INTELLIGENT PERCEPTION SYSTEMS FOR MOBILE ROBOT: FROM SEMANTIC-AWARE PLANNING TO HYBRID QUANTUM-CLASSICAL VIEW OPTIMISATION

XIAOTONG YU

PhD

The Hong Kong Polytechnic University

2025

The Hong Kong Polytechnic University
Department of Computing

Intelligent Perception Systems for Mobile Robot: From Semantic-aware Planning to Hybrid Quantum-Classical View Optimisation

Xiaotong Yu

A thesis submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy
May 2025

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgment has been made in the text.

Signature: _____

Name of Student: Xiaotong Yu

Abstract

Intelligent perception represents a critical challenge for mobile robotic systems operating in complex, unknown environments. Current approaches face fundamental limitations in semantic understanding, optimisation quality for viewpoint selection, and robust sensing under variable conditions. This thesis investigates these challenges and proposes novel solutions to enhance the perceptual capabilities of autonomous mobile robots.

The research is motivated by three key observations: first, traditional Next-Best-View planning algorithms typically optimize for geometric coverage without considering semantic significance, resulting in inefficient exploration when specific objects hold particular importance; second, classical optimization methods for viewpoint selection often converge to suboptimal solutions due to the vast, high-dimensional solution space; and third, the predominant reliance on RGB imagery limits robustness in challenging lighting conditions and raises privacy concerns in sensitive applications.

To address these challenges, this thesis presents three complementary contributions. The first introduces a semantic-aware Next-Best-View (S-NBV) framework that incorporates semantic information alongside visibility metrics in a unified information gain formulation, enabling efficient search-and-acquisition manoeuvres. Experimental validation demonstrates up to 27.46% enhancement in region-of-interest reconstruction efficiency compared to state-of-the-art methods.

The second contribution develops a Hybrid Quantum-Classical Next-Best-View (HQC-

NBV) framework that leverages quantum computing principles to more effectively navigate the complex solution space of viewpoint selection. Using a novel Hamiltonian formulation and bidirectional entanglement patterns, this approach achieves up to 49.2% higher exploration efficiency than classical methods, establishing a pioneering connection between quantum computing and robotic perception.

The third contribution presents the Cross Shallow and Deep Perception Network (CSDNet), a lightweight architecture designed for integrating low-coherence depth and thermal modalities. Through spatial information prescreening, implicit coherence navigation, and Segment Anything Model (SAM)-assisted encoder pre-training, CSDNet achieves performance comparable to triple-modality (RGB-D-T) methods while using only depth and thermal data, and reducing computational requirements by orders of magnitude, demonstrates that effective integration of low-coherence modalities can achieve robust perception in challenging conditions without relying on RGB data, offering both efficiency and inherent privacy advantages.

Extensive experiments across diverse scenarios validate the effectiveness of these approaches. The research advances the capabilities of robotic perception systems, enabling more intelligent exploration through semantic awareness, superior viewpoint selection through hybrid quantum-classical optimisation, and robust operation in challenging environmental conditions through effective multi-modal integration.

Keywords: Robotic Perception, Next-Best-View Planning, Semantic-Aware View Planning, Quantum Variational Algorithm, Hybrid Quantum-Classical View Planning, Low-Coherence Modality Integration

Publications Arising from the Thesis

1. Xiaotong Yu and Chang Wen Chen, "Semantic-aware Next-Best-View for Multi-DoFs Mobile System in Search-and-Acquisition based Visual Perception", in *Proceedings of the 32nd ACM International Conference on Multimedia. 2024.*

2. Xiaotong Yu and Chang Wen Chen, "HQC-NBV: A Hybrid Quantum-Classical View Planning Approach", manuscript submitted to *International Conference on Computer Vision, ICCV 2025.*

3. Xiaotong Yu, Ruihan Xie, Zhihe Zhao, and Chang Wen Chen, "CSDNet: Detect Salient Object in Depth-Thermal via A Lightweight Cross Shallow and Deep Perception Network", manuscript submitted to *IEEE Transactions on Multimedia.*

Acknowledgments

My heartfelt appreciation goes to my country.

I would like to express my deep gratitude to my supervisor, Professor Chang Wen Chen, for his invaluable guidance, unwavering support, and insightful mentorship throughout my PhD journey. The freedom he provided to explore innovative research directions, while offering critical feedback at key moments, has made this thesis possible.

I am particularly grateful to Ms. Cynthia Cheng from the Student Affairs Office, whose consistent support and guidance throughout my four years of study have been invaluable. Her encouragement and practical assistance have helped me navigate both academic and personal challenges with greater confidence. Her help has made a profound difference in my doctoral experience.

I owe an immense debt of gratitude to my family for their unconditional love, support, and patience throughout this journey. To my parents, whose support and encouragement have been the bedrock of my academic pursuits—thank you for believing in me even when the path seemed uncertain. Their constant support, both emotional and practical, has sustained me through the most difficult phases.

Special thanks go to the Hong Kong Polytechnic University and the Department of Computing, as well as the colleagues at the general office, who have provided detailed instructions and assistance at every step throughout these four years.

Finally, I am grateful to all those friends who have supported me throughout this journey, whether through technical discussions, emotional support, or simply being there when needed.

Thankful for the trust bestowed upon me by my nation.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The Intelligent Perception in Robotics

In the rapidly evolving landscape of robotics, intelligent perception represents the foundational capability that bridges the gap between autonomous systems and their effective operation in real-world environments. Perception—the ability to sense, interpret, and understand the surrounding world—forms the critical first stage in the perception-planning-control pipeline (as is shown in Figure 1.1) that governs robotic behaviour. Without robust and comprehensive perception, even the most sophisticated planning algorithms and precise actuation mechanisms remain fundamentally limited in their utility and application scope.

The significance of intelligent perception becomes particularly evident in challenging scenarios such as search and rescue operations, autonomous exploration of unknown environments, and human-robot interaction systems. In disaster response situations, for instance, mobile robots must rapidly develop accurate environmental representations while identifying victims, assessing structural stability, and planning safe traversal paths—all under severe time constraints and in environments characterised by visual degradation, structural irregularities, and possible dynamic conditions. Such

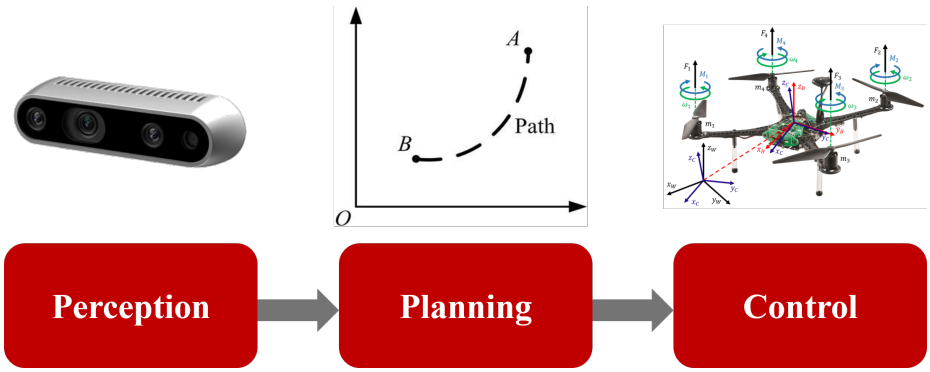

Figure 1.1: A conceptual diagram showing the perception-planning-control cycle in robotics, highlighting how perception feeds into decision-making and actions.

scenarios demand perception systems that transcend mere geometric reconstruction to incorporate semantic understanding, uncertainty management, and adaptive viewpoint selection.

Current approaches to robotic perception face several fundamental limitations. First, traditional Next-Best-View planning algorithms typically optimise for geometric coverage or information gain without considering the semantic significance of observed elements, resulting in inefficient exploration trajectories when specific objects or regions hold particular importance. Second, classical approaches for viewpoint selection often struggle to navigate the complex, high-dimensional solution space effectively, frequently resulting in suboptimal solutions due to their reliance on heuristics or limited sampling approaches. Third, the predominant reliance on RGB imagery creates significant vulnerabilities to lighting conditions, with performance degrading substantially in low-light, high-contrast, or variable illumination environments commonly encountered in real-world applications such as search and rescue operations, industrial inspection, and round-the-clock monitoring.

Moreover, the integration of perception and planning remains a significant challenge, with most systems treating these as separate, loosely coupled modules rather than deeply integrated components. This separation often results in exploration strategies that fail to leverage the rich semantic information available in the scene, leading to

inefficient utilisation of limited sensing resources and computational capabilities. As autonomous systems become increasingly deployed in complex human environments, the need for perception systems that are simultaneously semantically aware, able to find optimal solutions to complex planning problems, and robust to challenging environmental conditions becomes increasingly critical.

These challenges motivate the development of a new generation of intelligent perception approaches that can operate effectively under the constraints of mobile robotics while providing richer environmental understanding. By addressing the limitations in semantic integration, solution optimality in planning, and multi-modal fusion under challenging conditions, such approaches can significantly advance the capabilities of autonomous systems across diverse application domains. The work presented in this thesis aims to address precisely these challenges through three complementary research directions: semantic-aware planning for targeted exploration, hybrid quantum-classical optimisation for superior viewpoint selection, and robust multi-modal perception leveraging depth and thermal sensing.

1.2 Research Gaps and Challenges

Building upon the fundamental limitations of current robotic perception systems described previously, this section identifies specific research gaps that motivate the contributions presented in this thesis. Despite significant advances in robotic perception over the past decade, several critical challenges remain unaddressed at the intersection of semantic understanding, optimal viewpoint solution in robot view planning, and low-coherence modal integration for robust perception in challenging environments.

The semantic understanding gap represents perhaps the most significant limitation in current view planning approaches. While state-of-the-art Next-Best-View (NBV) algorithms have demonstrated impressive capabilities in geometric reconstruc-

tion and exploration of unknown environments [12, 84], they operate primarily on low-level geometric representations such as occupancy grids or Truncated Signed Distance Fields (TSDFs). These approaches typically define information gain solely in terms of visibility metrics—the number of unknown voxels that can be observed from a candidate viewpoint or the expected entropy reduction in the environmental map. Such formulations fundamentally fail to incorporate the semantic significance of different scene elements, leading to exploration trajectories that are efficient in covering space but inefficient in acquiring information about objects of interest. This limitation becomes particularly problematic in applications such as search and rescue operations, where rapidly locating and thoroughly examining specific objects (e.g., victims, hazardous materials) takes precedence over complete environmental mapping.

A related **solution optimality gap** emerges when considering the inherent complexity of view planning optimisation. The selection of optimal viewpoints represents a combinatorial optimisation problem with a vast solution space, particularly in scenarios involving multiple degrees of freedom in sensing platform mobility. Current approaches predominantly rely on sampling-based methods or heuristics that, while computationally tractable, often converge to suboptimal solutions due to their inherent limitations in navigating complex parameter interdependencies. More sophisticated classical optimisation methods typically struggle with the high dimensionality and non-convexity of the solution space. This fundamental limitation in the optimisation approach has remained a persistent challenge, restricting the exploration efficiency of autonomous systems in complex environments. Recent advances in quantum computing suggest potential pathways to address this optimisation challenge through inherently different computational paradigms, but their application to robotic perception remains largely unexplored.

The third significant research gap concerns **low-coherence modal integration for robust perception** in robotic systems. Contemporary perception systems predominantly rely on RGB cameras, which provide rich textural and appearance informa-

tion but perform poorly in challenging lighting environments such as darkness, glare, or highly variable illumination. Alternative sensing modalities such as depth cameras and thermal sensors offer potential robustness to these conditions, but integrating these low-coherence modalities—which exhibit reduced correlation in information content—presents significant technical challenges. While multi-modal perception systems have been extensively studied [89, 18], current approaches typically use RGB as the primary modality and struggle to efficiently integrate depth and thermal data without substantial computational overhead. The effective integration of these low-coherence modalities is particularly crucial for applications requiring operation in extreme lighting conditions, such as search and rescue operations and 24-hour surveillance. Moreover, these alternative modalities offer inherent privacy advantages as an additional benefit. The development of lightweight, efficient fusion architectures that can leverage complementary information from these low-coherence modalities represents an important open challenge, particularly for resource-constrained mobile platforms that must operate reliably across diverse environmental conditions.

These three gaps—in semantic awareness, solution optimality, and low-coherence modal integration—are deeply interconnected. For instance, incorporating semantic information into view planning increases the complexity of the solution space, while alternative modalities like depth and thermal that offer robustness to lighting variations often provide less direct semantic information than RGB imagery. Addressing these challenges requires novel approaches across algorithm design, optimisation frameworks, and sensor fusion architectures. The research presented in this thesis aims to bridge these gaps through complementary contributions that collectively advance the state of the art in intelligent perception for mobile robotic systems operating in complex environments.

1.3 Research Objectives and Contributions

In response to the critical research gaps identified in semantic understanding, solution optimality in view planning, and low-coherence modal integration for robotic perception, this thesis presents three complementary approaches designed to advance intelligent perception systems for mobile robots operating in complex environments. The overarching objective of this research is to develop novel solutions that enhance semantic awareness in exploration, leverage quantum computational paradigms for superior viewpoint selection, and enable robust integration of low-coherence modalities for perception in challenging lighting conditions. This objective is pursued through three complementary research directions, each addressing a specific aspect of the intelligent perception challenge.

The first major contribution of this thesis is the development of a semantic-aware Next-Best-View (NBV) planning framework for multi-degree-of-freedom mobile systems. This framework fundamentally reimagines view planning by explicitly incorporating semantic information alongside traditional visibility metrics in the utility function for viewpoint evaluation. Unlike previous approaches that treat all unknown regions equally, our formulation distinguishes between different environmental elements based on their semantic significance, enabling more purposeful exploration trajectories. The key innovation lies in the novel information gain formulation that integrates both visibility gain and semantic gain in a unified mathematical framework, allowing the system to dynamically balance between global exploration and targeted investigation of objects of interest. This approach is further enhanced through an adaptive strategy with termination criteria that facilitates efficient two-stage search-and-acquisition manoeuvres, first locating objects of interest and then acquiring comprehensive perceptual data about them. Experimental results demonstrate significant improvements in exploration efficiency, achieving up to 27.46% enhancement in region-of-interest reconstruction and dramatically improved perspective directivity

compared to state-of-the-art methods.

Building upon the challenge of solution optimality in viewpoint selection, the second major contribution is the development of a hybrid quantum-classical framework (HQC-NBV) for robotic view planning. This paradigm-shifting approach leverages recent advances in quantum computing to address the inherent limitations of classical optimisation methods in navigating the complex solution space of view planning. Rather than relying on heuristics or approximate sampling methods that often lead to suboptimal solutions, our approach formulates the NBV problem as a quantum optimisation task through a carefully designed multi-component Hamiltonian that encodes exploration objectives, environmental constraints, and complex parameter interdependencies. The key innovation lies in the parameter-centric variational ansatz with bidirectional alternating entanglement patterns that capture the hierarchical dependencies between viewpoint parameters. This hybrid quantum-classical approach enables more effective exploration of the vast solution space by leveraging quantum superposition and entanglement to simultaneously evaluate multiple movement strategies while encoding complex spatial relationships, leading to significantly improved optimisation outcomes. Comprehensive experimental validations demonstrate that quantum-specific components provide measurable performance advantages, with up to 49.2% higher exploration efficiency compared to classical methods, establishing a pioneering connection between quantum computing and robotic perception.

The third major contribution addresses the challenge of robust perception through multi-modal integration of low-coherence sensing data. The proposed Cross Shallow and Deep Perception Network (CSDNet) represents a lightweight architecture specifically designed to integrate depth and thermal modalities—two sensing approaches that preserve privacy by not capturing RGB information. Unlike conventional multi-modal methods that use RGB as the primary modality, our approach maximises scene interpretation by leveraging the complementary nature of depth and thermal data despite their relatively low coherence. The key innovations include a spatial information

prescreening mechanism and implicit coherence navigation across shallow and deep network layers. This architecture is further enhanced through a Segment Anything Model (SAM)-assisted encoder pre-training framework that guides effective feature mapping to a generalised feature space. Experimental results demonstrate that our approach achieves state-of-the-art performance while reducing computational requirements by orders of magnitude compared to triple-modality methods, making it particularly suitable for deployment on resource-constrained mobile platforms.

These three contributions each represent significant advancements for next-generation intelligent perception systems, addressing key challenges from different perspectives. The semantic-aware planning approach enables more efficient exploration by focusing on objects of interest; the hybrid quantum-classical optimisation achieves superior solution quality in viewpoint selection; and the low-coherence modality integration architecture provides robust perception capabilities while preserving privacy. While developed as distinct solutions to specific challenges, these advances collectively push forward the state of the art in robotic perception for applications ranging from search and rescue operations to autonomous exploration and privacy-preserving surveillance. The research presented in this thesis not only makes notable contributions to each of these domains individually but also demonstrates how innovations across these different aspects of perception can substantially improve the capabilities of autonomous mobile systems in complex environments.

1.4 Thesis Outline

This thesis is organised into six chapters, structured to progressively develop the concepts, methodologies, and experimental validations of the proposed intelligent perception systems for mobile robots.

Chapter 1 introduces the research context, motivation, challenges, and objectives of

the thesis. It establishes the importance of intelligent perception in robotics, identifies key research gaps in current approaches, and outlines the main contributions of the research.

Chapter 2 provides a comprehensive review of the foundational concepts and related work across the three main research themes. It covers mobile robot perception fundamentals, Next-Best-View planning approaches, semantics in robot perception, quantum computing applications in robotics, multi-modal perception systems, and foundation models for perception. This chapter establishes the theoretical and technical background necessary for understanding the subsequent contributions.

Chapter 3 presents the Semantic-aware Next-Best-View for Multi-DoFs Mobile Systems, the first major contribution of this study, focusing on the integration of semantic information into view planning for mobile robots.

Chapter 4 presents the Hybrid Quantum-Classical Approach for View Planning. This chapter introduces the novel hybrid quantum-classical framework for viewpoint optimisation.

Chapter 5 presents the Cross Shallow and Deep Perception Network for Multi-Modal Fusion, the third major contribution of this study, detailing the lightweight architecture for integrating low-coherence modalities.

Chapter 6 is the final chapter that summarises the research contributions, discusses the limitations of the current approaches, and outlines promising directions for future work.

Chapter 2

Background and Literature Review

2.1 Robotic Perception and Environmental Representation Fundamentals

Mobile robots require reliable sensing and effective environmental representations to perceive and interact with their surroundings. This section examines fundamental sensing modalities and representational frameworks that underpin advanced robotic perception systems, establishing the theoretical foundation for the perception-based planning and decision-making approaches proposed in this thesis.

2.1.1 Robotic Sensing Modalities

Robotic perception relies on various sensing modalities to capture different aspects of the environment. Each modality offers unique capabilities and limitations, making the selection and integration of appropriate sensors critical to the overall performance of robotic systems.

Visual Perception

RGB cameras remain the primary sensing modality for most robotic systems due to their rich information content, high spatial resolution, and low cost. These sensors provide colour and texture information crucial for object recognition, feature tracking, and general scene understanding. The pinhole camera model forms the mathematical foundation for projecting 3D world points onto a 2D image plane:

$$\lambda \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} R & t \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{2.1}$$

where (u, v) are the image coordinates, (X, Y, Z) are the world coordinates, f_x and f_y are the focal lengths, (c_x, c_y) is the principal point, R is the rotation matrix, t is the translation vector, and λ is a scaling factor. This model enables the interpretation of 2D image measurements in the context of 3D spatial reasoning. Despite their utility, visual sensors are highly sensitive to environmental conditions such as illumination variations, shadows, and reflections. Additionally, monocular RGB cameras lack direct depth information, necessitating computational approaches to infer 3D structure from 2D observations.

Depth Perception

Depth sensors directly measure the distance to objects in the environment, providing explicit 3D structural information that complements visual data. Common depth sensing technologies include:

Time-of-Flight (ToF) cameras, which measure the time required for light pulses to travel to objects and return to the sensor. The depth Z is calculated as:

$$Z = \frac{c \cdot \Delta t}{2} \tag{2.2}$$

where c is the speed of light and Δt is the time difference between emitted and received light.

Structured light sensors, which project known patterns onto the scene and analyse their deformation. The disparity between the projected and observed patterns enables depth triangulation:

$$Z = \frac{f \cdot B}{d} \tag{2.3}$$

where f is the focal length, B is the baseline between the projector and camera, and d is the disparity.

Stereo vision systems, which estimate depth from the disparity between corresponding points in two camera images:

$$Z = \frac{f \cdot B}{x_L - x_R} \tag{2.4}$$

where x_L and x_R are the x-coordinates of corresponding points in the left and right images. Depth sensors provide valuable information for navigation, obstacle avoidance, and object manipulation tasks. However, they often struggle with transparent, reflective, or highly absorptive surfaces, as well as in outdoor environments with strong ambient light.

Thermal Imaging

Thermal cameras detect infrared radiation emitted by objects based on their temperature, enabling perception independent of visible light conditions. The thermal

radiance L at wavelength λ for an object at temperature T follows Planck's law:

$$L_\lambda(T) = \frac{2hc^2}{\lambda^5} \frac{1}{e^{\frac{hc}{\lambda k_B T}} - 1} \tag{2.5}$$

where h is Planck's constant, c is the speed of light, k_B is Boltzmann's constant, and T is the absolute temperature. Thermal imaging excels in conditions where visual perception is challenging, such as darkness, smoke, or fog. It is particularly effective for detecting living beings, as they typically exhibit temperature signatures distinct from the background environment. This capability makes thermal sensing valuable for search and rescue operations, surveillance, and human-robot interaction. The limitations of thermal imaging include lower spatial resolution compared to RGB cameras, sensitivity to ambient temperature conditions, and difficulty in distinguishing objects with similar thermal properties. To be more general, the qualitative comparison of sensing modalities can be represented as shown in Table 2.1.

Table 2.1: Qualitative Comparison of Sensing Modalities

Performance Metric	RGB	Depth	Thermal
Low Light Performance	⋆	⋆⋆⋆	⋆⋆⋆⋆⋆
Texture Recognition	⋆⋆⋆⋆⋆	⋆⋆	⋆⋆
Geometric Information	⋆⋆	⋆⋆⋆⋆⋆	⋆⋆
Privacy Preservation	⋆	⋆⋆⋆⋆	⋆⋆⋆
Computational Efficiency	⋆⋆	⋆⋆⋆	⋆⋆⋆⋆
Weather Robustness	⋆	⋆⋆⋆	⋆⋆⋆⋆
Power Consumption	⋆⋆⋆	⋆⋆	⋆⋆⋆
Material Differentiation	⋆⋆⋆⋆	⋆⋆	⋆⋆⋆⋆⋆
Cost-Effectiveness	⋆⋆⋆⋆⋆	⋆⋆⋆	⋆⋆

Sensor Uncertainty Modeling

All sensor measurements contain inherent uncertainties that must be properly modeled for robust perception. Figure 2.1 illustrates the complete process of transform-

ing raw sensor readings into a geometric model through probabilistic estimation. This process encompasses sensor reading acquisition, probabilistic sensor modeling, Bayesian estimation, and final decision rule application.

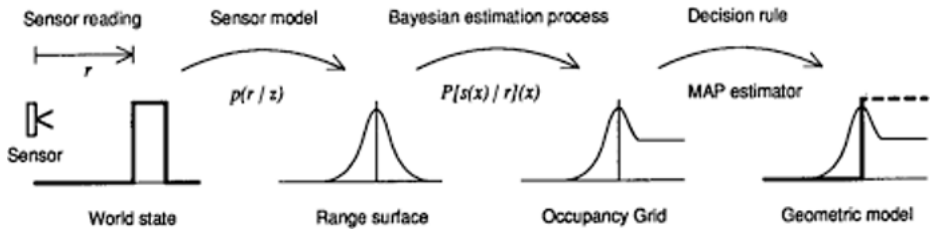

Figure 2.1: The sensor uncertainty modeling pipeline: from raw sensor reading to geometric model construction via probabilistic estimation. The process involves modeling the sensor reading with an appropriate probability distribution, applying Bayesian estimation to derive the occupancy grid, and using a maximum a posteriori (MAP) estimator as the decision rule to construct the final geometric model.

The Bayes filter provides a formal probabilistic framework for state estimation under uncertainty [94]:

$$p(x_t|z_{1:t}, u_{1:t}) = \eta \cdot p(z_t|x_t) \cdot \int p(x_t|x_{t-1}, u_t) \cdot p(x_{t-1}|z_{1:t-1}, u_{1:t-1}) dx_{t-1} \tag{2.6}$$

where $p(x_t|z_{1:t}, u_{1:t})$ is the posterior probability of the state x_t given all measurements $z_{1:t}$ and control inputs $u_{1:t}$, $p(z_t|x_t)$ is the measurement model shown in the second step of Figure 2.1, $p(x_t|x_{t-1}, u_t)$ is the motion model, and η is a normalization constant.

For sensors with Gaussian noise characteristics, the Kalman filter provides an efficient recursive estimation solution for the third step in Figure 2.1:

$$\begin{aligned}
\hat{x}_{t|t-1} &= F_t \hat{x}_{t-1|t-1} + B_t u_t \\
P_{t|t-1} &= F_t P_{t-1|t-1} F_t^T + Q_t \\
K_t &= P_{t|t-1} H_t^T (H_t P_{t|t-1} H_t^T + R_t)^{-1} \\
\hat{x}_{t|t} &= \hat{x}_{t|t-1} + K_t (z_t - H_t \hat{x}_{t|t-1}) \\
P_{t|t} &= (I - K_t H_t) P_{t|t-1}
\end{aligned} \tag{2.7}$$

where $\hat{x}$ is the state estimate, P is the covariance matrix, F is the state transition matrix, H is the observation matrix, K is the Kalman gain, and Q and R are the process and measurement noise covariance matrices, respectively. The final MAP estimator shown in the fourth step of Figure 2.1 selects the most likely state from this posterior distribution.

2.1.2 Environmental Representation Techniques

Environmental representations form the foundation for robotic understanding of the surrounding world, enabling tasks such as mapping, localisation, planning, and interaction. This section explores key representation techniques, focusing on their mathematical foundations and practical applications.

Occupancy Grid Maps

Occupancy grid maps represent the environment as a discrete grid of cells, each associated with a probability of being occupied. Figure 2.2 illustrates this concept, showing how sensor measurements from a single viewpoint update the occupancy probabilities of cells in the map. Black cells indicate high occupancy probability (0.95), white cells represent low occupancy probability (0.05), and gray cells denote unknown or uncertain regions (0.5).

