic tracker: A spatio-temporal trajectory planner for flexible aerial tracking, *in* '2022 International Conference on Robotics and Automation (ICRA)', IEEE, pp. 47–53.

Ku, J., Mozifian, M., Lee, J., Harakeh, A. and Waslander, S. L. (2018), Joint 3d proposal generation and object detection from view aggregation, *in* 'IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)', IEEE, pp. 1–8.

Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Qiao, Y. and Dai, J. (2022), Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers, *in* 'European Conference on Computer Vision (ECCV)', Springer, pp. 1–18.

Liao, M., Lu, F., Zhou, D., Zhang, S., Li, W. and Yang, R. (2020), 'Dvi: Depth guided video inpainting for autonomous driving', *arXiv preprint arXiv:2007.08854* .

Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D. L. and Han, S. (2023), Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation, *in* 'IEEE International Conference on Robotics and Automation (ICRA)', IEEE, pp. 2774–2781.

Ma, Y., Zhu, X., Zhang, S., Yang, R., Wang, W. and Manocha, D. (2019), Trafficpredict: Trajectory prediction for heterogeneous traffic-agents, *in* 'Proceedings of the AAAI Conference on Artificial Intelligence'.

Nam, H. and Han, B. (2015), 'Learning multi-domain convolutional neural networks for visual tracking', *arXiv preprint arXiv:1510.07945* .

Philion, J. and Fidler, S. (2020), Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d, *in* 'European Conference on Computer Vision (ECCV)', Springer, Glasgow, UK, pp. 194–210.

Qi, C. R., Litany, O., He, K. and Guibas, L. J. (2019), Deep hough voting for 3d object detection in point clouds, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)', IEEE/CVF, pp. 9277–9286.

Qi, C. R., Su, H., Mo, K. and Guibas, L. J. (2017), Pointnet: Deep learning on point sets for 3d classification and segmentation, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)', IEEE, pp. 652–660.

Qi, C. R., Yi, L., Su, H. and Guibas, L. J. (2017), 'Pointnet++: Deep hierarchical feature learning on point sets in a metric space', *Advances in Neural Information Processing Systems (NeurIPS)* **30**.

Redmon, J., Divvala, S., Girshick, R. and Farhadi, A. (2016), You only look once: Unified, real-time object detection, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)', IEEE, pp. 779–788.

Ren, S., He, K., Girshick, R. and Sun, J. (2015), 'Faster r-cnn: Towards real-time object detection with region proposal networks', *Advances in Neural Information Processing Systems (NeurIPS)* **28**.

Ronneberger, O., Fischer, P. and Brox, T. (2015), 'U-net: Convolutional networks for biomedical image segmentation', pp. 234–241.

Shi, S., Guo, C., Jiang, L., Wang, Z., Shi, J., Wang, X. and Li, H. (2020), Pv-rcnn: Point-voxel feature set abstraction for 3d object detection, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)', IEEE, pp. 10529–10538.

Sun, J., Xie, Y., Chen, L., Zhou, X. and Bao, H. (2021), Neuralrecon: Real-time coherent 3d reconstruction from monocular video, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)', IEEE/CVF, pp. 15598–15607.

Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B. et al. (2019), 'Scalability in perception for autonomous driving: Waymo open dataset', *arXiv preprint arXiv:1912.04838* .

Wang, S., Sun, Y. and Liu, C. e. a. (2020), 'Pointtracknet: An end-to-end network for 3-d object detection and tracking from point clouds', *IEEE Robotics and Automation Letters* **5**(2), 3206–3212.

Wojke, N., Bewley, A. and Paulus, D. (2017), 'Simple online and realtime tracking with a deep association metric', *arXiv preprint arXiv:1703.07402* .

Xu, D., Anguelov, D. and Jain, A. (2018), Pointfusion: Deep sensor fusion for 3d bounding box estimation, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)', IEEE, pp. 244–253.

Zhou, Y. and Tuzel, O. (2018), Voxelnet: End-to-end learning for point cloud based 3d object detection, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)', IEEE, pp. 4490–4499.