For a cell $m_{i,j}$, the posterior probability of occupancy given sensor measurements $z_{1:t}$ and robot poses $x_{1:t}$ is calculated using Bayes' rule:

$$p(m_{i,j}|z_{1:t}, x_{1:t}) = \frac{p(z_t|m_{i,j}, x_t) \cdot p(m_{i,j}|z_{1:t-1}, x_{1:t-1})}{p(z_t|z_{1:t-1}, x_{1:t-1})} \tag{2.8}$$

As shown in Figure 2.2, the sensor model $p(z_t|m_{i,j}, x_t)$ assigns different occupancy probabilities based on sensor readings. Cells along the sensor rays (shown in red) before an obstacle are typically assigned low occupancy probabilities, while cells where

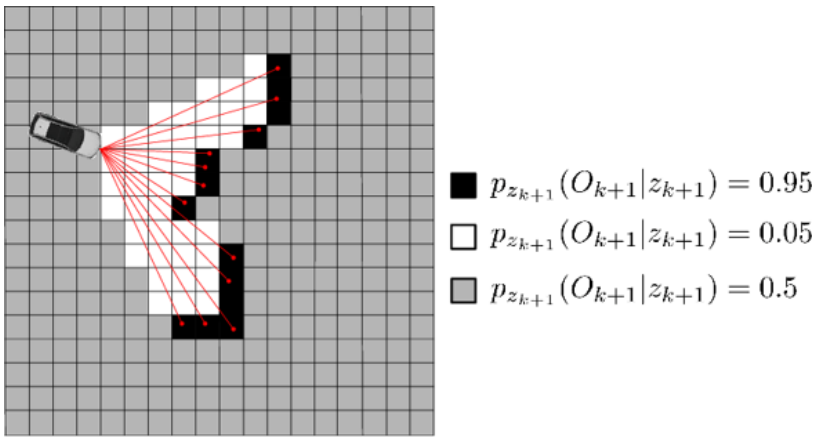

Figure 2.2: Occupancy grid mapping example showing a robot updating cell occupancy probabilities based on sensor measurements.

rays end are assigned high occupancy probabilities, consistent with the detection of an obstacle.

To simplify the computation, the log-odds representation is often used:

$$l(m_{i,j}|z_{1:t}, x_{1:t}) = l(m_{i,j}|z_{1:t-1}, x_{1:t-1}) + l(m_{i,j}|z_t, x_t) - l_0 \tag{2.9}$$

where $l(m_{i,j}) = \log \frac{p(m_{i,j})}{1-p(m_{i,j})}$ and l_0 is the prior log-odds ratio [94].

Occupancy grid maps are widely used in robotic navigation and planning due to their simplicity and effectiveness in representing both known obstacles and unexplored regions. However, they suffer from discretisation artifacts, memory inefficiency for large environments, and limited ability to represent uncertainty in object boundaries [43].

Truncated Signed Distance Fields

Truncated Signed Distance Fields (TSDFs) encode the environment by storing the distance to the nearest surface at each point in space, truncated to a maximum value

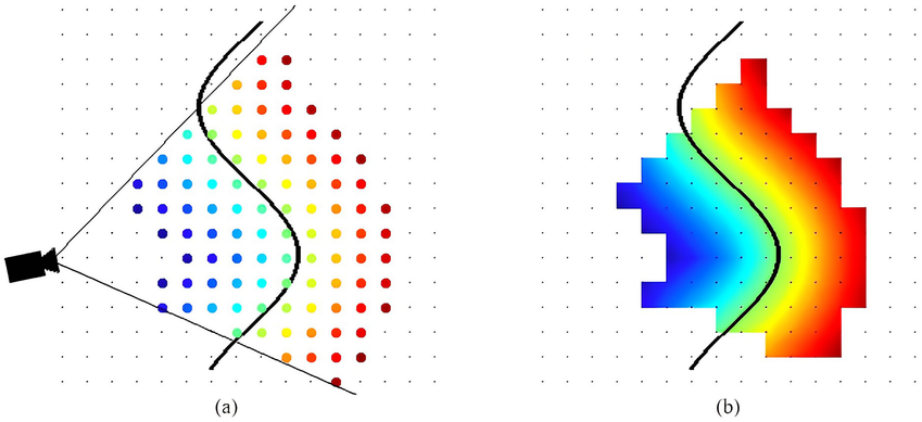

Figure 2.3: The black line represents the surface. Positive distances are indicated by the colour blue to green; negative distances are indicated by the colour green to red [70].

for computational efficiency [24]. The TSDF function $\Phi_\tau(\mathbf{x})$ is defined as:

$$\Phi_\tau(\mathbf{x}) = \begin{cases} \min(\tau, d(\mathbf{x}, \partial\Omega)) & \text{if } \mathbf{x} \text{ is outside } \Omega \\ \max(-\tau, -d(\mathbf{x}, \partial\Omega)) & \text{if } \mathbf{x} \text{ is inside } \Omega \end{cases} \tag{2.10}$$

where $d(\mathbf{x}, \partial\Omega)$ is the Euclidean distance from point $\mathbf{x}$ to the nearest point on the boundary $\partial\Omega$ of object Ω, and τ is the truncation distance. TSDFs can be incrementally updated with new measurements through weighted average fusion:

$$\Phi_t(\mathbf{x}) = \frac{W_{t-1}(\mathbf{x}) \cdot \Phi_{t-1}(\mathbf{x}) + w_t(\mathbf{x}) \cdot \Phi_{\text{new}}(\mathbf{x})}{W_{t-1}(\mathbf{x}) + w_t(\mathbf{x})} \tag{2.11}$$

$$W_t(\mathbf{x}) = W_{t-1}(\mathbf{x}) + w_t(\mathbf{x}) \tag{2.12}$$

where $W_t(\mathbf{x})$ is the cumulative weight at position $\mathbf{x}$ and time t, and $w_t(\mathbf{x})$ is the weight of the new measurement, typically inversely proportional to the measurement uncertainty. TSDF-based representations enable efficient surface reconstruction through methods like Marching Cubes [86] and support rapid ray-casting for synthetic view generation and collision detection. The KinectFusion system [72] demonstrated the effectiveness of TSDF for real-time 3D reconstruction, influencing numerous subse-

quent mapping systems. A visualised representation is presented in Figure 2.3.

Point Cloud Representations

Point clouds represent the environment as a collection of 3D points directly sampled from surfaces. A point cloud $\mathcal{P}$ can be defined as:

$$\mathcal{P} = \mathbf{p}_i = (x_i, y_i, z_i) \in \mathbb{R}^3 \mid i = 1, 2, \ldots, N \tag{2.13}$$

For colored point clouds with normals:

$$\mathcal{P}\text{colored} = \mathbf{p}i = (x_i, y_i, z_i, r_i, g_i, b_i, nx, i, ny, i, n_{z,i}) \mid i = 1, 2, \ldots, N \tag{2.14}$$

Point clouds provide a direct representation of sensor measurements without discretisation, making them suitable for preserving fine details. However, they lack explicit connectivity information and can be memory-intensive for dense representations. Processing point clouds often involves registration to align multiple scans, which can be formulated as finding the rigid transformation that minimises the distance between the source point cloud and the target point cloud:

$$R^*, \mathbf{t}^* = \arg\min_{R,\mathbf{t}} \sum_{i=1}^{N} |R\mathbf{p}_i + \mathbf{t} - \mathbf{q}_i|^2 \tag{2.15}$$

where $\mathbf{p}_i$ and $\mathbf{q}_i$ are corresponding points, R is a rotation matrix, and $\mathbf{t}$ is a translation vector. The Iterative Closest Point (ICP) algorithm and its variants are commonly used for point cloud registration, iteratively refining the transformation estimate by alternating between correspondence identification and transformation optimisation [83].

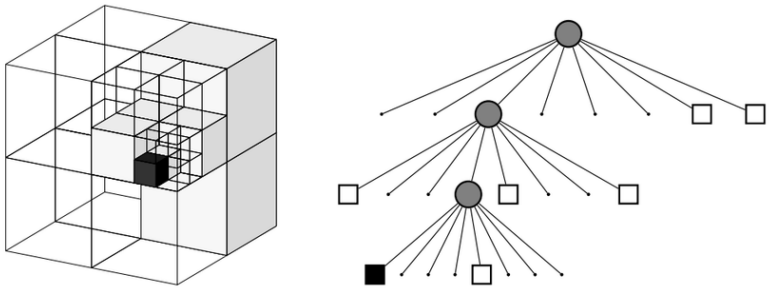

Figure 2.4: An illustration of the Octomap octree representation [39].

Octree-Based Representations

Octrees provide a hierarchical spatial partitioning of the environment, offering a compromise between the resolution of occupancy grids and the memory efficiency of sparse representations [43]. Each node in the octree represents a cubic volume of space, which is recursively subdivided into eight octants if needed, as is illustrated in Figure 2.4:

$$\text{volume}(n) = \begin{cases} \text{volume}(n_{\text{parent}})/8 & \text{if } n \text{ is not root} \\ \text{volume}(entirespace) & \text{if } n \text{ is root} \end{cases} \tag{2.16}$$

Octrees enable multi-resolution representation of the environment, with higher resolution in areas of interest and lower resolution in empty or homogeneous regions. This property makes them particularly suitable for large-scale outdoor environments and scenarios with varying detail requirements. The OctoMap framework [43] combines octrees with probabilistic occupancy estimation, enabling efficient 3D mapping with the ability to represent unknown space explicitly:

$$p(n|z_1) = \frac{p(z_t|n) \cdot p(n|z_1)}{p(z_t|z_1)} \tag{2.17}$$

where $p(n|z_1)$ is the posterior occupancy probability of node n given all measurements z_1.

2.1.3 Multi-Modal Perception

Multi-modal perception systems integrate information from multiple sensing modalities to achieve more robust and comprehensive environmental understanding. This section presents theoretical frameworks and practical approaches for multi-modal fusion.

Theoretical Foundations of Multi-Modal Fusion

From an information-theoretic perspective, the benefit of multi-modal fusion arises from complementary information across modalities. The mutual information between multiple modalities $X^1, X^2, \ldots, X^M$ and the environment state Y can be decomposed as:

$$I(X^1, X^2, \ldots, X^M; Y) = \sum_{i=1}^{M} I(X^i; Y) - \sum_{i=1}^{M} I(X^i; X^1, \ldots, X^{i-1}, X^{i+1}, \ldots, X^M; Y) \tag{2.18}$$

where $I(X^i; Y)$ represents the information content of modality i with respect to Y, and the second term represents redundancy across modalities. The effectiveness of multi-modal fusion depends on both the complementarity of information across modalities and the quality of integration methods. Fusion approaches can be categorised based on the level at which information from different modalities is combined [3].

Early, Late, and Intermediate Fusion

Early fusion combines raw data or low-level features from different modalities before processing:

$$F = \phi([X^1, X^2, \ldots, X^M]) \tag{2.19}$$

where $[X^1, X^2, \ldots, X^M]$ represents the concatenation of features from different modalities, and ϕ is a function that maps the concatenated features to a joint representation

F. Late fusion combines predictions or decisions from modality-specific models:

$$\hat{Y} = g(f_1(X^1), f_2(X^2), \ldots, f_M(X^M)) \tag{2.20}$$

where f_i is the model for modality i, and g is the fusion function [3]. Intermediate fusion combines features at multiple levels of abstraction:

$$F^l = \phi^l([F_1^l, F_2^l, \ldots, F_M^l]) \tag{2.21}$$

where F_i^l represents features from modality i at level l [79]. Each fusion approach offers distinct advantages and limitations. Early fusion can capture low-level correlations between modalities but may struggle with modalities of different dimensionalities or sampling rates. Late fusion is more modular and can leverage pre-trained unimodal models but may miss cross-modal interactions. Intermediate fusion aims to balance these trade-offs by integrating information at multiple levels but often requires more complex architectures [3].

Modality Coherence and Integration Challenges

The coherence between modalities—defined as the degree of correlation in the information they provide—significantly influences fusion strategies [3]. High-coherence modality pairs exhibit strong correlations in their information content, facilitating simpler fusion approaches. In contrast, low-coherence modalities may contain more complementary information but require more sophisticated integration methods [46]. Modality coherence can be quantified using mutual information:

$$C(X^i, X^j) = \frac{I(X^i; X^j)}{\sqrt{H(X^i) \cdot H(X^j)}} \tag{2.22}$$

where $I(X^i; X^j)$ is the mutual information between modalities i and j, and $H(X^i)$ is the entropy of modality i.

Challenges in multi-modal integration include:

- Alignment and registration: Different modalities may have different spatial resolutions, fields of view, or sampling rates, necessitating careful alignment.
- Missing or corrupted data: Sensors may fail or provide unreliable data under certain conditions, requiring robust fusion methods that can handle incomplete information.
- Conflicting information: Different modalities may provide contradictory information about the same scene element, necessitating conflict resolution strategies.
- Computational efficiency: Processing multiple data streams increases computational requirements, which can be challenging for resource-constrained robotic platforms.

Advanced fusion methods address these challenges through techniques such as attention mechanisms, uncertainty modelling, and adaptive weighting of modalities based on their reliability in different contexts [3]. These approaches enable robust multimodal perception across a wide range of environmental conditions, supporting advanced robot perception and decision-making capabilities.

2.2 Next-Best-View Planning

Next-Best-View (NBV) planning addresses the fundamental challenge of determining optimal sensor configurations to maximise information gain about an environment or object. This section examines the theoretical foundations, algorithmic approaches, and evaluation metrics for NBV planning, providing the background for the semantic-aware and hybrid quantum-classical approaches proposed in this thesis.

2.2.1 Theoretical Foundations

Next-Best-View planning originated in the context of unknown environment exploration and reconstruction, with pioneering work by Connolly et al. [22] and Maver et al. [67] establishing the conceptual framework.

Problem Formulation

The NBV problem, initially approached through heuristic methods in early work by Connolly [22], has evolved to be formalised as an optimisation problem over possible viewpoints. In modern formulations, selecting from a set of candidate viewpoints $\mathcal{V} = \{v_1, v_2, \ldots, v_n\}$, the objective is to select the next viewpoint v^* that maximizes an information gain function $I(v)$ [85]:

$$v^* = \arg\max_{v \in \mathcal{V}} I(v) \tag{2.23}$$

For model-based NBV, where a coarse model of the object is available, the information gain can be computed directly from the model [20]. In contrast, non-model-based approaches operate without prior knowledge, requiring sequential decision-making as information is accumulated [25].

At each step t of non-model-based exploration, the next best viewpoint v_t^* is selected based on the current knowledge state $\mathcal{K}_{t-1}$:

$$v_t^* = \arg\max_{v \in \mathcal{V}} I(v|\mathcal{K}_{t-1}) \tag{2.24}$$

The knowledge state $\mathcal{K}_t$ is updated after each observation:

$$\mathcal{K}_t = \mathcal{K}_{t-1} \cup O(v_t^*) \tag{2.25}$$

where $O(v_t^*)$ represents the observation obtained from viewpoint v_t^*.

Mathematical Properties

The NBV problem exhibits several mathematical properties that influence algorithm design and performance:

Non-convexity: The information gain function $I(v)$ is generally non-convex, making global optimisation challenging and often necessitating sampling-based or heuristic approaches.

Submodularity: Under certain conditions, the information gain function exhibits submodularity, meaning that the marginal benefit of adding a new viewpoint decreases as more viewpoints are selected:

$$I(A \cup v) - I(A) \geq I(B \cup v) - I(B), \quad \forall A \subseteq B, v \notin B \tag{2.26}$$

This property enables near-optimal greedy solutions with theoretical performance guarantees.

Diminishing returns: The incremental information gain typically decreases as more observations are collected, following a logarithmic or exponential decay pattern [12]:

$$\frac{dI(t)}{dt} \propto -\alpha I(t) \tag{2.27}$$

where α is a decay constant and t represents time or the number of observations.

2.2.2 Algorithmic Approaches

NBV planning algorithms can be categorised based on their search strategy, information gain metrics, and application domain. This section examines key algorithmic approaches and their relative strengths and limitations.

Sampling-Based Methods

Sampling-based NBV methods generate candidate viewpoints through random sampling or structured sampling strategies [96, 12]. These approaches are particularly effective in large or complex environments where exhaustive evaluation of all possible viewpoints is computationally infeasible. The Rapidly-exploring Random Tree (RRT) algorithm [60] and its variant RRT* [53] are commonly employed for exploring the configuration space. The RRT algorithm incrementally builds a tree $\mathcal{T} = (V, E)$ in the configuration space $\mathcal{C}$, starting from an initial configuration q_{init}. At each iteration, a random configuration q_{rand} is sampled, and the nearest node q_{near} in the tree is identified:

$$q_{\text{near}} = \arg\min_{q \in V} |q - q_{\text{rand}}| \tag{2.28}$$

A new node q_{new} is generated by moving from q_{near} towards q_{rand} by a step size Δq:

$$q_{\text{new}} = q_{\text{near}} + \min\left(1, \frac{\Delta q}{|q_{\text{rand}} - q_{\text{near}}|}\right) \cdot (q_{\text{rand}} - q_{\text{near}}) \tag{2.29}$$

The node q_{new} is added to the tree if it passes collision checks against known obstacles [60]. RRT* improves upon RRT by incorporating two additional steps [53]:

Parent selection: Choose the parent node that minimises the cost to reach q_{new}:

$$q_{\text{parent}} = \arg\min_{q \in Q_{\text{near}}} c(q) + d(q, q_{\text{new}}) \tag{2.30}$$

where $c(q)$ is the cost to reach node q from the root, $d(q, q_{\text{new}})$ is the distance between q and q_{new}, and Q_{near} is the set of nodes within a specified radius of q_{new}.

Rewiring: Update the parent of existing nodes if reaching them through q_{new} results in a lower cost:

$$\forall q \in Q_{\text{near}} : \text{if } c(q_{\text{new}}) + d(q_{\text{new}}, q) < c(q) \text{ then } \text{parent}(q) \leftarrow q_{\text{new}} \tag{2.31}$$

These modifications ensure asymptotic optimality, guaranteeing convergence to the optimal solution as the number of samples approaches infinity.

In the context of NBV planning, RRT-based methods grow trees in the known free space and evaluate candidate viewpoints along the branches based on expected information gain and movement cost. For unknown environment exploration, the Receding Horizon Next-Best-View (RH-NBV) approach [12] has demonstrated particular effectiveness by combining RRT-based exploration with receding horizon control:

$$b^* = \arg\max_{b \in \mathcal{B}} \sum_{v \in b} I(v) \cdot \gamma^{d(v)} \tag{2.32}$$

where $\mathcal{B}$ is the set of all branches in the RRT, $I(v)$ is the information gain at viewpoint v, γ is a discount factor, and $d(v)$ is the depth of node v in the tree.

Deterministic Approaches

Deterministic NBV methods rely on analytical criteria to determine optimal viewpoints [35, 85]. These approaches often employ geometrical reasoning, voxel visibility analysis, or information-theoretic metrics to evaluate viewpoint utility.

Frontier-based exploration [104] is a popular deterministic approach that directs the robot towards boundaries between known and unknown regions. The frontier $\mathcal{F}$ is defined as the set of known free cells that are adjacent to unknown cells:

$$\mathcal{F} = c \in \mathcal{M}_{\text{free}} \mid \exists c' \in \mathcal{N}(c) : c' \in \mathcal{M}_{\text{unknown}} \tag{2.33}$$

where $\mathcal{M}_{\text{free}}$ and $\mathcal{M}_{\text{unknown}}$ are the sets of free and unknown cells in the map, respectively, and $\mathcal{N}(c)$ is the set of neighboring cells of c. The next best viewpoint is typically selected to minimise the distance to the nearest frontier:

$$v^* = \arg\min_{v \in \mathcal{V}} \min_{f \in \mathcal{F}} |v - f| \tag{2.34}$$

Frontier-based methods have shown particular effectiveness in high-speed flight and fast exploration tasks [21], but may struggle to generalise to other applications that require more sophisticated information gain formulations.

For object reconstruction tasks, volumetric approaches evaluate viewpoints based on the visibility of unknown voxels [96]. The information gain for a viewpoint v can be formulated as:

$$I(v) = \sum_{x \in \mathcal{M}_{\text{unknown}} \cap \text{FOV}(v)} \text{Visible}(x, v) \tag{2.35}$$

where $\text{FOV}(v)$ is the field of view from viewpoint v, and $\text{Visible}(x, v)$ is a binary function indicating whether voxel x is visible from viewpoint v.

2.2.3 Information Gain Metrics

Information gain metrics quantify the expected utility of a viewpoint in terms of new information about the environment or object of interest. Different metrics emphasise various aspects of exploration, such as coverage, precision, or task relevance.

Visibility-Based Gain

Visibility-based metrics measure the number or proportion of unknown voxels that would become visible from a candidate viewpoint [4]:

$$I_{\text{vis}}(v) = \sum_{x \in \mathcal{M}} \mathbf{1}(x \in \mathcal{M}_{\text{unknown}} \text{ and } \text{Visible}(x, v)) \tag{2.36}$$

where $\mathbf{1}(\cdot)$ is the indicator function, and $\text{Visible}(x, v)$ checks if voxel x is visible from viewpoint v. Raycasting is commonly used to determine visibility:

$$\text{Visible}(x, v) = \begin{cases} 1 & \text{if ray } r_{v,x} \text{ does not intersect before reaching } x \\ 0 & \text{otherwise} \end{cases} \tag{2.37}$$

where $r_{v,x}$ is the ray from viewpoint v to voxel x [12]. While visibility-based gain is computationally efficient, it treats all unknown voxels equally, regardless of their significance for the task at hand or their spatial relationship to the viewpoint.

Information-Theoretic Gain

Information theory provides a rigorous framework for quantifying information gain. Entropy reduction measures the expected decrease in uncertainty about the environment:

$$I_{\text{entropy}}(v) = H(\mathcal{M}) - \mathbb{E}_{z|v}[H(\mathcal{M}|z)] \tag{2.38}$$

where $H(\mathcal{M})$ is the entropy of the current map, and $H(\mathcal{M}|z)$ is the conditional entropy after obtaining observation z from viewpoint v. For occupancy grid maps, the entropy can be computed as:

$$H(\mathcal{M}) = -\sum_{x \in \mathcal{M}} [p(x) \log p(x) + (1 - p(x)) \log(1 - p(x))] \tag{2.39}$$

where $p(x)$ is the probability of voxel x being occupied. Mutual information between the map $\mathcal{M}$ and a potential observation z from viewpoint v provides another information-theoretic metric:

$$I_{\text{MI}}(v) = I(\mathcal{M}; z|v) = H(z|v) - H(z|\mathcal{M}, v) \tag{2.40}$$

where $H(z|v)$ is the entropy of the observation given the viewpoint, and $H(z|\mathcal{M}, v)$ is the entropy of the observation given both the map and the viewpoint. Information-theoretic metrics provide a principled approach to uncertainty reduction but can be computationally intensive for large environments or complex sensor models.

Proximity-Based Volumetric Information

Delmerico et al. [25] introduced several proximity-based volumetric information gain metrics that consider the spatial relationship between viewpoints and voxels:

Proximity Count (PC):

$$I_{\text{PC}}(v) = \sum_{x \in \mathcal{M}_{\text{unknown}} \cap \text{FOV}(v)} \frac{1}{|x - v|^2} \tag{2.41}$$

Occlusion Aware (OA):

$$I_{\text{OA}}(v) = \sum_{x \in \mathcal{M}_{\text{unknown}} \cap \text{FOV}(v)} \frac{P(\text{Visible}(x, v))}{|x - v|^2} \tag{2.42}$$

Area Factor (AF):

$$I_{\text{AF}}(v) = \sum_{x \in \mathcal{M}_{\text{unknown}} \cap \text{FOV}(v)} \frac{\cos \theta_{x,v}}{|x - v|^2} \tag{2.43}$$

where $\theta_{x,v}$ is the angle between the viewing ray and the surface normal at voxel x. These metrics account for the fact that closer observations generally provide more accurate information, addressing a limitation of simpler visibility-based approaches.

Utility Functions with Cost Considerations

In practice, NBV planning must balance information gain with the cost of reaching a viewpoint. Utility functions typically combine gain and cost terms:

$$U(v) = I(v) - \lambda C(v) \tag{2.44}$$

where $I(v)$ is the information gain, $C(v)$ is the cost (e.g., distance, energy, time), and λ is a weighting factor that determines the relative importance of gain versus cost [12]. Schmid et al. [84] proposed a ratio-based utility function that eliminates the

need for parameter tuning:

$$U(v) = \frac{I(v)}{C(v)} \tag{2.45}$$

This formulation naturally balances information gain against cost, avoiding the sensitivity to parameter selection present in weighted-sum approaches.

2.2.4 Receding Horizon and Multi-Step Planning

While greedy NBV approaches select the single best next viewpoint, receding horizon and multi-step planning methods consider sequences of viewpoints to avoid local optima and improve global exploration efficiency.

Receding Horizon NBV

Receding Horizon Control (RHC), also known as Model Predictive Control (MPC), optimises over a finite planning horizon but only executes the first step before replanning:

$$P^* = \arg\max_{P \in \mathcal{P}} I_{\text{path}}(P) \tag{2.46}$$

where $\mathcal{P}$ is the set of feasible paths, and $I_{\text{path}}(P)$ is the cumulative information gain along path P. The robot executes only the first action to reach the first viewpoint $v_1 \in P^*$, then replans based on the updated knowledge state. This approach allows the robot to adapt to new observations while still considering multi-step consequences of its actions.

The RH-NBV algorithm [13] combines RRT-based exploration with receding horizon control. It grows an RRT from the current position, evaluates the information gain along each branch, selects the best branch, and executes the first edge before

replanning. The information gain for a path is computed as:

$$I_{\text{path}}(P) = \sum_{v \in P} I(v) \cdot \gamma^{d(v)} \tag{2.47}$$

where γ is a discount factor, and $d(v)$ is the depth of node v in the tree. The discount factor prioritises near-term gains over long-term ones, addressing the increasing uncertainty in future observations. The RH-NBV approach has demonstrated superior performance compared to greedy methods, particularly in complex environments with local minima [12, 68]. By considering multiple steps ahead, it can identify paths that may have low immediate gain but lead to high-gain regions later.

POMDP Formulations

The NBV problem can be formulated as a Partially Observable Markov Decision Process (POMDP), providing a principled framework for sequential decision-making under uncertainty:

$$\text{POMDP} = \langle \mathcal{S}, \mathcal{A}, \mathcal{Z}, T, O, R, \gamma \rangle \tag{2.48}$$

where: $\mathcal{S}$ is the state space (e.g., robot pose and environment map), $\mathcal{A}$ is the action space (e.g., robot movements), $\mathcal{Z}$ is the observation space (e.g., sensor measurements), $T(s, a, s') = P(s'|s, a)$ is the transition function, $O(s, a, z) = P(z|s, a)$ is the observation function, $R(s, a)$ is the reward function, γ is the discount factor.

The objective is to find a policy π that maximises the expected discounted reward:

$$V^{\pi}(b) = \mathbb{E}\pi \left[\sum t = 0^{\infty} \gamma^t R(s_t, a_t) \mid b_0 = b \right] \tag{2.49}$$

where b is the belief state (probability distribution over states). Exact POMDP solutions are computationally intractable for large state spaces typical in robotics. However, approximate methods such as Monte Carlo Tree Search (MCTS) [87] and Point-Based Value Iteration (PBVI) [77] have shown promise for NBV planning in

specific scenarios.

Information-Theoretic Planning

Information-theoretic planning approaches optimise sensor trajectories to maximise information gain over extended horizons. The objective is to find a trajectory $\tau = (v_1, v_2, \ldots, v_T)$ that maximizes the cumulative information gain:

$$\tau^* = \arg\max_{\tau} \sum_{t=1}^{T} I(v_t | v_1, \ldots, v_{t-1}) \tag{2.50}$$

subject to dynamic constraints:

$$v_{t+1} = f(v_t, u_t) \tag{2.51}$$

where f is the system dynamics, and u_t is the control input. Hollinger et al. [42] proposed the use of Gaussian processes to model the information content in the environment, enabling more effective planning by capturing the spatial correlation between measurements:

$$I(v_t; v_{t+1} | v_1, \ldots, v_{t-1}) \approx \frac{1}{2} \log \frac{|\Sigma_{t+1|t}|}{|\Sigma_{t+1|t+1}|} \tag{2.52}$$

where $\Sigma_{t+1|t}$ is the predicted covariance matrix, and $\Sigma_{t+1|t+1}$ is the updated covariance matrix after incorporating the measurement at v_{t+1}.

2.2.5 Current Limitations and Research Gaps

Despite significant progress in NBV planning, several limitations and research gaps remain, particularly in the context of complex environments and specialised tasks.

Semantic Awareness

Traditional NBV approaches focus on geometric information gain without considering the semantic content of the environment. All unknown voxels are treated equally, regardless of their significance for the task at hand. This limitation becomes particularly apparent in scenarios where certain objects or regions are more important than others, such as search and rescue operations or object-specific exploration. Semantic information, such as object categories, functional areas, or points of interest, could be integrated into the information gain metric to guide exploration towards task-relevant regions:

$$I_{\text{semantic}}(v) = I_{\text{geometric}}(v) + \lambda_S \cdot S(v) \tag{2.53}$$

where $S(v)$ is a semantic relevance function, and λ_S is a weighting factor. Few works have explored semantic-guided NBV planning, leaving room for novel approaches that leverage rich semantic understanding for more efficient exploration.

Adaptation to Dynamic Environments

Most NBV approaches assume static environments, where the only changes result from the robot's observations. However, many real-world scenarios involve dynamic elements, such as moving objects, changing lighting conditions, or evolving objectives. Adapting NBV planning to dynamic environments requires modeling and predicting changes in the environment:

$$p(s_{t+1}|s_t, a_t) = \int p(s_{t+1}|s_t, a_t, w_t) \cdot p(w_t|s_t) dw_t \tag{2.54}$$

where w_t represents external disturbances or environmental dynamics. Few works have addressed truly dynamic NBV planning, leaving opportunities for novel approaches that explicitly consider environmental changes in the planning process.

Task-Specific Exploration

Different tasks may require different exploration strategies and information gain metrics. For instance, object search requires focusing on potential object locations, while mapping aims for comprehensive coverage. Task-specific NBV planning could incorporate task objectives directly into the utility function:

$$U_{\text{task}}(v) = \mathbb{E}z|v[\Delta U\text{task}(z)] \tag{2.55}$$

where $\Delta U_{\text{task}}(z)$ is the expected improvement in task utility after obtaining observation z. The semantic-aware NBV approach proposed in this thesis addresses several of these limitations by incorporating semantic information to guide viewpoint selection, enabling more efficient and task-relevant exploration in complex environments. Additionally, the hybrid quantum-classical NBV approach tackles the computational complexity challenge by leveraging quantum computing advantages for the combinatorial aspects of viewpoint planning.

2.3 Quantum Computing in Computer Vision

Quantum computing represents a paradigm shift in computational approaches, offering potential advantages for solving complex problems in computer vision and robotic perception. This section explores the principles of quantum computing relevant to visual perception, presents algorithmic approaches, and investigates emerging applications in perception and planning, establishing the theoretical foundation for the hybrid quantum-classical NBV algorithm proposed in this thesis.

2.3.1 Quantum Computing Fundamentals

Quantum computing leverages quantum mechanical phenomena to perform computational tasks with potentially exponential speedups compared to classical approaches for certain problems. Understanding these fundamental principles is essential for developing effective quantum algorithms for computer vision tasks.

Quantum Bits and Quantum States

The fundamental unit of quantum information is the quantum bit or qubit, which exists in a superposition of states until measured. Unlike classical bits that can only be 0 or 1, a qubit's state is represented as:

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle \tag{2.56}$$

where $\alpha, \beta \in \mathbb{C}$ are complex probability amplitudes satisfying $|\alpha|^2 + |\beta|^2 = 1$, and $|0\rangle$ and $|1\rangle$ are the computational basis states. The state of an n-qubit system resides in a 2^n-dimensional Hilbert space and can be written as:

$$|\psi\rangle = \sum_{i=0}^{2^n-1} \alpha_i |i\rangle \tag{2.57}$$

where $\sum_{i=0}^{2^n-1} |\alpha_i|^2 = 1$, and $|i\rangle$ represents the computational basis state corresponding to the binary representation of integer i. This exponential scaling of the state space with the number of qubits is a key source of quantum computational advantage, allowing quantum systems to represent and process vast amounts of information with relatively few qubits.

Key Quantum Phenomena

Several quantum phenomena provide potential computational advantages for computer vision applications:

Superposition allows qubits to exist in multiple states simultaneously, enabling parallel computation. When n qubits are placed in superposition, the system can represent 2^n classical states simultaneously. For image processing, this could theoretically allow simultaneous operations on all pixels.

Entanglement creates non-classical correlations between qubits. For a two-qubit system, an entangled state cannot be factored as a product of individual qubit states:

$$|\psi_{\text{entangled}}\rangle \neq |\phi_1\rangle \otimes |\phi_2\rangle \tag{2.58}$$

Entanglement enables quantum systems to exhibit correlations stronger than any classical system, providing resources for quantum teleportation, superdense coding, and certain speedups in quantum algorithms.

Quantum Interference allows probability amplitudes to constructively or destructively interfere, directing the quantum system toward desired states and away from undesired ones. This phenomenon is central to quantum algorithms like Grover's search and quantum walks.

Quantum Circuits and Gates

Quantum computations are implemented using quantum gates, represented mathematically as unitary matrices that act on quantum states. These gates form the fundamental building blocks of quantum circuits, analogous to logic gates in classical computing but with significantly different properties and capabilities.

Pauli Operators The Pauli operators (X, Y, Z) are fundamental single-qubit gates that correspond to rotations around the respective axes of the Bloch sphere, a geometrical representation of a qubit's state space:

$$X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad Y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad Z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{2.59}$$

The Pauli-X gate (quantum NOT) flips the computational basis states:

$$X|0\rangle = |1\rangle, \quad X|1\rangle = |0\rangle \tag{2.60}$$

The Pauli-Y gate performs a bit flip with an additional phase shift:

$$Y|0\rangle = i|1\rangle, \quad Y|1\rangle = -i|0\rangle \tag{2.61}$$

The Pauli-Z gate (phase flip) leaves $|0\rangle$ unchanged but applies a phase of -1 to $|1\rangle$:

$$Z|0\rangle = |0\rangle, \quad Z|1\rangle = -|1\rangle \tag{2.62}$$

These Pauli operators satisfy important algebraic properties:

$$X^2 = Y^2 = Z^2 = I \tag{2.63}$$

$$XY = iZ, \quad YZ = iX, \quad ZX = iY \tag{2.64}$$

$$[X, Y] = 2iZ, \quad [Y, Z] = 2iX, \quad [Z, X] = 2iY \tag{2.65}$$

where I is the identity operator and $[A, B] = AB - BA$ is the commutator.

Rotation Gates Rotation gates implement rotations around the Bloch sphere axes by arbitrary angles. For a rotation angle θ, these gates are defined as:

$$R_x(\theta) = e^{-i\theta X/2} = \cos(\theta/2)I - i\sin(\theta/2)X = \begin{pmatrix} \cos(\theta/2) & -i\sin(\theta/2) \\ -i\sin(\theta/2) & \cos(\theta/2) \end{pmatrix} \tag{2.66}$$

$$R_y(\theta) = e^{-i\theta Y/2} = \cos(\theta/2)I - i\sin(\theta/2)Y = \begin{pmatrix} \cos(\theta/2) & -\sin(\theta/2) \\ \sin(\theta/2) & \cos(\theta/2) \end{pmatrix} \tag{2.67}$$

$$R_z(\theta) = e^{-i\theta Z/2} = \cos(\theta/2)I - i\sin(\theta/2)Z = \begin{pmatrix} e^{-i\theta/2} & 0 \\ 0 & e^{i\theta/2} \end{pmatrix} \tag{2.68}$$

These rotation gates are particularly important in variational quantum algorithms, where their rotation angles serve as tunable parameters.

Hadamard Gate The Hadamard gate (H) creates superposition states and is defined as:

$$H = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \tag{2.69}$$

It transforms the computational basis states into equal superpositions:

$$H|0\rangle = \frac{|0\rangle + |1\rangle}{\sqrt{2}} = |+\rangle \tag{2.70}$$

$$H|1\rangle = \frac{|0\rangle - |1\rangle}{\sqrt{2}} = |-\rangle \tag{2.71}$$

The Hadamard gate is self-inverse: $H^2 = I$, and when applied to all qubits in a register, it creates a uniform superposition of all computational basis states:

$$H^{\otimes n}|0\rangle^{\otimes n} = \frac{1}{\sqrt{2^n}} \sum_{x=0}^{2^n-1} |x\rangle \tag{2.72}$$

Phase Gates The phase (S) and $\pi/8$ (T) gates implement phase rotations:

$$S = \begin{pmatrix} 1 & 0 \\ 0 & i \end{pmatrix} = \sqrt{Z}, \quad T = \begin{pmatrix} 1 & 0 \\ 0 & e^{i\pi/4} \end{pmatrix} = \sqrt{S} \tag{2.73}$$

These gates, together with the Hadamard gate and the CNOT gate, form a universal set for quantum computation, meaning any unitary operation can be approximated to arbitrary precision using only these gates.

Multi-Qubit Gates Multi-qubit gates enable interactions between qubits, which is essential for creating entanglement and implementing complex quantum algorithms. The CNOT (Controlled-NOT) gate is a two-qubit gate that performs an X operation on the target qubit if the control qubit is in state $|1\rangle$:

$$\text{CNOT} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \tag{2.74}$$

The action of CNOT on computational basis states is:

$$\text{CNOT}|00\rangle = |00\rangle, \quad \text{CNOT}|01\rangle = |01\rangle \tag{2.75}$$

$$\text{CNOT}|10\rangle = |11\rangle, \quad \text{CNOT}|11\rangle = |10\rangle \tag{2.76}$$

The CNOT gate is central to quantum computing as it can create entanglement. For example, applying CNOT to $|+\rangle|0\rangle$ creates a maximally entangled Bell state:

$$\text{CNOT}(H \otimes I)|00\rangle = \text{CNOT}\frac{|00\rangle + |10\rangle}{\sqrt{2}} = \frac{|00\rangle + |11\rangle}{\sqrt{2}} \tag{2.77}$$

The SWAP gate exchanges the states of two qubits:

$$\text{SWAP} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{2.78}$$

The Toffoli gate (CCNOT) is a three-qubit gate that performs an X operation on the target qubit if both control qubits are in state $|1\rangle$. It is a universal gate for classical reversible computation and plays a key role in quantum error correction and fault-tolerant quantum computing.

Parameterized Quantum Circuits Parameterised quantum circuits, central to variational quantum algorithms, combine fixed gate structures with variable rotation angles:

$$U(\boldsymbol{\theta}) = \prod_{l=1}^{L} U_l(\boldsymbol{\theta}_l) = U_L(\boldsymbol{\theta}_L) \cdots U_2(\boldsymbol{\theta}_2) U_1(\boldsymbol{\theta}_1) \tag{2.79}$$

where each layer $U_l(\boldsymbol{\theta}_l)$ typically includes parameterized rotation gates and fixed entangling operations:

$$U_l(\boldsymbol{\theta} l) = E_l \prod i = 1^n R_i(\theta_{l,i}) \tag{2.80}$$

Here, $R_i(\theta_{l,i})$ represents a rotation gate applied to qubit i with angle $\theta_{l,i}$, and E_l is an entangling operation such as a layer of CNOT gates between adjacent qubits. These parameterised circuits form the basis for quantum machine learning and optimisation applications, including those in computer vision and viewpoint planning.

Quantum Circuit Measurement Quantum computation concludes with measurement, which collapses the quantum state to a classical state. Measurement in the

computational basis is represented by projection operators:

$$P_0 = |0\rangle\langle 0|, \quad P_1 = |1\rangle\langle 1| \tag{2.81}$$

For a state $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$, measurement yields outcome 0 with probability $|\alpha|^2$ and outcome 1 with probability $|\beta|^2$. In quantum algorithms for computer vision, measurements are typically performed multiple times (shots) to estimate expectation values of observables:

$$\langle O\rangle = \langle\psi|O|\psi\rangle \tag{2.82}$$

where O is an observable, often expressed as a sum of Pauli operators.

2.3.2 Quantum Algorithms for Image and Vision

The unique properties of quantum computing provide opportunities to develop algorithms that can potentially outperform classical approaches for specific computer vision tasks. This section examines key quantum algorithms with applications in computer vision.

Quantum Search and Optimization

Grover's algorithm [38] provides a quadratic speedup for unstructured search problems, requiring $O(\sqrt{N})$ operations to find a marked item in a database of size N, compared to $O(N)$ for classical algorithms. In robotics, Grover's algorithm can be effectively applied to path planning problems [16].

For a robotic path planning task, the Grover operator $G = U_\psi U_\omega$ is constructed by combining two key components: the diffuser $U_\psi = 2|\psi\rangle\langle\psi| - I$ and the oracle U_ω, which tests potential solutions by marking desired states with phase inversion. The

diffuser can be implemented as:

$$U_\psi = H^{\otimes k} X^{\otimes k} (MCZ) X^{\otimes k} H^{\otimes k} \tag{2.83}$$

where H is the Hadamard gate, X is the Pauli-X gate, and MCZ is the multi-controlled-Z operation.

For path planning in an $n \times n$ grid map, the algorithm encodes both positions using $\eta_{MAP}(n) = \lceil \log_2 n^2 \rceil$ qubits and movement directions with additional qubits. The oracle U_ω is decomposed into M and T blocks, where M performs quantum moves and T tests if the output contains the target cell. This quantum approach has been shown to find optimal paths with only $O(\sqrt{E/S})$ time complexity, where E represents the number of elements in the search space and S the number of solutions [16].

The Quantum Approximate Optimization Algorithm (QAOA) [29] has shown promising applications in image segmentation problems that can be formulated as quadratic unconstrained binary optimisation (QUBO) [61]. For geometric-constrained image segmentation, the problem is first modeled as finding a minimum s-t cut in a directed graph. This is converted to a QUBO formulation with an objective function:

$$F_C = x^T Q x = \sum_{i,j} Q_{ij} x_i x_j \tag{2.84}$$

where $x = [x_1, x_2, \ldots]$ is a binary vector representing graph nodes, and Q is derived from the adjacency matrix of the graph with appropriate modifications. This QUBO problem is then mapped to finding the ground state of a quantum Hamiltonian:

$$H_C = \sum_{i,j} Q_{ij} \frac{1 - Z_i}{2} \frac{1 - Z_j}{2} \tag{2.85}$$

where Z_i and Z_j are Pauli-Z operators. QAOA approximates the ground state through

a parameterised quantum circuit:

$$|\psi(\gamma,\beta)\rangle = U_M(\beta_p)U_C(\gamma_p)\cdots U_M(\beta_1)U_C(\gamma_1)|\psi_0\rangle \tag{2.86}$$

where $U_C(\gamma) = e^{-i\gamma H_C}$ is the cost unitary, $U_M(\beta) = e^{-i\beta H_M}$ is the mixer unitary, and $|\psi_0\rangle$ is an initial state. The parameters γ and β are optimized classically to minimize the expectation value $\langle\psi(\gamma,\beta)|H_C|\psi(\gamma,\beta)\rangle$, which corresponds to the QUBO objective function value. This hybrid quantum-classical approach has been shown to successfully identify optimal segmentation surfaces in both 2D and 3D images, while incorporating smoothness constraints that are essential for realistic delineation of anatomical structures [61]. Importantly, the quantum implementation can identify multiple globally minimal solutions, providing alternative valid segmentations that classical algorithms might miss. QAOA and related variational quantum algorithms represent a promising direction for quantum advantage in computer vision tasks that involve combinatorial optimisation, especially when constraints and domain knowledge need to be incorporated into the segmentation process.

Quantum Machine Learning for Vision

Quantum machine learning algorithms leverage quantum computation to enhance classical machine learning tasks applicable to computer vision [11]. Quantum Principal Component Analysis (QPCA) [64] exponentially reduces the dimensionality of quantum data:

$$\rho \approx \sum_{i=1}^{k} \lambda_i |\psi_i\rangle\langle\psi_i| \tag{2.87}$$

where ρ is the density matrix of the input data, λ_i are the k largest eigenvalues, and $|\psi_i\rangle$ are the corresponding eigenvectors. QPCA can be applied to feature extraction and dimensionality reduction in image processing, potentially offering exponential speedup compared to classical PCA for certain data structures. Quantum Support

Vector Machines [81] use quantum algorithms to compute kernel functions exponentially faster than classical algorithms:

$$K(x_i, x_j) = |\langle \phi(x_i) | \phi(x_j) \rangle|^2 \tag{2.88}$$

where $|\phi(x)\rangle$ is a quantum feature map encoding classical data into quantum states. Quantum Neural Networks [30] use parameterized quantum circuits as models:

$$f_{\boldsymbol{\theta}}(x) = \langle 0 | U^\dagger(\boldsymbol{\theta}) O U(\boldsymbol{\theta}) | x \rangle \tag{2.89}$$

where $U(\boldsymbol{\theta})$ is a parameterized quantum circuit, $|x\rangle$ is the input state, and O is an observable. These quantum machine learning approaches offer potential advantages for image classification, object detection, and other computer vision tasks, particularly as quantum hardware continues to advance [11].

2.3.3 Quantum Approaches for Computer Vision Tasks

Recent research has applied quantum computing to specific computer vision tasks, demonstrating potential advantages over classical approaches for certain problem instances.

Quantum Algorithms for Feature Extraction and Matching

Quantum algorithms for feature extraction leverage quantum properties to achieve computational speedups in image processing tasks. For feature matching and correspondence problems, quantum computing has shown significant promise. Benkner et al. [7] pioneered quantum approaches to graph matching problems, which are fundamental in computer vision for establishing correspondences between features. They reformulated quadratic assignment problems (QAPs) with permutation matrix constraints as quadratic unconstrained binary optimisation (QUBO) problems suitable

for quantum annealing:

$$\min_{X \in \mathcal{P}_n} f(x) := x^T W x + c^T x \tag{2.90}$$

where $\mathcal{P}_n$ is the set of permutation matrices, and W and c encode the matching costs. Their Quantum Graph Matching (QGM) method efficiently implements permutation constraints through innovative quantum Hamiltonian formulations that maximise the spectral gap, increasing the probability of measuring valid permutation matrices in a single run. The approach was successfully demonstrated on a D-Wave quantum annealer for 3D shape matching applications, showing competitive performance against classical relaxation methods.

Building on this foundation, Benkner et al. [8] later introduced Q-Match, an iterative quantum method for solving correspondence problems that addresses key limitations of previous quantum approaches. Unlike earlier methods that directly enforced permutation constraints via penalty terms—which significantly limited success probability on quantum hardware—Q-Match employs a cyclic α-expansion strategy inspired by classical computer vision algorithms. This novel formulation allows the method to update current correspondence estimates through a series of smaller QUBO problems that implicitly enforce the permutation constraints. By solving:

$$\min_{\alpha \in \{0,1\}^m} \alpha^T \tilde{W} \alpha \tag{2.91}$$

where α determines whether to apply specific permutation cycles and $\tilde{W}$ encodes the energy changes, Q-Match efficiently navigates the solution space while guaranteeing valid permutations. This approach enabled the matching of substantially larger point sets (up to 502 vertices compared to previous quantum methods' limit of 3-4 points) and demonstrated performance comparable to classical state-of-the-art methods on the FAUST dataset [14]. The ability to scale to problems an order of magnitude larger represents a significant advancement in practical quantum computing applications for computer vision tasks.

Most recently, Bhatia et al. [9] proposed CCuantuMM, advancing quantum shape matching to handle multiple shapes with guaranteed cycle consistency. Cycle consistency ensures that following correspondences around a chain of shapes returns to the original point—a critical property for multi-shape alignment that previous quantum methods could not guarantee. CCuantuMM introduces a novel approach that reduces the N-shape matching problem to a series of three-shape matching subproblems, enabling linear scaling with the number of shapes. The method integrates visibility and semantic information while carefully managing higher-order quantum terms that would exceed current hardware capabilities. By formulating shape triplet matching as QUBOs that preserve cycle consistency by construction, CCuantuMM can match up to 100 shapes—a significant improvement over previous quantum methods—while producing results competitive with classical state-of-the-art approaches. The work demonstrates that by designing algorithms within hardware constraints and discarding negligible higher-order terms, quantum methods can effectively address complex computer vision problems that were previously inaccessible to quantum computing.

Quantum Approaches for Multi-Model Fitting and Motion Analysis

Recent advances have extended quantum computing to complex computer vision tasks, leveraging adiabatic quantum computing (AQC) to solve combinatorial optimisation problems in visual perception.

Quantum Multi-Model Fitting (QUMF) [31] reformulates geometric model fitting as a quadratic unconstrained binary optimisation (QUBO) problem suitable for quantum annealing. QUMF addresses the task of selecting the best subset of models to explain data points as a set cover problem:

$$\min_{\mathbf{z}\in\{0,1\}^m} \mathbf{z}^T Q \mathbf{z} + \mathbf{s}^T \mathbf{z} \tag{2.92}$$

where $\mathbf{z}$ represents model selection variables, and Q and $\mathbf{s}$ encode both data fidelity

and constraint satisfaction. The method demonstrates competitive performance on multi-homography estimation and motion segmentation tasks, with a decomposed version (DEQUMF) handling larger problem instances by iterative pruning.

Quantum Motion Segmentation [2] introduces the first algorithm for motion segmentation using quantum optimisation. It maps the synchronisation formulation of motion segmentation to a QUBO problem:

$$\min_{\mathbf{z}\in\{0,1\}^d} \mathbf{z}^T(I_{d\times d} \otimes (2Z - 1_{p\times p}))\mathbf{z} \tag{2.93}$$

where Z is the preference-consensus matrix encoding which points belong to the same motion. This approach achieves competitive accuracy with classical state-of-the-art methods while potentially offering advantages for large-scale problems.

Adiabatic Quantum Computing for Multi-Object Tracking [106] formulates tracking as an assignment problem between detections and tracks across frames:

$$\min_{\mathbf{z}} \mathbf{z}^T Q' \mathbf{z} + \mathbf{b}'^T \mathbf{z} \tag{2.94}$$

where $\mathbf{z}$ represents detection-to-track assignments, Q' incorporates pairwise similarity costs, and $\mathbf{b}'$ encodes linear terms and constraints. A unique feature is the adaptive Lagrangian multiplier optimisation to improve the spectral gap and solution probability. The authors demonstrate that their approach is already solvable on current quantum hardware for small examples and achieves competitive performance with state-of-the-art methods.

Common to these approaches is the formulation of vision problems in terms of an Ising model Hamiltonian:

$$H = \sum_{i,j} J_{ij}\sigma_i\sigma_j + \sum_i h_i\sigma_i \tag{2.95}$$

where $\sigma \in \{-1, +1\}$ represents spin variables, and parameters J_{ij} and h_i encode problem-specific costs and constraints. The potential advantage of quantum approaches lies in efficiently exploring the exponentially large solution space of these NP-hard problems, although current hardware limitations restrict experimental validation to small-scale problems.

These approaches represent significant steps toward leveraging quantum computing for vision tasks that involve combinatorial optimisation, indicating the potential for quantum advantage as hardware capabilities advance.

Chapter 3

Semantic-aware Next-Best-View for Multi-DoFs Mobile System in Search-and-Acquisition Based Visual Perception

3.1 Abstract

Efficient visual perception using mobile systems is crucial, particularly in unknown environments such as search and rescue operations, where swift and comprehensive perception of objects of interest is essential. In such real-world applications, objects of interest are often situated in complex settings, making the selection of the 'Next Best' view based solely on maximizing visibility gain suboptimal. We argue that incorporating semantics—providing a higher-level interpretation of perception—can significantly contribute to the selection of viewpoints for various perception tasks. In this study, we formulate a novel information gain that integrates both visibility and semantic gain in a unified form to select the semantic-aware Next-Best-View.

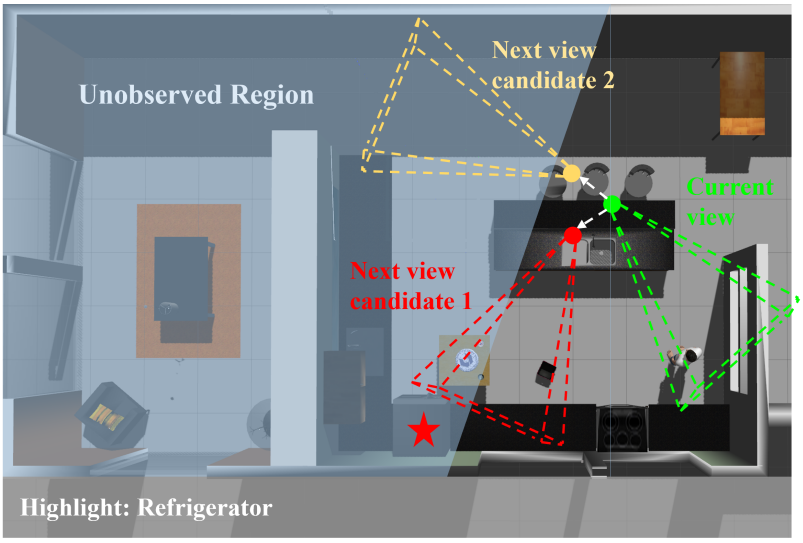

Figure 3.1: When the refrigerator is designated as the object of interest, next view candidate 1 provides higher semantic gain while next view candidate 2 offers higher visibility gain.

We also design an adaptive strategy with termination criterion to facilitate the two-stage search-and-acquisition manoeuvre on multiple objects of interest aided by a multi-degree-of-freedoms (Multi-DoFs) mobile system. To evaluate our approach, we introduce several semantically relevant reconstruction metrics, including perspective directivity and the region of interest (ROI)-to-full reconstruction volume ratio. Simulation experiments demonstrate that our approach outperforms the existing methods by up to 27.46% in the ROI-to-full reconstruction volume ratio and 0.88234 in average perspective directivity. Furthermore, the planned motion trajectory exhibits better perceiving coverage toward the target.

3.2 Introduction

Efficient visual acquisition is a crucial aspect of unknown scene perception using mobile platforms, providing essential information for various manipulation tasks, such as search and rescue operations. Multi-DoFs mobile systems equipped with cameras have

become increasingly popular due to their high mobility and agility, making them well-suited for a wide range of applications. Specifically, autonomous visual acquisition by multi-DoF mobile systems (e.g. unmanned aerial vehicles, UAVs) in unknown and inaccessible environments has proven to be an effective means in search and rescue, reducing the need for professional remote control skills among emergency personnel. However, exhaustive observation is a time-consuming and resource-intensive process. To ensure an efficient visual perception process, it is vital to select adaptive views that provide the most information. Next-Best-View (NBV) was initially presented for an unknown area exploration using the mobile robot [12, 68, 86, 103], usually a finite iteration random tree is grown in the known free space, e.g., Rapidly-exploring Random Tree (RRT), RRT* [60, 53] then the best branch is selected by maximizing the gain (e.g., the amount of unobserved space that can be observed) while minimizing the moving cost (e.g., distance or time cost). After that, it was also adopted to the path planning for single object surface reconstruction [57, 58], online inspection [90, 91, 71] and so on. However, the existing studies determining the next best view focus on information gain by evaluating the visibility of unknown voxels, regardless of their semantics. Unlike the previously mentioned scenarios, visual perception on the objects of interest under complex environments should be semantically selective rather than solely focused on perceiving the unknowns. In other words, the "Next Best" viewpoint in a complex environment cannot be evaluated effectively without the relevant semantic information. In Figure 3.1, semantically informative views should be selected as a higher priority to ensure the efficiency of visual perception on the specific target using mobile systems.

In this work, we propose a semantic-aware NBV scheme for efficient visual perception under complex environments and implement it in a two-stage search-and-acquisition manoeuvre aided by the multi-DoFs mobile system. We develop a novel information gain formulation which integrates both semantic gain and visibility gain. We also design an adaptive strategy to balance these two components so that the mobile

robot can perform both search and acquisition operations on specified semantically important objects. We evaluate the proposed approach using different self-build scenarios in the simulation environment. The results we obtained demonstrate that the proposed approach significantly improves the efficiency of visual perception on specified objects under complex environments through evaluating the reconstruction progress against region of interest (ROI) in volume, ROI-to-full reconstruction volume ratio and perspective directivity. Both the motion planning and reconstruction are implemented based on the voxblox [74] as the map representation, which employs Truncated Signed Distance Fields (TSDFs) to represent the object surface. Then, the RRT* is generated in the observed free space. To the best of our knowledge, this is the first work that investigates semantic-aware NBV for search-and-acquisition-based visual perception by mobile systems, which integrates the contribution from both semantic gain and visibility gain in a unified form for evaluating and selecting the next viewpoint. We demonstrate its capability in the application of different complex environments.

The main contributions of this work include:

1. We present a novel information gain formulation for evaluating the candidate viewpoints that integrates both semantic gain and visibility gain. Such novel formulation can be applied to many other application scenarios in which the visual data acquired contain rich semantics of the complex environment.

2. We design an adaptive strategy with termination criterion to balance the semantic and visibility terms so that the mobile platform can perform an effective two-stage search-and-acquisition manoeuvre on the specified object or multiple objects under the complex environment. The principle behind this two-stage approach can also be applied to scenarios in which the objective of the task can be properly decomposed to facilitate effective implementation.

3. To assess this novel formulation, we also introduce several evaluation metrics to

characterize the system performance and demonstrate the efficiency in perceiving the specific objects under the complex environment while the data acquisition mobile system is undergoing multi-DoFs motion.

The paper content is organized as follows: an overview of the related work and how we step further is presented in Section 3.3. We introduce the proposed system and showcase its effectiveness in visual acquisition on the objects of interest in simulation experiments in Sections 3.4 and 3.5. Finally, we analyze the results obtained and draw conclusions in Sections 3.6 and 3.7.

3.3 Related Work

3.3.1 Mobile System Informative Path Planning for Visual Acquisition

Real-time informative path planning is typically the approach to tackle the non-model-based visual acquisition problem that has no prior information or knowledge of the environment or the target object. Thus, the non-model-based reconstruction needs to plan each view in real-time, which is different from the model-based approach that can be planned offline. There are two main approaches for evaluating new viewpoints in 3D reconstruction: surface-based methods and volumetric methods. Surface-based approaches represent the 3D shape as a mesh and evaluate new views by analyzing the mesh surface [20]. For example, Krainin et al. [53] used a surface-based approach that modelled uncertainty with a Gaussian distribution along each camera ray and measured information gain as the total entropy reduction weighted by surface area. Surface-based methods can evaluate the quality of the 3D model during reconstruction but are computationally expensive due to complex visibility calculations [85]. And more recently, dynamic objects can be accurately reconstructed

by surface-based method [93]. Volumetric methods, on the other hand, represent the 3D shape with voxels, which allow for simple visibility calculations and estimating the probability that each voxel is occupied [43]. Volumetric view evaluation casts rays from the candidate next views through the voxel space to simulate how a camera would sample the scene. Volumetric approaches are computationally more efficient but may not directly provide a surface model of the 3D shape. After that, hybrid methods [58] have combined both surface and volumetric representations to gain the benefits of each. In summary, surface-based 3D reconstruction evaluates new views by analyzing an estimated 3D mesh surface [20, 56], while volumetric methods evaluate new views by casting rays through the voxel representation [43]. The hybrid methods use both representations to improve the efficiency of 3D modelling [58].

3.3.2 Next-Best-View and Related Applications

Next-Best-View is a widely-used greedy method to find local solutions from incomplete information. It was first addressed in the 1980s [22, 67]. It determines the next viewpoint that can observe the largest information gain from the current map iteratively, finally resulting in a completed observation. The information gain metric depends on the specific application and requirements. In order to perceive the unknown volumetric information, besides the early stage approach [4] which simply counts the number of unknown voxels that can be seen, Kriegel et al. [58] use information theoretic entropy to estimate the expected observation. To achieve high completeness of reconstruction, Delmerico et al. [25] proposed proximity count and area factor volumetric information, optimizing the expected gain on a probabilistic map. In 2016, Bircher et al. [12] presented the receding horizon NBV (RH-NBV) that adopts the core idea of model predictive control (MPC). It only executes the first edge in the best branch of RRT to avoid the dilemma of local minima. It also introduces the exponential discount term to penalize the long-distance path. In [84], a novel utility function is formulated as the ratio of gain and cost, minimizing the

number of parameters that need to be fine-tuned. Different from the NBV, which belongs to the sampling-based informative viewpoint planning method, the frontier-based method [104] consistently pursues the boundaries between the explored free and unexplored areas in the occupancy map. The frontier-based method is widely employed in high-speed flight and fast exploration tasks [21, 6], but it is difficult to be generalized to other applications since it cannot have a flexible information gain formulation in NBV fashion. In [50], an uncertainty-guided mapless NBV scheme is proposed, leading to more accurate scene reconstruction. In [69], the predicted fruit shapes are explicitly used to compute information gain for fruit mapping and reconstruction.

Due to the simplicity of the purpose or the environment, there is no existing research that has focused on the contribution of semantics on viewpoint selection in NBV fashion in the application of either unknown exploration or single-object reconstruction. However, under challenging environments (e.g., search and rescue), searching and perceiving the semantic informative views can help us model the object and its surroundings more efficiently. For a more closely related work [54] that proposed a semantically informed scheme for reconstruction. It presents the utility term multiplied by the entropy-formed gain, but does not formulate the semantic term explicitly. It may result in penalization on the unknown exploration capability and would be difficult to generalize to different tasks such as the search–and-acquisition mission.

3.4 Proposed Method

3.4.1 Problem Description

The problem considered in this work is that there exist one or more specific targets $A = \{A_1, ..., A_N\}$, located at the unknown positions in a 3D space $V \subset \mathbb{R}^3$. Unlike the other exploration approaches, the focus is not on observing all the free and occupied

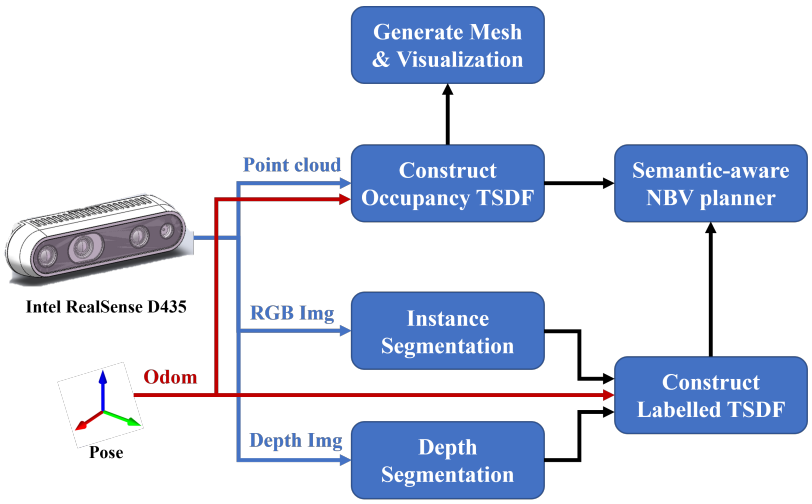

Figure 3.2: Diagram of the system overview: Both the occupancy map and labelled map are constructed in parallel. The Semantic-aware NBV planner takes two maps as the input. The reconstructed mesh is visualized using the occupancy TSDF map.

space (V_{free} and $V_{occ} \subset V$) to achieve $V_{free} \cup V_{occ} = V$. Instead, our approach focuses on searching for and observing each target $A_k \in A$ sequentially, We begin by exploring space V and once we identify a set of occupied voxels $V_{obA-k} \subset V_{occ}$ that have been labelled as c_{tgt-k}, it indicates that the target A_k has been found. Then the acquisition mode is initiated to retrieve not only the volume V_{tgt-k} of each target A_k, but also its surroundings ($V_{sur-k} \subset V_{free}$ or $V_{sur-k} \subset V_{occ}$), with the objective of effectively enlarging the observed volume V_{obA-k} s.t. $\min |V_{res-k}| = \min |V_{tgt-k} - V_{obA-k}|$ utilizing the most extensive accessible perspective coverage within the limited time, where V_{res-k} represents the residue voxels of the target A_k. The searching and acquisition process will be switched to the next target A_{k+1} after achieving the maximum observation on A_k.

3.4.2 System Overview

Two maps are constructed to support the two-stage search-and-acquisition scheme of the proposed semantic-aware NBV framework, an occupancy TSDF map and a labelled TSDF map. Figure 3.2 illustrates the overall system, where the occupancy

TSDF map is incrementally updated from the observations. This is achieved by utilizing the point cloud input from the Intel RealSense D435i depth camera and the real-time pose of the UAV, following the approach proposed in voxblox [74]. The occupancy TSDF map provides information about the occupancy status of the environment, which is essential for the planner to generate RRT* in free space and calculate visibility gain. Additionally, we constructed a labelled TSDF map inspired by the work of Grinvald et al. [37]. This map is generated by raycasting the overlap of the segmentation results from Mask R-CNN [41] based on the RGB input and the depth segments from the depth image input. Depth segments are identified by finding the convex area of depth discontinuity in the depth image. The labelled TSDF map provides a detailed representation of the environment's geometry with semantics, which is useful for the planner to identify the semantic gain. The different types of map representations used in the system are organized into separate layers, with each layer consisting of a set of blocks that are indexed based on their position in the map. It is the same as the structure adopted in voxblox [74]. The mapping between the block positions and their locations is stored in the hash table adopting voxel hashing [73]. Finally, the acquisition result is visualized by the surface model generated from the occupancy TSDF map.

3.4.3 Semantic-aware NBV Framework

From the representation of input TSDF maps, the space V is divided into separate layers of unit-volume cubical voxel $m_o \in \mathcal{M}_o$, $m_l \in \mathcal{M}_l$, where $\mathcal{M}_o$ and $\mathcal{M}_l$ denote the occupancy and labelled map respectively. Each voxel m_{oi} in the occupancy map $\mathcal{M}_o$ consists of an associated centre position p_i, distance d_i, weight w_i and state s_i. The centre position is represented by p_i using the coordinate of its geometric centre, and the voxel's distance from the surface boundary is represented by d_i. In order to minimize the quadratic sensing error of the 3D sensor (e.g., depth camera), we adopt the distance d_i updating approach in [84]. The weight w_i is a metric that refers to

the reliability of the distance's measurement. Here we employ the weighting method formulated in voxblox [74]. The state s_i of each voxel can be marked as "FREE", "OCCUPIED", or "UNKNOWN". For the voxel m_{li} in the labelled map $\mathcal{M}_l$, there are three additional associated properties instance label l_i, semantic category c_i and label confidence l_{ci}. In which the instance label is the index with the highest overlap probability between the binary mask result m_i from Mask R-CNN and the result r_i from depth segmentation. The corresponding semantic category is assigned to c_i if available; otherwise, the default semantics is the background. The label confidence is the number of times the voxel has been labelled as l_i divided by the observation times.

Visibility Gain Formulation

In order to perceive the unknown area and search for the target we are interested in, we define the visibility gain of a branch b associated n nodes $\{b_1, b_2, ..., b_n\}$ in Equation 3.1.

$$Visible(\mathcal{M}_o, b) = \sum_{j}^{n} Visible(\mathcal{M}_o, b_j) \tag{3.1}$$

The visible voxels $\{m_{o1}, m_{o2}, ..., m_{om}\}$ at node b_j are obtained using the intrinsic and extrinsic parameters of the camera. Thus,

$$Visible(\mathcal{M}_o, b_j) = \sum_{i}^{m} V\_gain(\mathcal{M}_o, m_{oi}) \tag{3.2}$$

For simply perceiving the unknown in the 'search' stage, we employ the conventional V_gain formulation that applies a unit increase in gain if the s_i is "UNKNOWN", and there is no gain for "OCCUPIED" or "FREE" voxel.

Semantic Gain Formulation

Similar to the visibility gain, we have the semantic gain for each branch:

$$Semantic(\mathcal{M}_l, b) = \sum_{j}^{n} Semantic(\mathcal{M}_l, b_j) \tag{3.3}$$

Again for each node b_j on the branch,

$$Semantic(\mathcal{M}_l, b_j) = \sum_{i}^{m} S\_gain(\mathcal{M}_l, m_{li}) \tag{3.4}$$

The S_gain for each visible voxel m_{oi} at the specific node b_j is formulated intuitively favours the viewpoints that can observe the new area around the labelled target voxel. As is shown in Equation 3.5.

$$S\_gain(\mathcal{M}_l, m_{li}) = \begin{cases} exp(-\lambda_1 d_{li}),\ if\ s_i = Unknown \\ \eta_{tgt} \cdot f(m_{li}),\ if\ s_i = Occupied \\ \qquad\qquad \&\&\ c_i = c_{tgt-k} \\ exp(-\lambda_2 d_{li}),\ if\ s_i = Occupied \\ \&\&\ c_i != c_{tgt-k}\ \&\&\ c_i != background \\ 0,\ otherwise \end{cases} \tag{3.5}$$

Where η_{tgt} denotes the influence factor that refers to the significance or priority of the voxel with the target label. The exponential term represents the exponential discount on the influence regarding the distance d_{li} of the current voxel to the target volume V_{obA-k} of the target A_k. λ_1, λ_2 are the weight term. In order to minimize the sensing error and refine the voxel that has already been labelled as c_{tgt-k}, we also introduced the function f in Equation 3.6 as its gain.

$$f(m_{li}) = (1 - \frac{|N_{rays}(m_{li}) - N_{exp}|}{1 + |N_{rays}(m_{li}) - N_{exp}|}) \cdot (1 - \frac{w_i}{1 + w_i}) \tag{3.6}$$

Where $N_{rays}(m_{li})$ denotes the number of rays intersecting the m_{li}, which is usually proportional to the inverse of depth quadratically. N_{exp} represents the expected number of intersecting rays. The list $\mathcal{L}_{tgt}$ stores the voxels which have been labelled with the semantic category c_{tgt-k}, and $\mathcal{L}_{tgt}$ is maintained to serve the calculation of the shortest distance d_{li}. Inspired by [65], we maintain the listed voxels (i.e. V_{obA-k}) in a continuous and convex shape.

Adaptive Strategy with Termination Criterion

The proposed method integrates both visibility gain and semantic gain in a consistent format in Equation 3.7.

$$\begin{aligned} Gain(\mathcal{M}_o, \mathcal{M}_l, K) = K \cdot \sum_{j}^{n} Visible(\mathcal{M}_o, b_j) f_o(\delta_{b_{j-1}}^{b_j}) \\ +(1-K) \cdot \sum_{j}^{n} Semantic(\mathcal{M}_l, b_j) f_l(\delta_{b_{j-1}}^{b_j}) \end{aligned} \tag{3.7}$$

Where $\delta_{b_{j-1}}^{b_j}$ denotes the edge distance from node b_{j-1} to node b_j. K is a bool variable controlling the mode preference switching between 'search' and 'acquisition' in our case. f_o and f_l represent the cost function penalizing on the distance of the long edge. It could be in the form of exponential penalty [13, 86], linear penalty [23] or a reciprocal cost to reduce the complexity in tuning parameters [84]. Here, we employ the format in [84]. λ_o, λ_l are the constant parameters.

$$f_o(\delta_{b_{j-1}}^{b_j}) = 1/\lambda_o \delta_{b_{j-1}}^{b_j} \tag{3.8}$$

$$f_l(\delta_{b_{j-1}}^{b_j}) = 1/\lambda_l \delta_{b_{j-1}}^{b_j} \tag{3.9}$$

The state switching of the bool variable K ensures the smoothness of the stage changing between searching and acquisition in the manoeuvre. We also introduce a termination criterion for the acquisition stage to perform the target switching within a

manoeuvre. We separate the semantic gain for each branch into three parts:

$$S_{unknown}(\mathcal{M}_l, b) = \sum_{b_j} \sum_{m_{li}|con1} exp(-\lambda_1 d_{li}) \tag{3.10}$$

$$S_{refine}(\mathcal{M}_l, b) = \sum_{b_j} \sum_{m_{li}|con2} \eta_{tgt} \cdot f(m_{li}) \tag{3.11}$$

$$S_{surround}(\mathcal{M}_l, b) = \sum_{b_j} \sum_{m_{li}|con3} exp(-\lambda_2 d_{li}) \tag{3.12}$$

Where con1 refers to condition 1 $s_i = Unknown$, con2 refers to $s_i = Occupied$ && $c_i = c_{tgt-k}$ and con3 refers to $s_i = Occupied$ && $c_i != c_{tgt-k}$ && $c_i != background$. The planner starts with a zero-size $\mathcal{L}_{tgt}$, K is initially assigned to 1. Once the list $\mathcal{L}_{tgt}$ is expanded, K is switched to 0. The acquisition for one target is terminated if $S_{surround}$ is far greater than the summation of $S_{unknown}$ and S_{refine} for c_{thre} branches. Then K flips to 1, the $\mathcal{L}_{tgt}$ is cleared, and meanwhile, the target label is switched to $c_{tgt-(k+1)}$. Once the list $\mathcal{L}_{tgt}$ is further expanded, K will be set to 0 again. This described strategy can also be represented as Algorithm 1 below.

Algorithm 1 Semantic-aware NBV adaptive strategy with termination criterion

```
K = 1;
while Occupancy TSDF and Labelled TSDF is updated do
  last_size = L_tgt.size();
  Maintain the list L_tgt;
  if L_tgt.size() > 0 then
    Calculate S_unknown, S_target, S_refine
    if L_tgt.size() > last_size then
      K = 0;
    else if Count(S_surround ≫ S_unknown + S_refine) > c_thre then
      K = 1;
      L_tgt.clear();
      Target switch to c_tgt-(k+1);
    end if
  end if
  gain = Gain(M_o, M_l, K);
end while
```

3.5 Experiments and Results

3.5.1 Experimental Setup

Since the planner operates within a perception-planning-execution loop, realistic simulation is compulsory for the evaluation of the proposed scheme. The proposed approach is tested in the simulated world scenes in Gazebo, a 3D dynamic physical robotics simulator. The developed behaviour of the UAV is operating on the Robot Operating System (ROS) [78]. Gazebo-based simulator RotorS [33] is employed to provide an accurate model of the UAV's physics. The underlying control hierarchy of UAV is presented in [52].

The experiments are conducted in the simulation environment, three different settings with two self-build scenes (a narrow collapsed scene with an uneven lighting condition and a larger indoor house scene with ideal lighting condition) in Gazebo with the aid of the individual models by Open Robotics [75] and Google Research [26]. All the experiment results are collected on the machine with an Intel 8C16T Core i7-11700KF at 3.6 GHz × 16 and an NVIDIA GeForce RTX 3060 graphic card.

Collapsed Room Scene

The Collapsed Room Scene used in the experiment is a 10 m × 10 m × 2.5 m map with various furniture, industrial tools and a standing person within the obstacles. The standing person is highlighted as the specific target.

Kitchen and Dining Room Scene

The Kitchen and Dining Room Scene used in the experiment is a 16 m × 10 m × 3.5 m map with common facilities in the family house and a standing person in the corner. The standing person is highlighted as the specific target.

Table 3.1: System parameters for all experiments

Max. velocity	0.8 m/s	Camera RGB FOV	68° × 42°
Max. acceleration	0.8 m/s^2	Camera depth FOV	87° × 58°
Max. yaw rate	$\pi/4$ rad/s	Camera ray length	5 m

Kitchen and Dining Room with Multiple Specified Objects

The third one has the same environment as the Kitchen and Dining Room Scene, but the refrigerator and sink are highlighted as the specific targets.

The basic system parameters are consistent throughout all the experiments described in this study, including the motion dynamics constraints of the UAV and the camera parameters for acquisition, as shown in Table 3.1. For each experiment, the UAV starts at the initial pose where the target is not within the field of view (FOV).

3.5.2 Evaluation Metrics

Since the proposed scheme is designed to execute the search-and-acquisition manoeuvre for the specific target, once the target is found, we aim to acquire the target from multiple accessible viewpoints and achieve maximum reconstruction coverage of the target itself and potential interactions with the surroundings. For most real-world applications, the perception, planning and motion control pipeline of the UAV is expected to execute fully onboard. Thus, algorithmic complexity is also a crucial aspect of the evaluations.

Perspective Directivity

In order to measure whether the UAV consistently perceives the target and its nearest surroundings during the acquisition stage, we calculated the perspective directivity in the target direction D_{tgt-k} for each selected view. The current position p^i, and

orientation o^i of the UAV at view i are obtained from the odometry. The ground truth position of the target p_{tgt-k} is a privileged knowledge that we defined based on the built scene and is not known to the planner. o_P^i and o_Y^i denote the pitch and yaw angle of view i. The directivity at the target direction D_{tgt-k}^i of view i is defined as the cosine of the angle between the camera optical axis O_{cam} and the line connecting the current position p^i with the target position p_{tgt-k}, i.e.

$$O_{cam}^i = [cos(o_P^i) \cdot cos(o_Y^i), cos(o_P^i) \cdot sin(o_Y^i), sin(o_P^i)] \tag{3.13}$$

$$D_{tgt}^i = cos < O_{cam}^i, (p_{tgt-k} - p^i) > \tag{3.14}$$

ROI Reconstruction Progress in Volume and ROI-to-full Reconstruction Ratio

In addition to the perspective directivity, we also periodically record the reconstructed map and analyze the global growth of the reconstruction volume as well as the growth within the region of interest. Again, the region of interest (ROI) is a privileged knowledge that is not known to the planner. It comprises the target V_{tgt-k} and the nearest surroundings V_{sur-k}. The volume ratio of reconstructed ROI over the full reconstructed map indicates the strength of the purpose. A higher ratio implies less reconstruction redundancy in perceiving the target under the complex environment, while a lower ratio indicates more storage consumption on the non-important reconstructions. The target perceiving coverage is analyzed and compared using the motion trajectories with each pose of the UAV and the frustum of the camera.

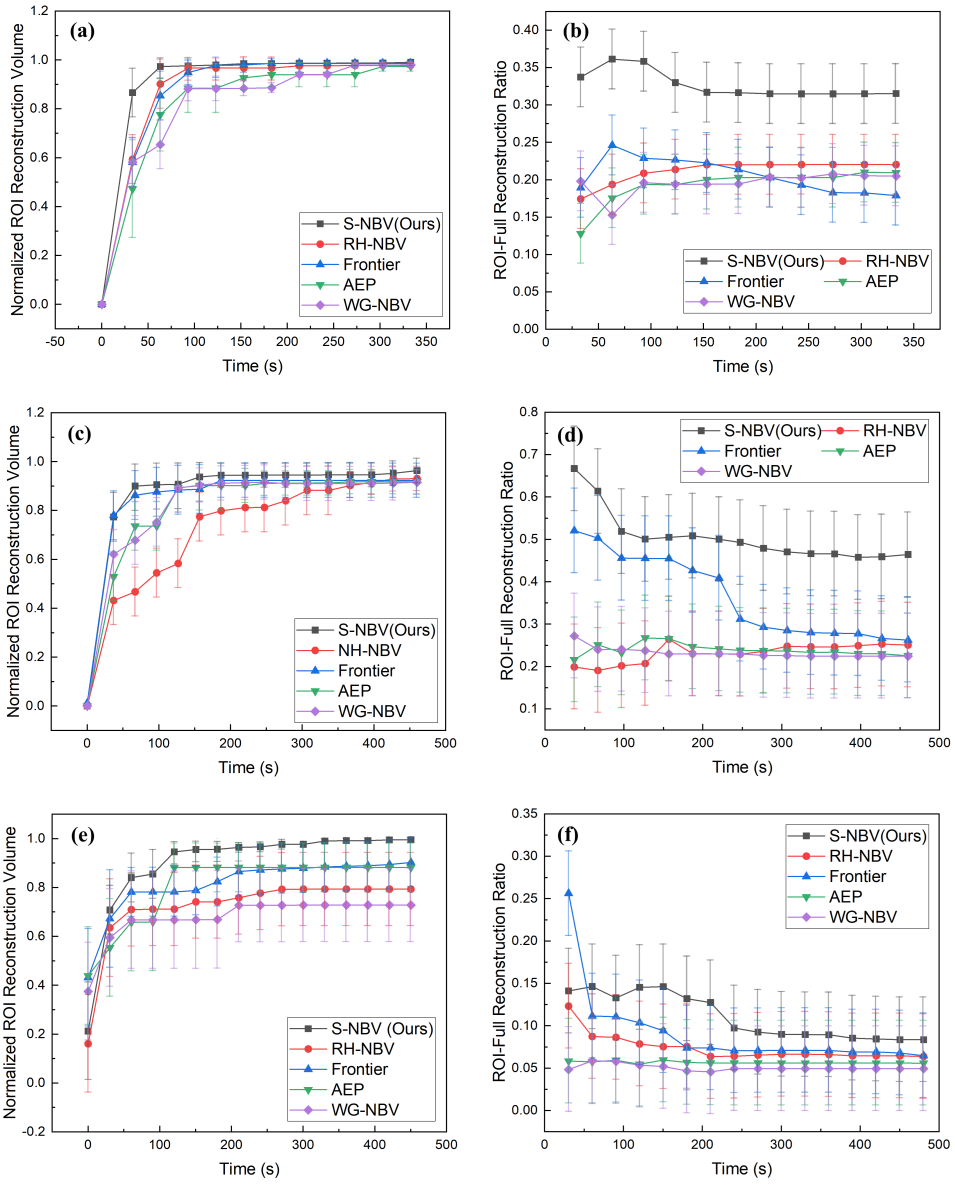

Figure 3.3: Sub-figures (a), (b) are the normalized ROI reconstruction volume and ROI-to-full reconstruction volume ratio verse the simulation time in the Collapsed Room scene. Sub-figures (c) and (d) are the corresponding results in the Kitchen and Dining Room experiment. Sub-figures (e) and (f) are the corresponding results in the Kitchen and Dining Room with Multiple Specified Objects. The performance comparisons between the proposed approach (S-NBV), RH-NBV [13], the frontier-based approach [105], AEP [86] and WG-NBV [71] are presented.

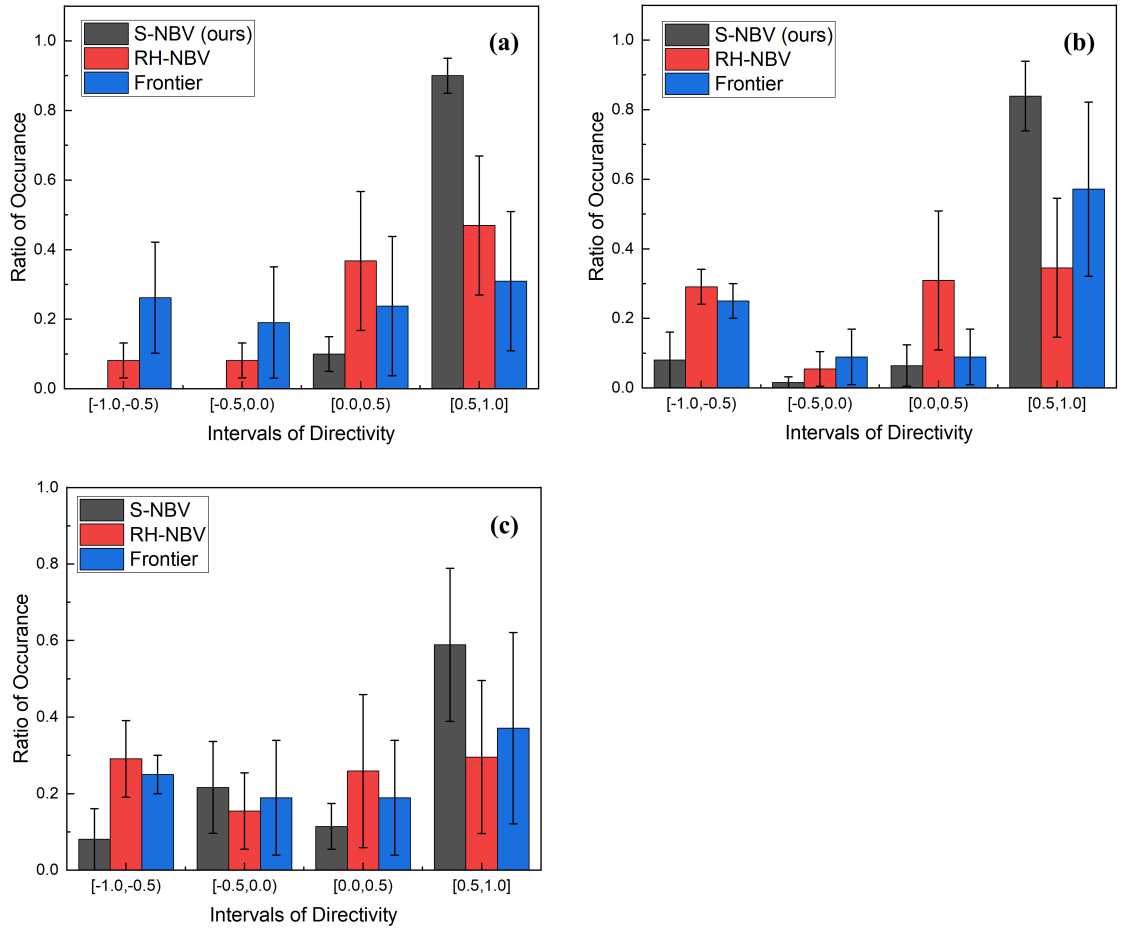

Figure 3.4: Sub-figure (a), (b) and (c) represent the distributions of directivity during the completed experiment in the Collapsed Room scene, Kitchen and Dining Room and Kitchen and Dining Room with Multiple Specified Objects, respectively

3.5.3 Experimental Results

The experiments in this study are conducted in two simulation scenes with three different settings in total. Compared to the smaller scene Collapsed Room, it typically takes longer for the UAV to locate the target in the larger one. In Figure 3.3(a) and (d), the target is well located within the first 30 s in each scenario. In complex scenes with more intricate structures, the proposed approach demonstrates significant advantages over the existing approaches, around 40% to 7% ahead at 60 s. Meanwhile, the frontier-based approach [105] also performed well in Figure 3.3(d) since it has a

strong pattern of exploring along the large and continuous entity, such as walls. It also can be seen in the trajectory in Figure 3.5(h) that the target person in Kitchen and Dining Scene is located closer to the corner of the wall. However, the proposed approach still exhibits 12% to 3% advantages over the frontier-based approach at 60 s. And in both single target reconstruction progress, the ROI reconstruction volume increases every 30 s with the semantic-aware approach, i.e. we are progressively perceiving the ROI, while other approaches show less progress or even remain the same when the normalized reconstruction volume is greater than 0.9. The proposed approach finally achieves the ROI reconstruction progresses at 99.03% and 96.31%, while the other four planners stop at 97.86%, 98.58%, 97.32%, 98.01% in Collapsed Room and 93.12%, 92.27%, 91.73%, 91.51% in Kitchen and Dining Room, respectively. In the experiment involving multiple specified objects, the proposed method demonstrates a significant improvement in reconstruction progress, outperforming the existing planners by up to 27.81% within the first 120 seconds. Ultimately, it achieves a more detailed ROI reconstruction, surpassing the other methods by up to 26.72%, as illustrated in Figure 3.3(g).

Table 3.2: Average Perspective Directivity of Entire Manoeuvre

Scene Name	S-NBV	RH-NBV	Frontier
Collapsed Room	0.90516±0.04	0.44037±0.20	0.02282±0.34
Kit & Din Room	0.76042±0.02	0.13461±0.15	0.45207±0.40
K&D Multi-Obj	0.64037±0.2	0.10375±0.14	0.12442±0.26

In Figure 3.3(b) and (e), the proposed planner prioritizes the ROI reconstruction once the target is located, while other planners focus on perceiving other unknowns, which may belong to semantically redundant areas. The proposed approach achieves an average ROI-to-full ratio of 0.3268 and 0.5046 for each scene, respectively. In comparison, the other approaches achieve averages of 0.2120, 0.2061, 0.1929, 0.1957 in Figure 3.3(b) and 0.2300, 0.3652, 0.2390, 0.2320 in Figure 3.3(e) respectively. For the multi-object scenario in Figure 3.3(h), our method still demonstrates a small

advantage range from 1.99% to 6.00% in ROI-to-full ratio, although multi-target searching and target switching require more observations of the environment.

The distribution of view directivity during the entire manoeuvre is shown in Figure 3.3(c) and (f). The proposed planner exhibits stronger directionality and purposiveness towards the target, with a significant amount of perspective directivity (90% and 83.871%) falling in the interval [0.5, 1.0]. The proposed approach planned more views which have directivity in the range of [-1.0, 0.5) in the Kitchen and Dining scene because it takes more views to search for the target. In Figure 3.3(i), The multi-object directivity distribution spends more views switching between different targets, but the ratio of occurrence in [0.5,1.0] still dominates. In Table 3.14, the proposed approach shows a significant advantage over the other two planners by up to 0.88234 and 0.62581 in the average perspective directivity.

Figure 3.5 shows the two original scenes in Gazebo and planned trajectories by each planner, where the green, red and blue trajectories denote the planned ones by our method, the RH-NBV and the Frontier-based one in order. The green trajectories exhibit the maximum coverage of the viewing angles of the target, circling around the target within the reachable region. Compared to the trajectories of the other two in Figure 3.5(c), (d), (g), and (h), the proposed method also shows strong directionality and purposiveness towards the target and the target's surroundings evidently with its trajectory in Figure 3.5(b) and (f), while the others seem like "wandering aimlessly and enjoying freedom". Due to space limitations, more visualization results are presented in the supplementary material.

In addition, the algorithmic complexity is analysed. The 3D space to be considered is denoted as V, and the resolution of the TSDFs is denoted as r. The number of nodes in the tree is N_T, and the maximum sensing range of the sensor l_{max}^{cam}. The proposed method queries in both the occupancy and semantic map, corresponding complexity $O(2log(V/r^3))$. The complexity of generating an RRT tree can be represented as $O(N_T log(N_T))$, while the query for the best node with O(N_T). Following

the complexity of the collision check in the occupancy TSDF can be denoted as $O(N_T/r^3 log(V/r^3))$. Moreover, both the visibility gain and semantic gain are calculated, considering the volume proportional to $(l_{max}^{cam})^3$, the complexity of evaluating every voxel on the ray $O(l_{max}^{cam}/r)$ by ray casting, resulting in $O((2l_{max}^{cam}/r)log(V/r^3))$. Thus, $O(2(l_{max}^{cam}/r)^4 log(V/r^3))$ for total gain calculation for once. Hence, the total complexity of single planning can be represented as:

$$O(N_T log(N_T) + N_T/r^3 log(V/r^3) + 2N_T((l_{max}^{cam}/r)^4 log(V/r^3))) \tag{3.15}$$

The complexities of RH-NBV and frontier-based method can be denoted as Equation (3.16) and (3.17), respectively. Where M denotes the frontier evaluation, which is proportional to the number of frontier voxels.

$$O(N_T log(N_T) + N_T/r^3 log(V/r^3) + N_T((l_{max}^{cam}/r)^4 log(V/r^3))) \tag{3.16}$$

$$O(N_T log(N_T) + N_T/r^3 log(V/r^3) + M) \tag{3.17}$$

The proposed method shows the highest complexity since we have both the occupancy and semantic status to be queried. We record the execution time of our method on NVIDIA Jetson Xavier NX. The average time costs in planning are shown in Table 3.3. It takes 3.79 s per planning on average on the mobile platform, which is acceptable.

Table 3.3: Measured Execution Time of the Proposed Method on NVIDIA Jetson Xavier NX

Task	Time (s)	Task	Time (s)
Tree Expansion	8.08×10^{-1}	Gain Calculation	4.98×10^{-1}
View Selection	6.97×10^{-1}	TSDFs Update	1.79

3.6 Discussion and Future Work

The proposed semantic-aware NBV scheme in this study demonstrates its advantages in search-and-acquisition manoeuvre under the complex environment over the existing informative path planners in the ROI reconstruction progress, ROI-to-full reconstruction volume ratio and perspective directivity. From the experimental results in Section 3.5, the RH-NBV planner also demonstrates good ROI reconstruction performance in the smaller scene but poor performance in the larger scene. The frontier-based planner exhibits a good ROI-to-full ratio at the beginning of the experiment, which profits from its pattern of pursuing the frontier voxels. After that, the planner intends to find the other unknowns, thus the ROI-to-full ratio drops. The significant difference here is that we are keen on perceiving the region we are interested in instead of pursuing unknown areas. More than 80% of the perspective directivity of the proposed approach falls in the interval [0.5, 1.0] in both scenes, while the other two distribute more average within four intervals. It means we are consistently looking towards the target's location once the target is well-located, while the other two are looking in all directions more evenly. The results also demonstrate the generalization potential of the proposed method by introducing the termination criterion to handle the multi-object search-and-acquisition. As the complexity of the manoeuvre increases, particularly when dealing with multiple objects, the proposed method offers a more exhaustive capture of the objects of interest.

However, there are some limitations to this work. First of all, both the planning and reconstruction processes are based on the same volumetric map. The choice of voxel size is a trade-off between reconstruction precision and planning efficiency, i.e. real-time smooth planning (e.g. around 3 to 6 seconds per planning) results in a compromise in the reconstruction precision. The second one is that the proposed approach is less aggressive in exploring or searching in the larger area than the frontier-based planner. Thus, it takes longer to locate the target as the area increases.

3.7 Conclusion

In this study, we presented a semantic-aware Next-Best-View aided by the multi-DoFs mobile system for autonomous visual perception under the complex and unknown environment. We formulate the novel semantic gain, combined with the conventional visibility gain in a unified form, to evaluate the "Next Best" view among the candidate views with the contribution of semantics. An adaptive strategy is introduced to control the mode switching between 'search' and 'acquisition' on the specific target under the challenging environment, and a termination criterion is designed to handle the target switching in multi-target visual acquisition. The capability of the proposed approach is demonstrated in three different settings in the simulation, achieving improvements of up to 27.46% in the ROI-to-full reconstruction volume ratio and 0.88234 in average perspective directivity. The planned motion trajectory is compared with the ones produced by existing planners, and a better target perceiving coverage is demonstrated evidently.

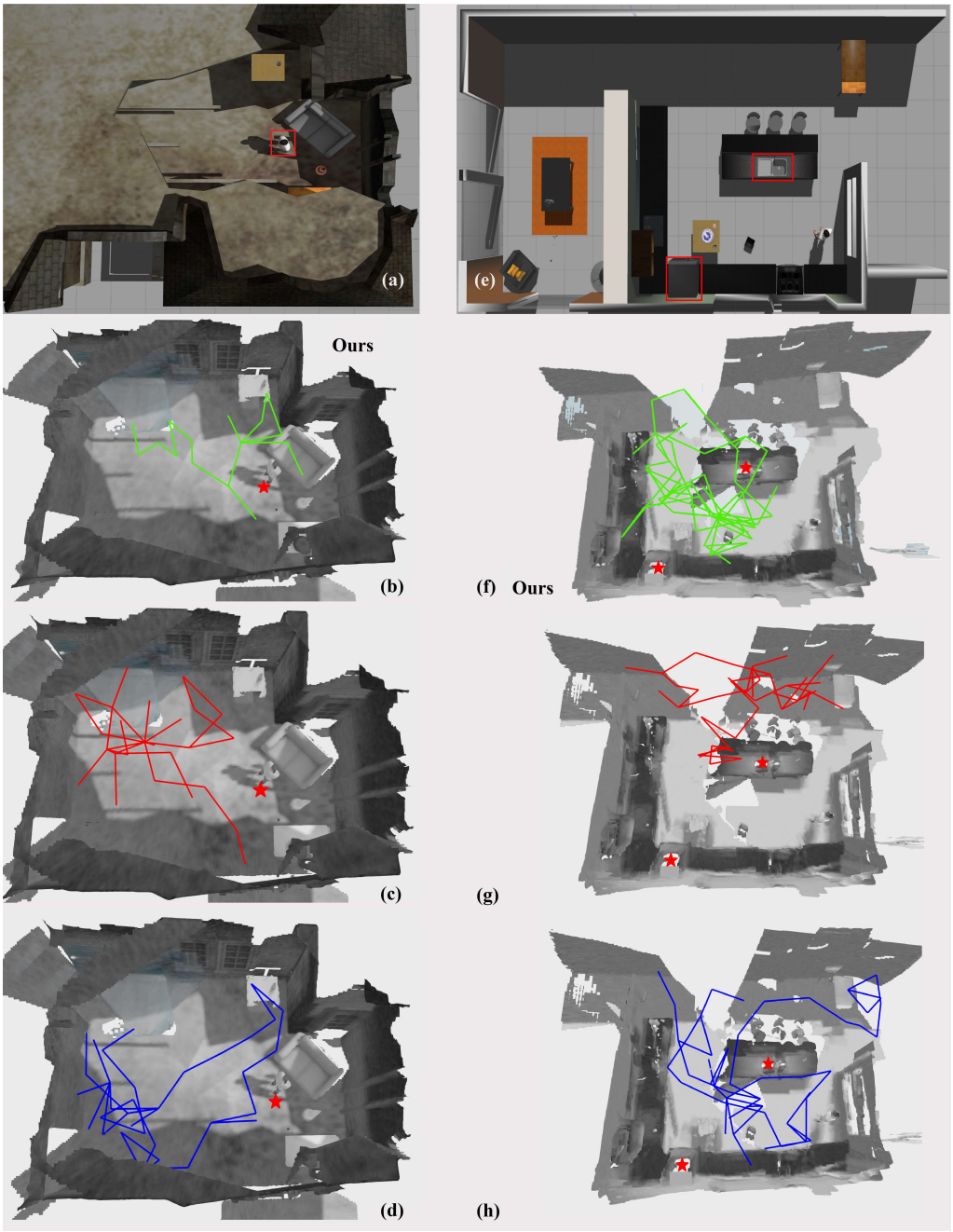

Figure 3.5: Original Scenes in Gazebo (the red square denotes the specified target): (a) Collapsed Room; (e) Kitchen and Dining Room; Sub-figures (b) and (f) show the motion trajectories planned by the proposed approach; (c) and (g) are the trajectories planned by RH-NBV [13]; (d) and (h) show the trajectories planned by the frontier-based approach [105]; The trajectories of different approaches are shown in the same global map, the trajectories of the proposed approach demonstrate the best target perceiving coverage around the target.

Chapter 4

HQC-NBV: A Hybrid Quantum-Classical View Planning Approach

4.1 Abstract

Efficient view planning is a fundamental challenge in computer vision and robotic perception, critical for tasks ranging from search and rescue operations to autonomous navigation. While classical approaches, including sampling-based and deterministic methods, have shown promise in planning camera viewpoints for scene exploration, they often struggle with computational scalability and solution optimality in complex settings. This study introduces HQC-NBV, a hybrid quantum-classical framework for view planning that leverages quantum properties to efficiently explore the parameter space while maintaining robustness and scalability. We propose a specific Hamiltonian formulation with multi-component cost terms and a parameter-centric variational ansatz with bidirectional alternating entanglement patterns that capture the hierarchical dependencies between viewpoint parameters. Comprehen-

sive experiments demonstrate that quantum-specific components provide measurable performance advantages. Compared to the classical methods, our approach achieves 7.9-49.2% higher exploration efficiency across diverse environments. Our analysis of entanglement architecture and coherence-preserving terms provides insights into the mechanisms of quantum advantage in robotic exploration tasks. This work represents a significant advancement in integrating quantum computing into robotic perception systems, offering a paradigm-shifting solution for various robot vision tasks.

4.2 Introduction

In unknown scene perception, determining where to move a camera next - known as the informative view planning problem - can mean the difference between success and failure in critical applications. For instance, in search and rescue operations, inefficient view planning can lead to crucial delays, where every minute matters for survival rates. Similar challenges exist in autonomous navigation and robotic manipulation, where systematic and efficient environment exploration directly impacts task completion time and resource utilization. The Next Best View (NBV) problem represents a fundamental challenge in computer vision and robotic exploration and perception, where the objective is to determine optimal sequential viewpoints to maximize information gained about the environment with each move. Solving the NBV problem effectively can significantly enhance the performance of robotic systems by ensuring that they gather the most relevant and useful visual data with minimal resources.

Next-Best-View was initially introduced for exploring unknown areas using mobile robots [12, 68, 86, 103]. Early approaches can be primarily categorized into sampling-based and deterministic methods. Sampling-based approaches [12, 82, 5] typically employ Rapidly-exploring Random Trees (RRT) or RRT* within known free space [60, 53], generating candidate views and selecting the optimal one based on information gain versus cost metrics. While these methods have shown success in simple

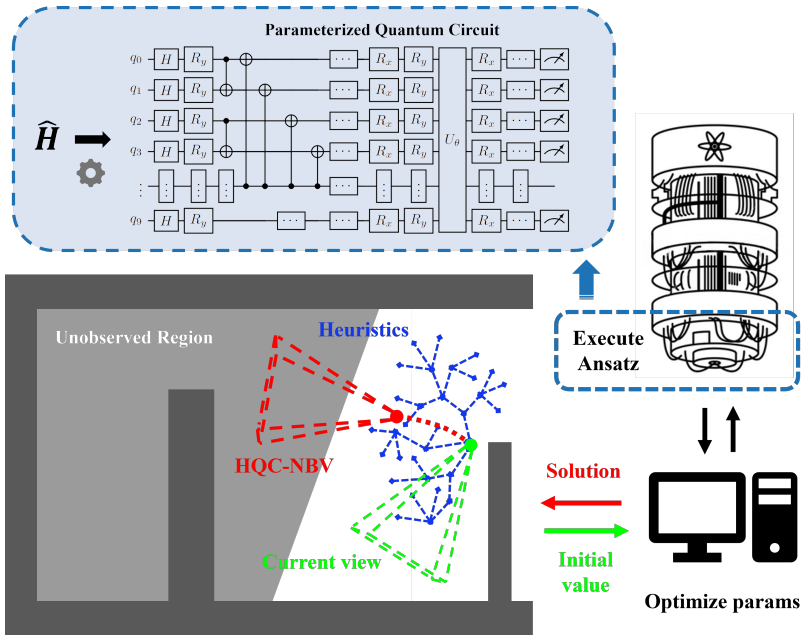

Figure 4.1: Execution logic of our HQC-NBV. Different from the classical approaches, we do not rely on heuristics but leverage quantum superposition to simultaneously evaluate multiple view parameters and quantum entanglement to capture complex dependencies between movement decisions.

environments, they face significant scalability challenges in complex scenarios, often requiring exponentially increasing computational resources with environment size. Deterministic methods [101, 107, 92], on the other hand, rely on heuristics to guide viewpoint selection, focusing on specific metrics or uncertainty minimization. Despite their convenient and widespread deployment in real-world mobile platforms, these classical approaches suffer from fundamental limitations. Heuristic-based methods often struggle to find global optima, particularly in large-scale environments, while sampling techniques frequently result in suboptimal solutions due to their approximative nature of the solution space.

To address these challenges, we explore the potential of quantum computing in solving the NBV problem. Quantum computing has recently demonstrated promising results in various computer vision tasks, including multi-model fitting [31], multi-object tracking [106], motion segmentation [2], and graph matching [7, 8]. The quantum ad-

vantage derives from its ability to leverage quantum phenomena such as superposition and entanglement, enabling efficient exploration of vast solution spaces. The NBV problem is particularly well-suited for quantum approaches due to its combinatorial nature and the presence of complex parameter interdependencies that can be naturally encoded in quantum entanglement structures. Recent quantum implementations in the related problem [31] have shown the potential of adiabatic quantum computing (AQC) in disjoint set cover problems, suggesting a similar potential for view planning optimization.

Our work introduces a novel hybrid quantum-classical framework that combines the computational advantages of quantum systems with the robustness of classical optimization techniques, as shown in Figure 4.1. This hybrid approach aims to overcome the limitations of traditional methods while maintaining collocation with current quantum computing devices. Specifically, our contributions are as follows:

- A novel hybrid quantum-classical approach for informative view planning featuring a Hamiltonian formulation of the NBV problem that effectively maps robotic navigation intuition into the quantum computing paradigm.
- A parameter-centric variational ansatz design with bidirectional alternating entanglement patterns that capture the hierarchical dependencies between view parameters, allowing simultaneous exploration of movement directions, distances, and orientations.
- Comprehensive experimental validations demonstrate the contribution of quantum-specific components (i.e., entanglement architecture and coherence-preserving terms), as well as the robustness and effectiveness of our approach in different experimental settings, achieving 7.9-49.2% higher exploration efficiency compared to the classical methods.

To the best of our knowledge, this is the first study to propose a hybrid quantum-

classical approach for informative view planning, opening new possibilities for efficient robot perception and navigation. The paper content is organised as follows: Section 4.3 and Section 4.4 provide an overview of the related work and essential quantum computing preliminaries. Section 4.5 presents the problem formulation of NBV and our proposed hybrid quantum-classical approach. Section 4.6 presents the implementation of the proposed framework and an additional local strategy. The experimental setup and results are presented in Section 4.7. Finally, we analyse the results and draw the conclusion in Section 4.8.

4.3 Related Work

4.3.1 Informative View Planning

Informative view planning, particularly in non-model-based visual acquisition scenarios where no prior environmental knowledge is available, requires real-time decision-making for each viewpoint. This planning process has evolved significantly since its introduction in the 1980s [22, 67], with approaches generally falling into three main categories: surface-based, volumetric, and hybrid methods. Surface-based approaches represent the 3D environment as a mesh and evaluate viewpoints by analyzing the mesh surface [20]. For instance, Krainin et al. [53] modelled uncertainty using Gaussian distributions along camera rays and quantified information gain through entropy reduction weighted by surface area. While these methods enable direct quality assessment during reconstruction and can handle dynamic objects accurately [93], they are computationally intensive due to complex visibility calculations [85]. Volumetric methods, alternatively, employ voxel-based representations that simplify visibility calculations and occupation probability estimation [43]. These methods evaluate potential views by ray-casting through voxel space, offering computational efficiency at the cost of direct surface modelling capability. To leverage the advantages of both

approaches, hybrid methods [58] combine surface and volumetric representations, achieving a balance between accuracy and efficiency. The Next-Best-View (NBV) paradigm has emerged as a dominant strategy in informative view planning, iteratively selecting viewpoints to maximize information gain. Information gain metrics have evolved from simple unknown voxel counting [4] to sophisticated measures such as information theoretic entropy [58] and proximity-based volumetric information [25]. A significant advancement came with Bircher et al. [12] receding horizon NBV (RH-NBV) approach, which incorporated model predictive control principles to avoid local minima through selective path execution and distance-based penalization. Recent developments include ratio-based utility functions [84] and uncertainty-guided schemes [50] that enhance reconstruction accuracy. While frontier-based methods [104] offer an alternative approach by focusing on boundaries between explored and unexplored areas, particularly effective in high-speed flight scenarios [21, 6], they lack the flexible information gain formulation characteristic of NBV methods. Recent work has extended view planning to specialized applications, such as fruit mapping [69], demonstrating the adaptability of these approaches to diverse scenarios.

4.3.2 Quantum Computer Vision

There is a growing interest in the potential of quantum computing for solving challenging problems in computer vision. The inherent advantages of quantum systems, e.g. superposition, entanglement, and quantum parallelism, offer unique opportunities to tackle computational intensive tasks more effectively. Farina et al. [31] propose Quantum Unconstrained Multi-Model Fitting (QUMF and DEQUMF) method effectively utilizes quantum annealing to optimize the selection of multiple geometric models as a combinatorial optimization, taking advantage of quantum superposition to explore multiple solutions simultaneously. Zaech et al. [106] map the multi-object tracking problem to an Ising model and utilize adiabatic quantum computing (AQC) to find optimal assignments through a quadratic unconstrained binary optimization

(QUBO) approach. Similarly, Arrigoni et al. [2] reformulate the motion segmentation problem into a quadratic unconstrained binary optimization format suitable for adiabatic quantum computing. Benkner et al. [7] reformulate quadratic assignment problems (QAPs) with permutation matrix constraints into a quadratic unconstrained binary optimization format suitable for quantum annealing. Later, they present an iterative method for solving the quadratic assignment problem in shape matching using quantum annealing, achieving high-quality correspondences between non-rigidly transformed shapes [8]. The existing studies focus on reformulating the problem to the Ising model or quadratic unconstrained binary optimization problems and solving them by adiabatic quantum computing. With the recent development of quantum technologies, we are currently within the Noisy Intermediate-Scale Quantum (NISQ) era [59], where new possibilities are emerging for solving complex problems using variational quantum algorithms, quantum approximate optimization algorithms, etc.

4.4 Quantum Computing Preliminaries

4.4.1 Basic Concepts and Properties

Quantum bit (qubit) is the basic computational element in quantum computers. Different from the classical bit, a qubit has the state of a superposition formed by two basis states $|0\rangle = [1\ 0]^T$ and $|1\rangle = [0\ 1]^T$. Qubits can be prepared via different kinds of approaches, including but not limited to photons, trapped ions, Si-based quantum dots, and superconducting circuits. In the NISQ era, the most widely used one is the superconducting circuit approach, leveraging its advantage in scalability.

Superposition refers to the property of the quantum state that can be a linear combination of the corresponding basis states. i.e. a qubit state $|\psi\rangle$ can be described as:

$$|\psi\rangle = c_1|0\rangle + c_2|1\rangle \tag{4.1}$$

c_1 and c_2 are complex numbers, named probability amplitudes, with $|c_1|^2 + |c_2|^2 = 1$.

Entanglement is the critical property for quantum computing. In an entangled system, the state of each qubit is interconnected with the states of the other qubits without space limit, meaning that no qubit can be described independently of the rest of the system.

Measurement the state of a qubit yields one of the basis states, either $|0\rangle$ or $|1\rangle$. The probability of measuring $|0\rangle$ and $|1\rangle$ are given by $|c_1|^2$ and $|c_2|^2$ respectively. Once a measurement is performed, it corresponds to an observation of the qubit, leading to the collapse of its wave function.

4.4.2 NISQ and AQC

Adiabatic Quantum Computing (AQC) is a paradigm that focuses on solving optimization problems by evolving a quantum system from a known initial state to a final state encoding the solution. This is achieved by slowly evolution the system's Hamiltonian $H(t)$, ensuring the ground state of the initial Hamiltonian H_0 evolves to the ground state of the final Hamiltonian H_f.

$$H(t) = (1 - t/T)H_0 + (t/T)H_f \tag{4.2}$$

The Noisy Intermediate-Scale Quantum (NISQ) era represents the current stage of quantum technology, characterized by quantum computers with a moderate number of qubits (typically a few dozen to a few hundred) that are prone to noise and errors. Despite these limitations, NISQ devices show promise in solving practical problems using variational quantum algorithms and quantum approximate optimization algorithms, which are resilient to noise and can be implemented on current hardware. Variational quantum algorithms (VQAs) are a class of hybrid quantum-classical algorithms designed to work on NISQ devices. They use a parameterized quantum circuit

$U(\theta)$ to prepare a quantum state $|\psi(\theta)\rangle$ and then measure an observable O. The goal is to minimize the expectation value $E(\theta)$ of a given Hamiltonian H:

$$E(\theta) = \langle\psi(\theta)|H|\psi(\theta)\rangle \tag{4.3}$$

The parameters θ are optimized using a classical optimizer to find the minimum energy state.

4.5 Methodology

4.5.1 Problem Formulation

CConsider a bounded 2D environment $S \subset \mathbb{R}^2$ containing obstacles $\mathcal{O} = \{O_1, O_2, ..., O_M\}$ to be explored. The camera's viewpoint is given by $v = (x, y, \theta) \in \mathcal{C}$. We aim to find the next best viewpoint that maximizes exploration while minimizing movement cost:

$$\min_{v \in \mathcal{C}} J(v) = -E(v) + \lambda_m M(v) \tag{4.4}$$

subject to:

$$P(v' \to v) \cap \mathcal{O} = \emptyset \tag{4.5}$$

where $E(v)$ denotes the exploration benefit function quantifying potential information gain, $M(v)$ denote the movement cost. λ_m is the weight parameter. And $P(v' \to v)$ represents the path between viewpoints. The exploration benefit function $E(v)$ measures the amount of new information gained by moving to viewpoint v, while $M(v)$ penalizes excessive movement. To evaluate $E(v)$ based on the historical and current views, we maintain an occupancy grid map $\mathcal{M}$ representing the accumulated knowledge of the environment, with each grid in a ternary state: unknown, free space, or occupied.

In practice, this formulation inherently becomes a combinatorial optimization problem. The continuous viewpoint space is discretized into a finite set $\mathcal{V} = \{v_1, ..., v_N\}$ of feasible positions. The collision-free constraint creates a finite feasible set $\mathcal{F} \subseteq \mathcal{V}$. The exploration benefit $E(v)$ depends on discrete visibility relationships derived from the $\mathcal{M}$, where each preference depends on the discrete distribution of unknown regions in the environment. Therefore, our problem becomes:

$$v^* = \arg\min_{v \in \mathcal{F}} J(v) \tag{4.6}$$

where $\mathcal{F}$ is a finite feasible set with solution space complexity $O(|\mathcal{F}|)$, making it naturally suited for quantum algorithms that can leverage superposition to explore multiple discrete solutions simultaneously.

4.5.2 Proposed Method

Problem Hamiltonian Formulation

We formulate the informative view planning problem as a combinatorial optimization task through a carefully designed Hamiltonian $\hat{H}$. This Hamiltonian is constructed as a weighted sum of Pauli operators, where each term encodes specific aspects of the exploration problem:

$$\hat{H} = \sum_i \alpha_i \hat{P}_i \tag{4.7}$$

Here, $\hat{P}_i$ represents a Pauli string (tensor product of Pauli matrices I, X, Y, Z), and α_i is the corresponding coefficient that determines the strength and direction of each term's contribution to the optimization objective. We decompose the Hamiltonian into five functional components:

$$\hat{H} = \hat{H}_{\text{dir}} + \hat{H}_{\text{dist}} + \hat{H}_{\text{adj}} + \hat{H}_{\text{orient}} + \hat{H}_{\text{coh}} \tag{4.8}$$

Our novel approach uses a 10-qubit system to encode viewpoint parameters for efficient exploration. The allocation of qubits is carefully designed to match the physical parameters' importance and range requirements. The main direction is encoded with the first 2 qubits. Distance and adjustment parameters are allocated 2 qubits each, providing sufficient precision for movement magnitudes while keeping the quantum circuit complexity manageable. The camera orientation angle receives 4 qubits to precisely direct the field of view toward information-rich regions. This allocation reflects the hierarchical nature of the exploration task while maintaining an efficient quantum representation with only 10 qubits total. This physical parameter encoding allows for effective quantum representation of navigation decisions.

The directional component $\hat{H}_{\text{dir}}$ encodes exploration preferences using Z operators on the first two qubits:

$$\hat{H}_{\text{dir}} = \sum_{i=0}^{1} \alpha_{dir,i} Z_i + \alpha_{ZZ} Z_0 Z_1 \tag{4.9}$$

Here Z_i denotes the Pauli-Z operator acting on qubit i. Identity operators on unspecified qubits are omitted for notational simplicity. The coefficients are directional exploration value derived from the occupancy map, e represents the unexplored density in each cardinal direction based on the area ratio, where E, N, W, and S correspond to East, North, West, and South directions, respectively:

$$\begin{aligned} \alpha_{dir,0} &= \lambda_{\text{dir},0} \cdot \tanh(e_W + e_S - e_E - e_N) \\ \alpha_{dir,1} &= \lambda_{\text{dir},1} \cdot \tanh(e_N + e_S - e_E - e_W) \\ \alpha_{ZZ} &= \lambda_{\text{diag}} \cdot \tanh(e_{SE} + e_{NE} - e_{SW} - e_{NW}) \end{aligned} \tag{4.10}$$

α_{ZZ} captures directional interdependencies. Here λ_{dir} and λ_{diag} are weighting parameters.

The distance component $\hat{H}_{\text{dist}}$ controls movement magnitude using Z operators on

distance-encoding qubits:

$$\hat{H}_{\text{dist}} = \sum_{i=0}^{1} \alpha_{Z_{\text{dist},i}} Z_{i+2} \tag{4.11}$$

with coefficients proportional to observed obstacle proximity and bit significance:

$$\alpha_{Z_{\text{dist},i}} = \lambda_{\text{dist}} \cdot 2^{-(i+1)} \cdot \frac{d_{\text{obs}}}{d_{\max}} \tag{4.12}$$

The adjustment component $\hat{H}_{\text{adj}}$ provides fine-tuning of the movement direction through Z operators on adjustment-encoding qubits:

$$\hat{H}_{\text{adj}} = \sum_{i=0}^{1} \alpha_{Z_{\text{adj},i}} Z_{i+4} \tag{4.13}$$

where the coefficients are scaled by both bit significance and the remaining unexplored area:

$$\alpha_{Z_{\text{adj},i}} = \lambda_{\text{adj}} \cdot 2^{-(i+1)} \cdot (1 - c) \tag{4.14}$$

Here, $(1 - c)$ represents the proportion of unexplored environment, ensuring that directional adjustments become more precise as exploration progresses. The exponential term $2^{-(i+1)}$ maintains the binary significance hierarchy, with higher-order bits contributing more substantially to the directional refinement.

The orientation component $\hat{H}_{\text{orient}}$ combines target direction terms with exploration-promoting terms:

$$\hat{H}_{\text{orient}} = \sum_{i=0}^{3} \alpha_{Z_{\text{orient},i}} Z_{i+6} + \sum_{i=0}^{3} \alpha_{X_{\text{orient},i}} X_{i+6} + \sum_{m,n} \alpha_{ZZ_{\text{couple}}} Z_m Z_n \tag{4.15}$$

The target direction coefficients are:

$$\alpha_{Z_{\text{orient},i}} = \lambda_{\text{orient-Z}} \cdot 2^{-(i+1)} \cdot \rho \cdot (1 - D) \cdot b_i \tag{4.16}$$

where ρ represents normalized point density, D is angular dispersion, and $b_i \in \{-1, 1\}$

encodes the target angle. Meanwhile, the ZZ operator handles the highest 2-bit coupling between direction and orientation with the coefficient:

$$\alpha_{ZZ_{\text{couple}}} = \lambda_{\text{orient-ZZ}} \cdot (1 - D) \tag{4.17}$$

For high dispersion scenarios, the exploration coefficients are:

$$\alpha_{X_{\text{orient},i}} = \lambda_{\text{orient-X}} \cdot D \cdot (\gamma)^i \tag{4.18}$$

where $\lambda_{\text{orient-X}}$ is a weighting parameter and γ is a decay factor for higher-order bits. Finally, the coherence component $\hat{H}_{\text{coh}}$ maintains quantum advantage through X operators and entangling terms:

$$\hat{H}_{\text{coh}} = \sum_i \alpha_{X_i} X_i + \sum_{(i,j)\in\mathcal{P}} \alpha_{XX_{i,j}} X_i X_j \tag{4.19}$$

where $\mathcal{P}$ represents selected qubit pairs (based on the physical meaning and correlation between view parameters. i.e., direction-adjustment, distance-adjustment, direction-orientation). The entanglement coefficients scale with unexplored area:

$$\alpha_{XX_{i,j}} = \lambda_{\text{coh-XX}} \cdot (1 - c) \tag{4.20}$$

And the X term promotes the exploration proportional to the coverage as well as the estimated information gain with Bresenham algorithm:

$$\alpha_{X_i} = \lambda_{\text{coh-X}} \cdot E_{exp} \cdot \gamma^i \cdot c \tag{4.21}$$

In our Hamiltonian formulation, the ground state of the Hamiltonian corresponds to the optimal next viewpoints for efficient environment exploration. This correspondence is established through the cost function encoding: since we minimize $J(v) = -E(v) + \lambda_m M(v)$ in the classical formulation, the Hamiltonian components

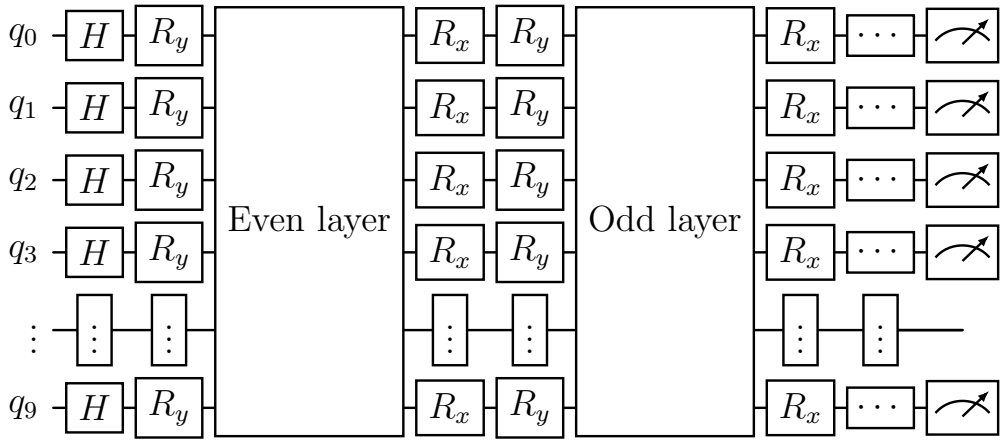

Figure 4.2: Block scheme of the proposed variational ansatz

are designed such that configurations yielding lower classical costs result in lower energy quantum states. Specifically, the directional terms $\hat{H}_{\text{dir}}$ favour movements toward unexplored regions, distance terms $\hat{H}_{\text{dist}}$ penalize excessive movement, and orientation terms $\hat{H}_{\text{orient}}$ promote information-rich viewing directions. The coherence terms $\hat{H}_{\text{coh}}$ maintain quantum superposition to explore multiple solutions simultaneously. Therefore, the ground state $|\psi_0\rangle$ satisfying $\hat{H}|\psi_0\rangle = E_0|\psi_0\rangle$ with minimum eigenvalue E_0 encodes the optimal viewpoint parameters that minimize the objective function $J(v)$. This ground state correspondence enables our variational quantum algorithm to approximate the optimal solution by minimizing the expectation value $\langle\psi(\vec{\theta})|\hat{H}|\psi(\vec{\theta})\rangle$, where the variational parameters $\vec{\theta}$ are optimized to approach the ground state configuration.

Variational Ansatz Design

We develop a multi-layered parameterized quantum circuit $U(\vec{\theta})$ that acts on n native qubits initialized in a uniform superposition state:

$$|\psi(\vec{\theta})\rangle = U(\vec{\theta})|+\rangle^{\otimes n} \tag{4.22}$$

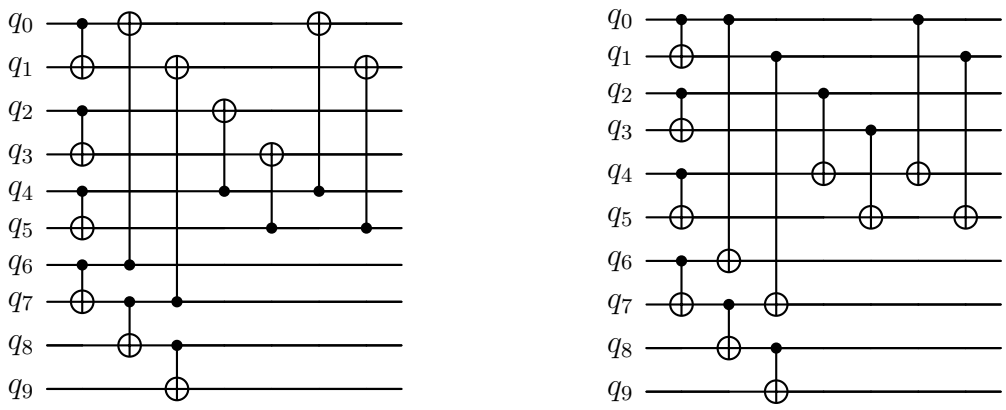

Figure 4.3: (Left) The block module with even index; (Right) The block module with odd index.

where $|+\rangle^{\otimes n}$ represents the uniform superposition state obtained by applying Hadamard gates to all qubits in the $|0\rangle^{\otimes n}$ state.

The circuit architecture consists of $L = 5$ alternating layers of parameterized rotations and structured entanglement operations, the overview is shown in Figure 4.2:

$$U(\vec{\theta}) = U_L(\vec{\theta}_L) \cdots U_2(\vec{\theta}_2) U_1(\vec{\theta}_1) \tag{4.23}$$

Each layer $U_l(\vec{\theta}_l)$ comprises three key components:

$$U_l(\vec{\theta}_l) = U_l^{\text{rx}}(\vec{\theta}_l^{\text{rx}}) \cdot U_l^{\text{ent}} \cdot U_l^{\text{rot}}(\vec{\theta}_l^{\text{rot}}) \tag{4.24}$$

The rotational component $U_l^{\text{rot}}(\vec{\theta}_l^{\text{rot}})$ applies R_y rotations to encode the parameters into the quantum state. These rotations are partitioned according to the parameter groups:

$$\begin{aligned} U_l^{\text{rot}}(\vec{\theta}_l^{\text{rot}}) = &\bigotimes_{i=0}^{1} R_y(\theta_{l,i}^{\text{dir}}) \otimes \bigotimes_{i=2}^{3} R_y(\theta_{l,i}^{\text{dist}}) \otimes \bigotimes_{i=4}^{5} R_y(\theta_{l,i}^{\text{adj}}) \\ &\otimes \bigotimes_{i=6}^{9} R_y(\theta_{l,i}^{\text{orient}}) \end{aligned} \tag{4.25}$$

This structured encoding allows the circuit to independently modulate each parameter while maintaining correlations through subsequent entanglement operations.

The entanglement component U_l^{ent} establishes quantum correlations between qubits following a two-level hierarchical strategy, as is shown in Figure 4.3: intra-group entanglement followed by inter-group entanglement. The intra-group entanglement creates linear chains of CNOT gates within each parameter group:

$$U_l^{\text{intra}} = \prod_{g \in \{\text{dir,dist,adj,orient}\}} \prod_{i=l_g}^{l_g+n_g-2} \text{CNOT}_{i,i+1} \tag{4.26}$$

where l_g and n_g represent the starting position and size of group g, respectively.

The inter-group entanglement establishes connections between parameter groups, with the pattern alternating between even and odd layers:

$$U_l^{\text{inter}} = \begin{cases} \text{CNOT}_{l_{\text{dir}},l_{\text{adj}}} \cdot \text{CNOT}_{l_{\text{dist}},l_{\text{adj}}} \cdot \text{CNOT}_{l_{\text{dir}},l_{\text{orient}}}, & l\%2 = 0 \\ \text{CNOT}_{l_{\text{orient}},l_{\text{dir}}} \cdot \text{CNOT}_{l_{\text{adj}},l_{\text{dist}}} \cdot \text{CNOT}_{l_{\text{adj}},l_{\text{dir}}}, & l\%2 \neq 0 \end{cases} \tag{4.27}$$

This bidirectional entanglement pattern ensures information flow between parameter groups in both forward and reverse directions, facilitating complex correlations while maintaining circuit depth efficiency.

The final component in each layer applies R_x rotations to all qubits:

$$U_l^{\text{rx}}(\vec{\theta}_l^{\text{rx}}) = \bigotimes_{i=0}^{n-1} R_x(\theta_{l,i}^{\text{rx}}) \tag{4.28}$$

These rotations introduce non-commutativity with respect to the R_y rotations and the Z-based measurement observables, enhancing the circuit expressivity and enabling exploration of a larger subspace of the Hilbert space.

Optimization Process

The variational quantum circuit is optimized to minimize the expectation value of the cost Hamiltonian:

$$\vec{\theta}^* = \arg\min_{\vec{\theta}} \langle \psi(\vec{\theta}) | \hat{H} | \psi(\vec{\theta}) \rangle \tag{4.29}$$

where $\hat{H}$ is the problem-specific cost Hamiltonian incorporating exploration objectives, environmental constraints, and quantum coherence requirements. The optimization is performed using an adaptive Simultaneous Perturbation Stochastic Approximation (SPSA) algorithm, which efficiently handles the high-dimensional parameter space of the variational circuit while being robust to the statistical noise inherent in quantum measurements. The adaptive learning rate mechanism adjusts the optimization step size based on progress metrics and stagnation detection:

$$\eta_{t+1} = \eta_t + \mu m_t + (1 - \mu)\Delta\eta_t \tag{4.30}$$

where μ is the momentum coefficient, m_t is the momentum term at iteration t, and $\Delta\eta_t$ is the learning rate adjustment based on recent optimization progress.

4.6 Implementation Details

The HQC-NBV system optimizes viewpoint selection for exploring unknown environments with obstacles. The process begins by initializing the scene with an initial viewpoint. In each iteration, before the coverage threshold is reached, the system updates the set of observed points by checking visibility from the current viewpoint. The variational ansatz is initialized for the current viewpoint and the problem Hamiltonian is constructed to encode the exploration objectives. The parameters are optimized using a Variational Quantum Eigensolver (VQE) with an adaptive SPSA optimizer, where the Hamiltonian-driven optimization is augmented with auxiliary constraints

Algorithm 2 Hybrid Quantum-Classical NBV System

Require: $S, v_0, \phi_{FOV}, d_{max}, \tau_{coverage}$
Ensure: $\mathcal{V} = \{v_0, v_1, \ldots, v_n\} : \mathcal{C}(\mathcal{V}) \geq \tau_{coverage}$
$\mathcal{V} \leftarrow \{v_0\}$, $\mathcal{M} \leftarrow$ InitializeMap(S)
$\mathcal{M} \leftarrow$ UpdateObservation($\mathcal{M}, v_0$)
while $\mathcal{C}(\mathcal{M}) < \tau_{coverage}$ **do**
  $\hat{H} \leftarrow$ ConstructHamiltonian($\mathcal{M}, v_t$)
  $U(\vec{\theta}) \leftarrow$ CreateParameterizedCircuit(n, L)
  $|\psi_0\rangle \leftarrow H^{\otimes n}|0\rangle^{\otimes n}$
  $\vec{\theta_0} \leftarrow$ InitializeParameters()
  for $i = 0$ **to** $N_{iter} - 1$ **do**
    $c_i \leftarrow \langle\psi(\vec{\theta_i})|\hat{H}|\psi(\vec{\theta_i})\rangle + f(\vec{\theta_i})$
    $g_i \leftarrow$ AdaptiveSPSA_GradientEsti.$(U, \hat{H}, \vec{\theta_i})$
    $\eta_i \leftarrow$ AdaptiveLearningRate$(c_0, ..., c_i)$
    $\vec{\theta}_{i+1} \leftarrow \vec{\theta_i} - \eta_i \cdot g_i$
  end for
  $\vec{\theta^*} \leftarrow \vec{\theta}_{N_{iter}}$
  $\vec{z} \leftarrow$ Measure_Z_Expectations$(U(\vec{\theta^*})|\psi_0\rangle)$
  $v_{next} \leftarrow$ DecodeParameters$(\vec{z}, v_t, \mathcal{M})$
  $valid \leftarrow$ ValidateTrajectory$(v_t, v_{next}, \mathcal{M})$
  if $valid$ **then**
    $v_{t+1} \leftarrow v_{next}$
  else
    $v_{t+1} \leftarrow$ ClassicalFallbackStrategy$(\mathcal{M}, v_t, v_{next})$
  end if
  $\mathcal{V} \leftarrow \mathcal{V} \cup \{v_{t+1}\}$
  $\mathcal{M} \leftarrow$ UpdateObservation$(\mathcal{M}, v_{t+1})$
end while
return $\mathcal{V}$

to ensure solution feasibility. The optimal parameters are decoded from Z expectation values to determine the next viewpoint, and trajectory validation ensures the new viewpoint lies within the observed area in $\mathcal{M}$ and avoids obstacle collisions. If the trajectory is invalid, we select the furthest valid position along the moving direc-

tion using a classical fallback strategy. The process iteratively executes to find out a sequence of optimal viewpoints, as detailed in Algorithm 2.

4.7 Experiments and Results

To demonstrate the effectiveness and robustness of the proposed HQC-NBV, we conduct a series of experiments on scenes with different areas and different obstacles. To examine the design of variational ansatz, we also conduct specialized experiments to isolate and quantify the contributions of key quantum components in our hybrid approach aiming to provide insights into how quantum characteristics—specifically entanglement patterns and coherence-preserving terms—impact exploration performance. In this study, the proposed method is implemented using the Qiskit framework, and all the experiments are performed on the Qiskit Aer backend simulator [49]. The camera parameters used in this study consistently respected a field of view (FOV) $2\pi/3$, and a maximum ray distance of 8 units. The starting view is initialized at a non-collision position.

4.7.1 Experimental Setup

We designed three distinct scenes with varying levels of complexity to comprehensively evaluate the robustness and scalability of the proposed methods.

- Scene 1 (S1): Surrounding obstacles in area 20×20 unit^2, Fig.4.4(a);
- Scene 2 (S2): Central obstacle in area 20×20 unit^2, Fig.4.4(b);
- Scene 3 (S3): Complex walls with surrounding obstacles in area 20×20 unit^2, Fig.4.4(c);

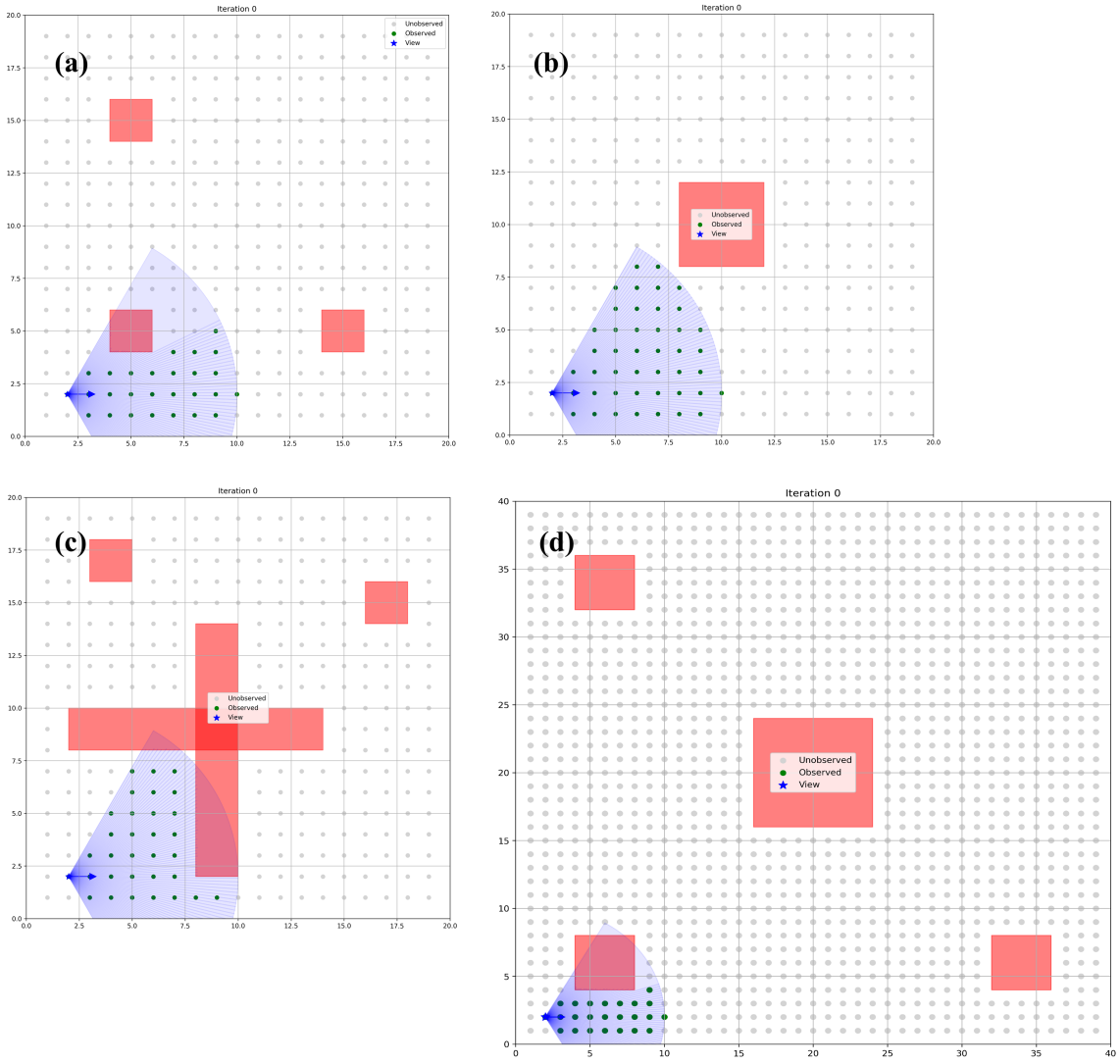

Figure 4.4: Visualized experimental scenes: (a) Scene 1; (b) Scene 2; (c) Scene 3; (d) Scene 4.

- Scene 4 (S4): Surrounding and central obstacles in larger area 40×40 unit^2, Fig.4.4(d).

To investigate the impact of entanglement structure in our approach, we implemented four variants of Ansatz architecture while maintaining identical Hamiltonian formulations and classical optimization procedures:

- Full Architecture (FA): Our proposed bidirectional alternating entanglement

pattern with both intra-group and inter-group CNOT gates;

- Non-Entangled (NE): A circuit with the same number of parameterized rotations but without any entangling gates, equivalent to independent qubit rotations;

- Intra-Group Only (IG): Preserving parameter group coherence through intra-group entanglement but removing connections between different parameter groups;

- Inter-Group Only (EG): Maintaining only the connections between parameter groups while removing intra-group entanglement.

To assess the contribution of quantum coherence-preserving terms in our cost Hamiltonian, we conducted a systematic ablation study by modifying the $\hat{H}_{\text{coh}}$ component:

- Complete Hamiltonian (CH): Including all coherence-preserving terms (X and XX operators with adaptive weights);

- No Coherence Terms (NC): Removing all $\hat{H}_{\text{coh}}$ components, retaining only the problem-encoding Z-based terms;

- Single-Qubit X Only (SQX): Preserving the $\sum_i \alpha_{X_i} \hat{X}_i$ terms while removing two-qubit XX interactions.

In addition to these, we also comprehensively evaluate the performance of our approach against two classical exploration approaches, RH-NBV and the frontier-based method, regarding the exploration coverage ratio, path length and exploration efficiency. Additionally, we conduct comparative experiments with established classical optimization algorithms to assess the advantage of our approach over general-purpose optimization approaches.

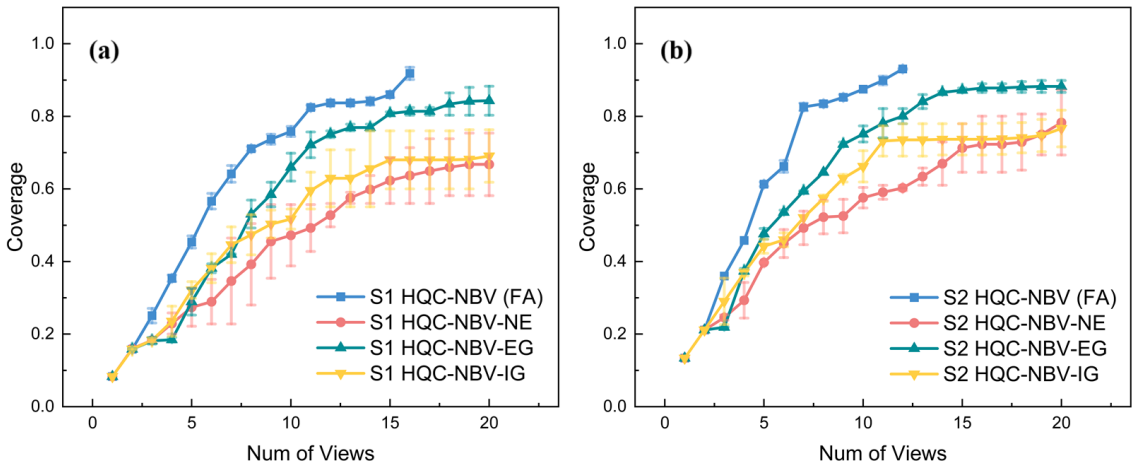

Figure 4.5: The effectiveness of entanglement architecture on the exploration performance: (a) coverage ratio in Scene 1; (b) coverage ratio in Scene 2.

4.7.2 Experimental Results

Figure 4.5 presents the comparative results, demonstrating that the full bidirectional entanglement architecture consistently outperformed reduced-entanglement variants. Notably, the non-entangled circuit required an average of 61.11% and 57.14% more views to achieve 65% coverage in S1 and S2 respectively, highlighting the significant role of quantum correlations in effective exploration planning. The inter-group-only variant performed better than the intra-group-only variant in both scenes, suggesting that maintaining cross-parameter entanglement between parameter groups is more critical than the entanglement within parameter groups. The intra-group entanglement architecture also contributes to the improvement compared to the non-entanglement variant because of the intrinsic connection between qubits within the logical groups of the informative view planning. Figure 4.6 illustrates that the absence of coherence-preserving terms led to frequent entrapment in local minima, with the no-coherence variant failing to achieve above 68.46% coverage and 65.77% on average in Scene 1 and Scene 2, respectively. The performance degradation was most pronounced in later exploration stages (coverage $>$ 50%), where remaining unexplored

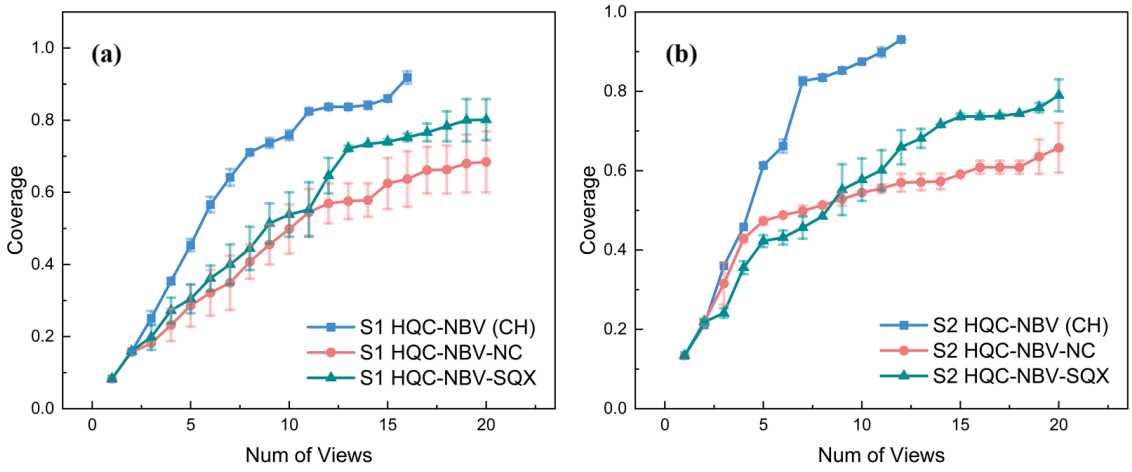

Figure 4.6: The evaluation of coherence-preserving term on the exploration performance: (a) coverage ratio in Scene 1; (b) coverage ratio in Scene 2.

regions became sparse and disconnected. The single-qubit X-only variant demonstrated intermediate performance, maintaining reasonable exploration capabilities but showing reduced ability to escape local minima in complex scenarios. This suggests that while single-qubit superposition maintenance contributes to exploration effectiveness, the two-qubit coherence terms play a crucial role in coordinating parameter updates across different aspects of the navigation decision. Figure 4.7 demonstrates the robustness and scalability of our approach. Our approach performs an efficient exploration in S1, S2 and S3 within 15 viewpoints. The coverage growth in S2 is more dramatic than that in S1 and S3 due to the simplicity of the scene. The exploration in S4 requires 56 viewpoints to achieve comparable coverage, which is roughly four times the number required for S1, S2 and S3. This scaling factor meets the simple 4:1 ratio of environment sizes (S4 is four times larger in area than S1, S2 and S3), suggesting that the proposed approach does not degrade with the increase in environment size and complexity. HQC-NBV consistently outperforms classical methods across all evaluation scenarios. In Scene 1, our method achieves 92.85% coverage within 16 viewpoints, while RH-NBV and frontier-based methods reach only 80.54% with the same views. Scene 2 demonstrates advantages with HQC-NBV achieving 93.02%

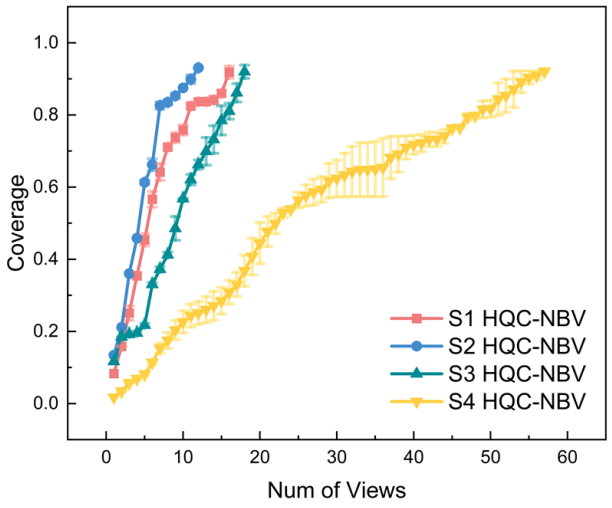

Figure 4.7: The coverage against the number of views of our approach in different scenes

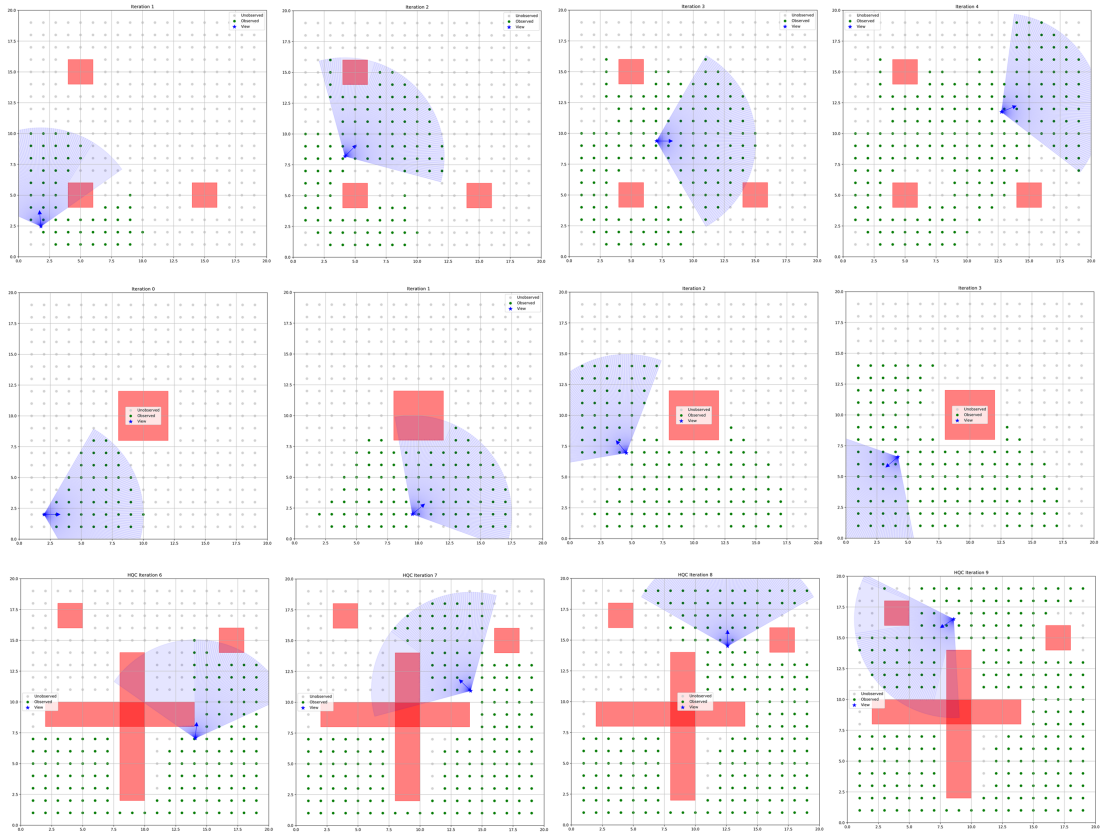

Figure 4.8: Sample continues views planned by HQC-NBV in S1, S2 and S3. The red rectangles denote the obstacles, the blue wedges represent the FOV of the viewpoint, and the green dots are the observed grid.

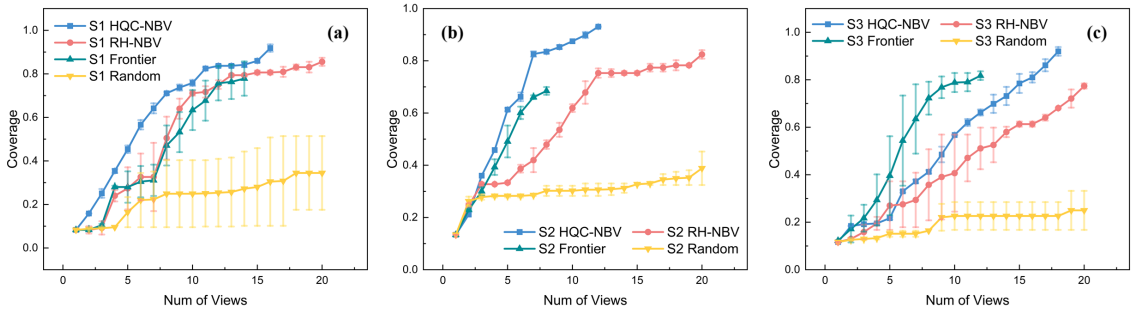

Figure 4.9: Comparison of coverage ratio progression between HQC-NBV, RH-NBV, and Frontier-based approaches in: (a) Scene 1; (b) Scene 2 and (c) Scene 3.

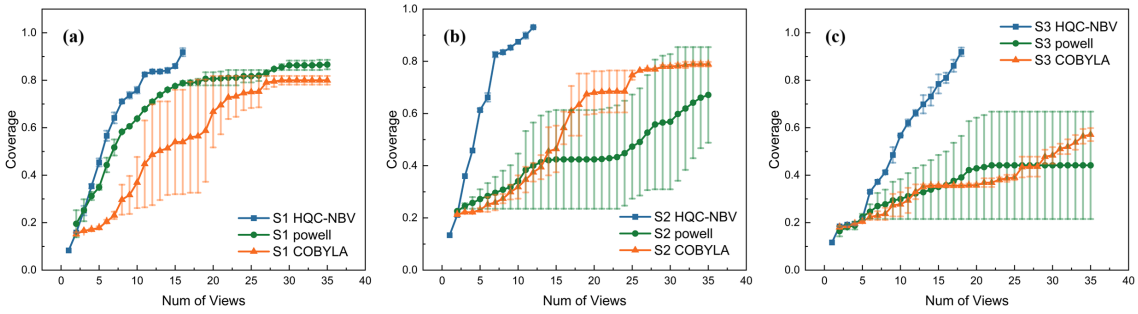

Figure 4.10: Comparison of coverage ratio progression between HQC-NBV and classical optimization methods Powell and COBYLA in: (a) Scene 1; (b) Scene 2 and (c) Scene 3.

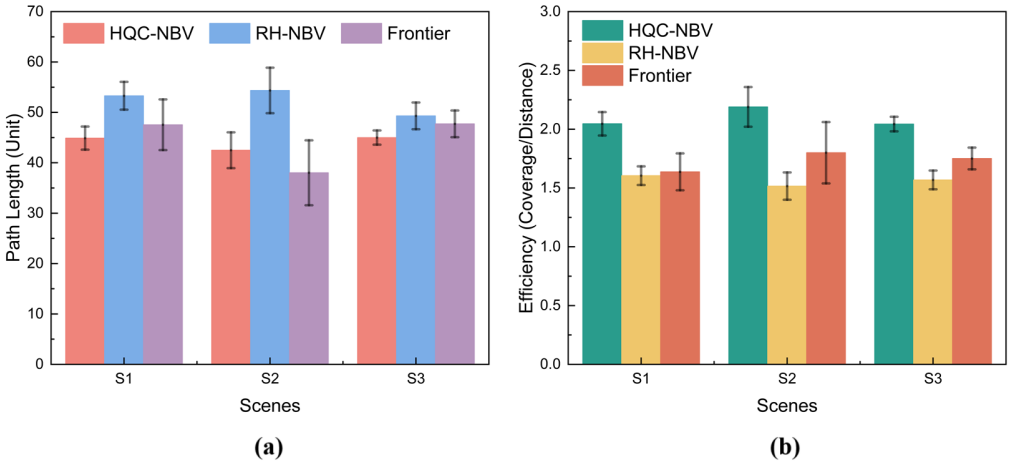

Figure 4.11: Performance metrics comparison: (a) Total path length across different scenes; (b) Exploration efficiency measured as coverage-to-distance ratio.

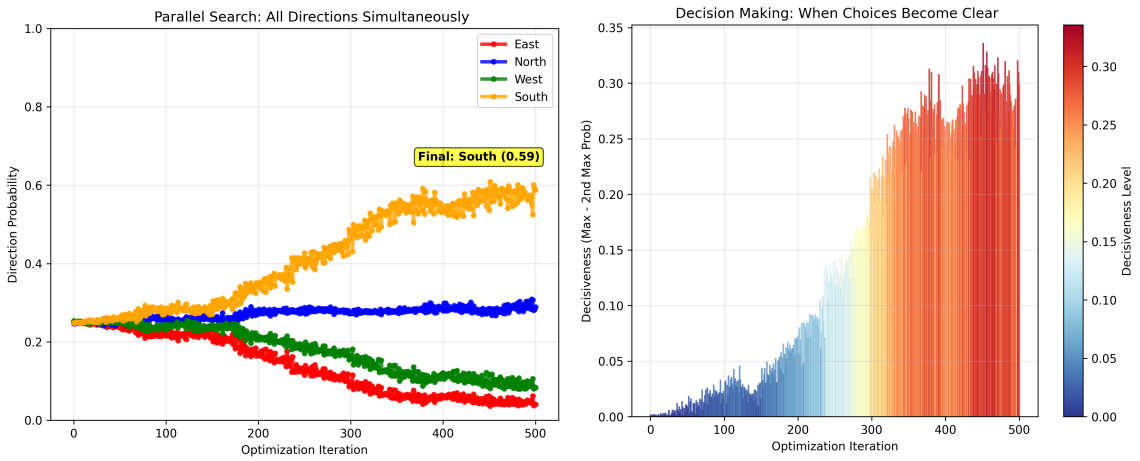

Figure 4.12: Optimization insight (direction qubits example): (Left) Evolution of directional probabilities; (Right) Decision formation process measured by decisiveness

coverage in 12 viewpoints compared to RH-NBV requiring twice as many viewpoints for 78.27% coverage. In the most challenging Scene 3, while frontier-based methods shows initial advantages in open spaces, they encounter early termination at 81.75% coverage due to disconnected unexplored regions. Our approach maintains consistent progress, achieving 91.97% coverage within 18 viewpoints as shown in Figure 4.9. Beyond coverage efficiency, HQC-NBV demonstrates 9.60-27.92% reduction in total path length and 16.19-30.75% higher exploration efficiency compared to classical approaches. The method also exhibits superior stability across multiple runs, particularly during the critical middle phase of exploration, as is shown in Figure 4.11. Comparison with traditional optimization methods (Powell and COBYLA) shows even more significant advantages, as shown in Figure 4.10. While classical optimizers frequently become trapped in local minima with coverage plateauing below 67.1%, HQC-NBV maintains superior performance across all complexity levels, highlighting the scalability advantages of our quantum-enhanced framework. The proposed approach outperforms traditional optimizers since we reformulate the conventional discrete information gain, which is ill-suited for optimization, into a continuous, differentiable Hamiltonian expectation. This expectation provides smooth gradients with respect to the variational parameters θ, while the underlying Hamiltonian encodes

the exploration objectives through structured Pauli operators with smooth coefficient functions.

Figure 4.12 provides crucial insights into the optimization process underlying HQC-NBV performance. The left figure demonstrates our method's parallel search capability, taking the directional qubits as the example, where all four directional choices (East, North, West, South) are simultaneously evaluated through quantum superposition. Unlike classical methods that evaluate directions sequentially, the quantum approach maintains probability distributions for each direction throughout the optimization process. The evolution shows how initially uniform probabilities (0.25 each) gradually converge toward the optimal choice, with South direction emerging as the final selection with 59% probability after 500 iterations. The right figure denotes the quantum decision formation process, showing how quantum superposition gradually resolves into a definitive choice. The color-coded bars represent decisiveness levels ($P_{max} - P_{second}$), progressing from superposition to definitive decisions. This visualization captures the optimization transitions from exploring all possibilities simultaneously to converging on the optimal solution. This figure effectively illustrates the optimization simultaneous exploration of all possibilities followed by gradual convergence to optimal solutions. The smooth probability evolution and stable decision formation explain the observed performance improvements and reduced variance in exploration outcomes.

Figure 4.8 presents two groups of continuous views planned by HQC-NBV in S1, S2 and S3, respectively. It demonstrates the effectiveness of our viewpoint planning algorithm in progressively expanding coverage across different scenes, over several iterations. The samples start from an initial status where only a few areas have been explored, the algorithm efficiently selects viewpoints that maximize the coverage of unobserved regions. With each subsequent iteration, the coverage area grows substantially with feasible movement.

4.8 Conclusion

In this paper, we present a paradigm-shifting scheme in view planning, namely, Hybrid Quantum-Classical Next-Best-View (HQC-NBV) for autonomous exploration tasks. Our approach features a multi-component quantum Hamiltonian and a variational circuit with bidirectional entanglement patterns. Experiments across various environments demonstrated that quantum-specific elements provide measurable contributions, with our entanglement architecture and coherence-preserving terms significantly enhancing exploration efficiency. Compared to the classical approaches, our method consistently achieved higher coverage rates (up to 95.8%) with 7.9-49.2% higher exploration efficiency against travel lengths. Moreover, our approach demonstrated excellent scalability and robustness across environments of increasing size and complexity. The framework achieves high-efficiency exploration while being compatible with current NISQ devices. This work paves the first step toward integrating quantum variational algorithms for solving robot vision problems.

Chapter 5

CSDNet: Detect Salient Object in Depth-Thermal via A Lightweight Cross Shallow and Deep Perception Network

5.1 Abstract

While we enjoy the richness and informativeness of multimodal data, it also introduces redundancy of information and distractions. To achieve optimal domain interpretation with limited resources, we propose CSDNet, a lightweight **C**ross **S**hallow and **D**eep Perception **Net**work designed to integrate two modalities with less coherence, thereby discarding redundant information or even modality. We implement our CSDNet for Salient Object Detection (SOD) tasks in robotic perception, emphasizing that effective integration of the depth-thermal (D-T) modality can facilitate mobile-friendly privacy-preserving visual tasks. The proposed method capitalises on spatial information prescreening and implicit coherence navigation across shallow and deep

layers of D-T modality, prioritising integration over fusion to maximise the scene interpretation. To further refine the descriptive capabilities of the encoder for the less-known D-T modalities, we also propose the Segment Anything Model (SAM) assist framework to guide an effective feature mapping to the generalised feature space. Our approach is tested on the VDT-2048 dataset, leveraging the D-T modality outperforms those of SOTA methods using RGB-T or RGB-D modalities for the first time, achieves comparable performance with the RGB-D-T triple-modality benchmark method with a runtime speed improvement of 5.97 times and a reduction in required FLOPs to 0.36% of the original.

5.2 Introduction

In recent decades, various multimodality techniques have shown significant advancements in this learning era, especially in the field of robotic perception. Within the scope of multimodality techniques, it is commonly observed that the performance of models tends to improve with an increasing number of modalities [46]. However, this inevitably leads to higher costs, higher computational demands, and unpredictable noise. In contrast, crossing off a modality with information density may also cause the loss of crucial scene interpretation [97], presenting a trade-off dilemma. Nevertheless, we identify an entirely new pivot beyond this lever: the integration of modalities with low coherence to achieve broader domain coverage, exploring the possibility of breaking the constraint of 'adding modalities for better interpretation'.

In the existing multimodal studies, RGB typically serves as the primary modality with high priority since it is widely recognised as a senior modality due to its rich texture information, colour and high-resolution spatial details. Depth and thermal modalities are considered subsidiary modalities since they provide relatively limited but more specialised information with particular significance. Conventional multimodal approaches on visual perception tasks by RGB-D [110, 115, 111, 112], RGB-T

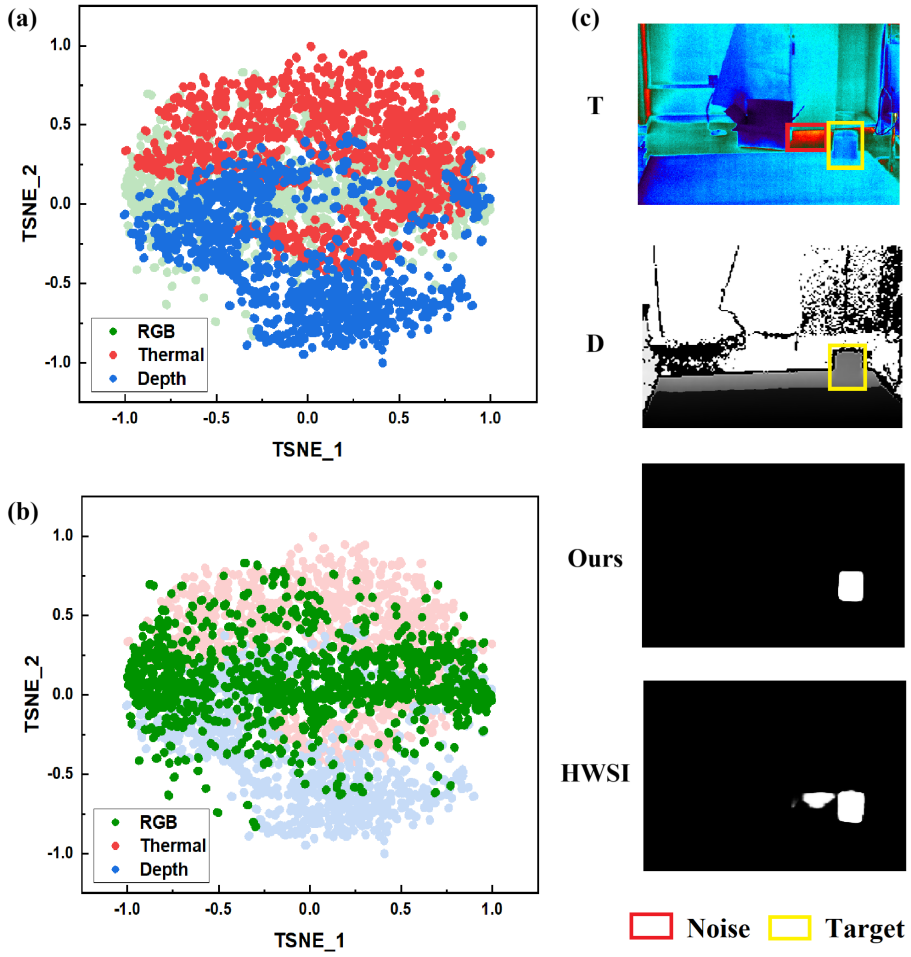

Figure 5.1: (a) The TSNE representation of different modalities with depth and thermal are highlighted; (b) TSNE representation with RGB modality is highlighted (c) The visualised results of existing methods on D-T modality, the RGB-dominated models show less capability in interpreting D-T data.

[114, 62, 102, 116] and RGB-D-T [88, 97] have already achieved commendable results. The addition of depth information (i.e., RGB-D) facilitates a more comprehensive understanding of 3D geometry, proving inherently superior to unimodal RGB, particularly in challenging environments such as low light [32]. Additionally, the inclusion of thermal images (i.e., RGB-T) accentuates specific temperature variances within the sensing domain, reducing distractions and enhancing domain awareness under clustered environments.

In this case, triple-modality data (RGB-depth-thermal) could potentially offer the most comprehensive information. Therefore, the triple-modality benchmark method [88] yields superior results compared to the other two multimodal combinations (RGB-D and RGB-T) [88, 97] at the cost of a substantial model size of 403.4 MB and a high computational complexity of 357.69G FLOPs. Indeed, in contrast to RGB-D-T methods, neither RGB-D nor RGB-T approaches can offer a complete representation.

However, indiscriminately incorporating additional modalities in multimodal tasks can lead to data redundancy, increased computational costs, and potential privacy concerns. In order to identify the redundant elements across these three modalities, we analyse the intrinsic characteristics of different modalities, revealing significant overlap between RGB and depth or thermal, while depth and thermal exhibit less overlap but offer broad domain coverage, as shown in Figure 5.1(a) and (b) indicating that D-T introduces less redundancy. Nevertheless, unlike the texture coherence of RGB-T and the spatial coherence of RGB-D, the lack of coherence between depth and thermal modalities poses a challenge for existing multimodal methods when dealing with depth-thermal data. Visualization results in Figure 5.1(c) show significant discrepancies in the description of the same region in D-T, which can confuse the model and lead to incorrect interpretations. To address the challenge of the information integration of the low coherence modalities, we present the Cross Shallow and Deep Perception Network, CSDNet. This two-stage approach explores synergies between the low-coherence modalities through saliency-aware prescreening at the shallow layer and implicit coherence activation at the deep layer to reasonably select similarities and distinctions among high-level features derived from these disparate inputs. Furthermore, considering that the selected backbone of our encoder, MobileNet-V2, was pre-trained on the expansive RGB dataset ImageNet, which shows less interpretation capability on depth and thermal data, we utilise the powerful and robust SAM [55] to guide the encoder in mapping the D-T into a generalised feature space. To the best of our knowledge, this is not only the first study to investigate the low coherence

modality synergies using depth and thermal data but also the first study of applying D-T modalities to the task of salient object detection. The main contribution of this work can be summarised as follows:

- We propose a novel cross shallow and deep perception scheme to maximise the scene interpretation by leveraging low-coherence modalities.
- We introduce an innovative SAM-assist encoder pre-training (SEP) framework to guide the encoder to extract more generalised features.
- The proposed method is implemented in the salient object detection network using only depth and thermal images. Comprehensive experiments are carried out to demonstrate the effective integration of depth and thermal modalities, which can benefit privacy-preserving visual applications, such as home service/care robots.

5.3 Related Works

5.3.1 Multi-modal Salient Object Detection

Recent advancements in image capture technologies have facilitated the integration of depth and thermal imaging in Salient Object Detection (SOD) tasks. Huang et al. [45] introduced an RGB-D saliency detection model using dual shallow subnetworks to extract unimodal RGB and depth features. Concurrently, Song et al. [89] presented a modality-aware decoder that includes feature embedding and modality reasoning. Bi et al. [10] developed a cross-modal hierarchical interaction network for progressively fusing multi-level features. The positional dependence of depth information complicates object identification, especially for targets near their background, leading to the incorporation of thermal imaging to enhance saliency detection. In [34], Gao et al.

proposed a depth-aware inverted refinement network innovatively structured to account for depth awareness, employing backward propagation to manage multimodal attributes across strata, thus preserving relevant details. Chen et al. [18] proposed a network that reduces modality discrepancies through various integrated modules. Similarly, He et al. [40] proposed a network centered on enhancement and feedback aggregation, with specialised blocks for inter-modal complementation. In existing dual-modality approaches, RGB serves as the primary modality, but neither RGB-D nor RGB-T provides a comprehensive representation compared to the RGB-D-T triple-modality.

Limited studies on RGB-D-T salient object detection exist due to the high costs of collecting and aligning triple-modality data. In [88], Song et al. introduced the VDT-2048 dataset for salient object detection, comprising 2048 image groups from 14 challenging scenes to evaluate multimodal SOD methods. They also presented a benchmark using a hierarchical weighted suppress interference (HWSI) architecture for effective feature fusion. Subsequently, Wan et al. [97] addressed the limitations of single and dual-modal methods by introducing MFFNet, which integrates RGB, depth, and thermal images through a deep fusion encoder and a progressive feature enhancement decoder, improving performance over dual-modality methods. However, they inevitably lead to larger model size and increasing computational burden. In this study, we present the low-coherence D-T modality integration method, aiming to achieve scene interpretation comparable to the triple-modality approach.

5.3.2 Segment Anything Model and Derived Works

The release of the Segment Anything Model (SAM) [55] by Meta has gained widespread attention. SAM is built around a powerful and robust image encoder, leveraging the strengths of the Vision Transformer (ViT) architecture, coupled with a lightweight decoder that generates prompt-guided masks which work in sequence. SAM was

trained on the SA-1B dataset, which comprises an impressive collection of over 1 billion masks on 11 million images. The extensive training offered the model with strong generalisation ability and can be easily adapted to a range of downstream vision tasks, positioning it as a pivotal foundation model in computer vision and it is believed to be a 'GPT moment' for vision. Following the release of SAM, numerous studies have been conducted to make it more mobile-friendly [113, 108, 109], as well as to tailor it for specific data types, such as medical imagery [99, 17, 36] and video stream [80]. In this work, we propose the SEP framework, which employs the SAM to guide the encoder in effectively interpreting depth and thermal data, thereby improving model performance.

5.4 Proposed Method

This section presents the overview of our proposed CSDNet for salient object detection relying only on thermal and depth images. For the integration of the modalities with low coherence, the prescreened spatial information is exchanged between the modalities at the shallow layer. Meanwhile, the middle layers select the relevant representations from the superposition features. And finally, the deep layer facilitates the implicit coherence between two modalities utilizing high-level features. Following this pipeline, the cross shallow and deep integration scheme consists of two modules. For the shallow layer spatial synergies, we introduce the CFAR detector-based saliency prescreening (CSP) module, while the implicit coherence activation navigator (ICAN) module is designed for the deep layer semantic synergies. The motivation and implementation of CSP, ICAN, and SEP will be presented in detail in the following subsections. Finally, the loss used in the proposed method is formulated. Figure 5.2 shows the overview of our proposed network. The overall network structure follows the conventional encoder-decoder framework. The adapted MobileNet-V2 is incorporated as the encoder backbone for both depth and thermal modalities. Representative

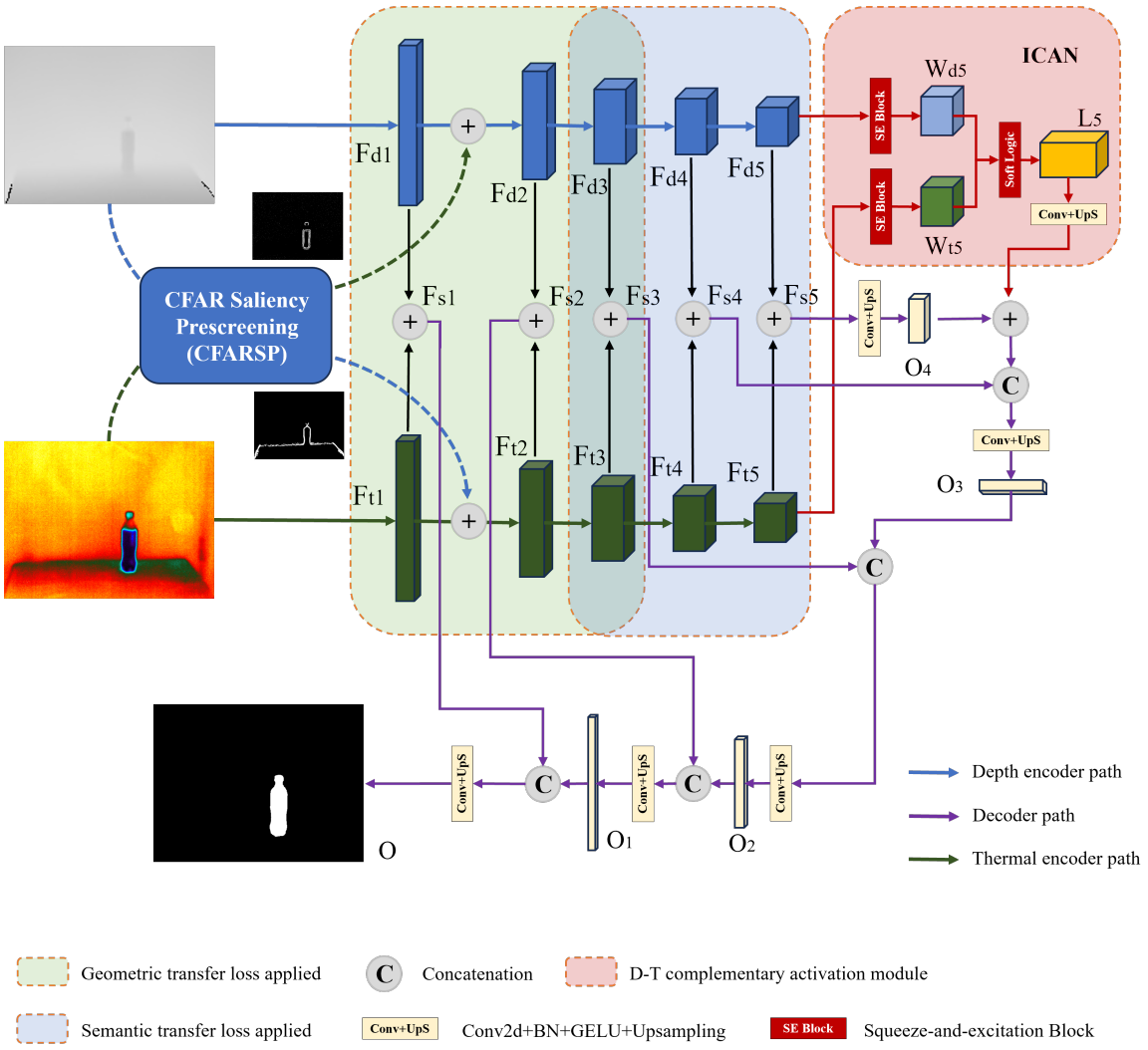

Figure 5.2: The overview of the proposed network CSDNet

features extracted by the encoder at five distinct scales are denoted as F_{di} and F_{ti}, where $i = 1, 2, 3, 4, 5$. The CSP module accepts the original depth and thermal images as inputs, yielding a saliency-aware prescreening mask. The ICAN module utilises F_{d5} and F_{t5} as input features and subsequently integrates the supplementary features at O_4. The ultimate output of the decoder is a refined saliency map, symbolised as O.

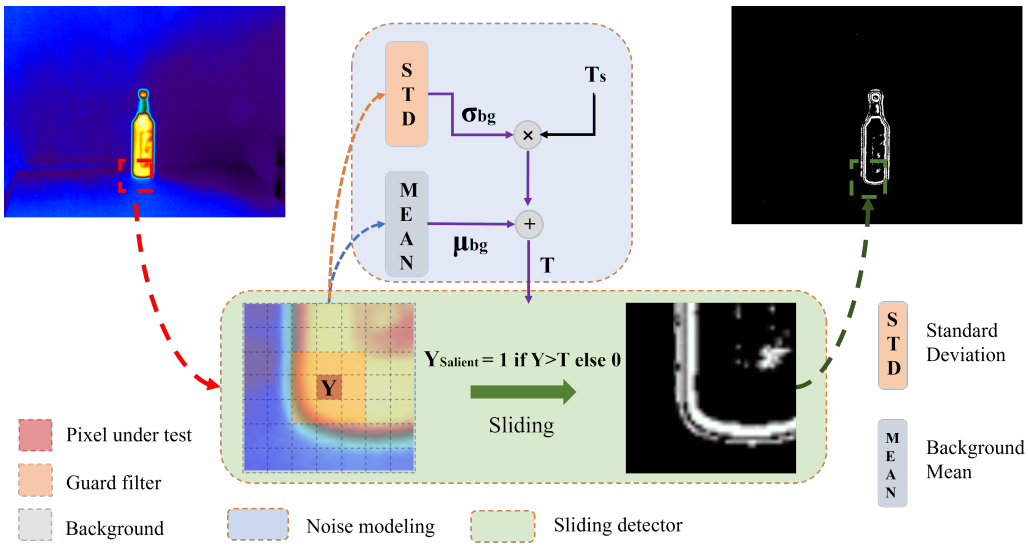

Figure 5.3: The schematic of CFAR Saliency Prescreening Module

5.4.1 CFAR Saliency Prescreening Module

As mentioned in the introduction, hastily fusing two low-coherence modalities in the shallow layers of the network can confuse the model on how to interpret them. To ensure the distinctiveness of each modality, and to exchange information between two modalities in shallow layers, we propose the CSP module. The overall architecture of the CSP module is shown in Figure 5.3. The constant false alarm rate (CFAR) detection is a technique widely used in radar systems to identify and eliminate false alarm signals caused by noise or other interference. Its binary nature implies that it imposes a lower computational burden and can serve as a screening mask to highlight significant features of the modality. We follow the modelling approach of CFAR using probability density functions to describe background clutter. In this work, the 2D probability of false alarm (PFA) for the threshold T can be represented as $PFA = 1 - \int_{-\infty}^{T} f(x)dx = \int_{T}^{\infty} f(x)dx$.

In instances of medium and lower resolution imagery, such clutter frequently adheres to a Gaussian distribution when examined within the intensity domain or, alternatively, conforms to a Rayleigh distribution within the amplitude domain, in accor-

dance with the principles of the central limit theorem [15]. In this case, we employ the commonly used Gaussian distribution model and use a sliding window to make saliency judgments on candidate points in the background. Thus the detector can be described as: $Y > \mu_{bg} + \sigma_{bg} T_s \Leftrightarrow target.$ where Y denotes the pixel under testing, μ_{bg} and σ_{bg} represent the mean value and standard deviation of the background. And T_s is the design parameter threshold scale which controls the sensitivity of CFAR detector.

5.4.2 Implicit Coherence Activation Navigation Module

Our analysis indicates that deep networks can have different semantic descriptions of the same scene in two modalities with lower coherence. In this context, the model can facilitate a more comprehensive scene interpretation by linking the semantic information from depth and thermal modalities wisely. Motivated by this potential for enhanced scene understanding, we introduce a new implicit coherence activation navigator aimed at activating hidden coherence relationships by emphasizing the consistency and difference between two semantics. The implementation logic is demonstrated in the upper-right corner in Figure 5.2(a). The highest level features F_{d5} and F_{t5} are weighted by the squeeze-and-excitation block [44], denoted as W_{d5} and W_{t5}. The soft logic operations can be represented as $AND_{w5} = min(W_{d5}, W_{t5})$, $OR_{w5} = max(W_{d5}, W_{t5})$, $XOR_{w5} = abs(W_{d5} - W_{t5})$, and then the results are concatenated along the channel axis $L_5 = concat(AND_{w5}, OR_{w5}, XOR_{w5})$. The concatenated logic result is added to O_4 with a higher resolution feature after sequential operations of convolution, batch normalisation, Gaussian Error Linear Unit (GELU) activation and a bilinear interpolation upsampling.

5.4.3 SAM-Assist Encoder Pre-training Framework

It is noticed that the pre-trained MobileNet-V2 on the ImageNet dataset has limited capability in extracting spatial information from depth images. Informed by the recent progress in vision foundation models, notably the Segment Anything Model (SAM) [55], which demonstrates exceptional aptitude in interpreting various types of images, we have incorporated a SAM-assisted depth encoder pre-training stage. This framework aims to augment the feature extraction capabilities specific to the depth modality, thereby enhancing the overall performance of the CSDNet. The schematic of the proposed SEP framework is depicted in Figure 5.4. The robust and powerful vision transformer (ViT)-based encoder of SAM yields an image embedding S_d dimensioned at [256, 64, 64] for a single input instance. To capitalise on the potential of these embeddings, they are strategically deployed to guide the depth encoder at the fourth phase feature output using SAM-assist loss (SAL), simultaneously ensuring the full depth encoder is weakly aligned with the thermal encoder using geometric transfer loss (GTL) and semantic transfer loss (STL) proposed in [116].

5.4.4 Loss Formulation

In the SAM-assist depth encoder pre-training stage, SAL integrate the Mean Square Error (MSE) loss with the STL. The feature from the depth encoder is denoted as F_{di}, and the image embedding from SAM is denoted as S_d. For instance, considering the STL loss transfers the feature from the $F_{di}, i = 4$ to S_d, the channel attention for attentive transfer involves a sequence of global average pooling (AP), two convolutions ($Conv$) and a sigmoid activation (Sig), followed by a global normalisation (GN) along the channel dimension after AP. The normalized feature w_{Fdi} from F_{di} can be represented as:

$$w_{Fdi} = GN(Sig(Conv(ReLU(Conv(AP(F_{di}^{detach})))))) \tag{5.1}$$

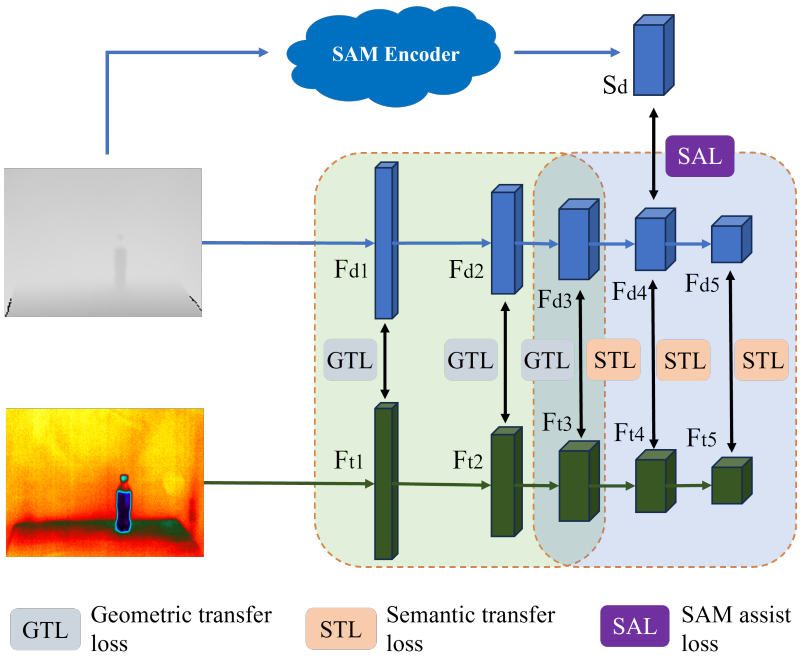

Figure 5.4: The schematic of SAM-assist depth encoder pre-training framework

The L2 norm is applied on the plane dimension to preserve the representations across different channels.

$$S_{d\_norm} = L2norm(S_d), \;\; F_{di\_norm}^{detach} = L2norm(F_{di}^{detach}) \tag{5.2}$$

Hence, the STL from the depth encoder feature to SAM embedding is represented as:

$$STL_{Fdi \to Sd} = w_{Fdi} \times MSE(F_{di\_norm}^{detach}, S_{d\_norm}) \tag{5.3}$$

The SAL is formulated as a weighted sum of MSE loss and STL, where w_1 and w_2 signify the respective weights:

$$SAL = w_1 \cdot MSE(F_{di}, S_d) + w_2 \cdot STL_{Fdi \to Sd} \tag{5.4}$$

Meanwhile, the GTL and STL are employed to anchor the depth encoder using the ImageNet pre-trained thermal encoder. Following the approach in [116]:

$$STL_{Fdi \to Fti} = \sum_{i=3}^{6} w_{Fdi} \times MSE(F_{di\_norm}^{detach}, F_{ti\_norm})$$
$$STL_{Fti \to Fdi} = \sum_{i=3}^{6} w_{Fti} \times MSE(F_{ti\_norm}^{detach}, F_{di\_norm})$$

Unlike STL, GTL calculates spatial attention rather than channel attention to obtain the global geometric weight and distinguish the significant features on the plane. The *L2norm* in GTL is computed along the channel axis to preserve geometric information of the spatial structure.

$$GTL_{Fdi \to Fti} = \sum_{i=1}^{3} w_{Fdi} \times MSE(F_{di\_norm}^{detach}, F_{ti\_norm})$$
$$GTL_{Fti \to Fdi} = \sum_{i=1}^{3} w_{Fti} \times MSE(F_{ti\_norm}^{detach}, F_{di\_norm})$$

The overall loss for the pre-training stage can be represented as follows, where w_3 and w_4 denote different weights.

$$\begin{aligned} L_{SEP} = SAL &+ w_3(GTL_{Fdi \to Fti} + STL_{Fdi \to Fti}) \\ &+ w_4(GTL_{Fti \to Fdi} + STL_{Fti \to Fdi}) \end{aligned} \tag{5.5}$$

After the depth encoder pre-training stage, the SOD loss functions are employed for joint encoder-decoder training. The predicted saliency region should not only be accurately aligned with the ground truth but also demonstrate a good fit on the boundary of the saliency region. Therefore, the saliency region boundaries are extracted from both GT and O_i from the decoder. The intersection-over-union (IOU)

and binary cross-entropy (BCE) are utilised to measure the accuracy and precision:

$$L_{SOD}^{O_i} = L_{ioubce}^{reg_{Oi}} + L_{ioubce}^{bou_{Oi}} \tag{5.6}$$

Where $L_{ioubce} = L_{iou} + L_{bce}$. To achieve a better performance, the L_{SOD} is calculated for the last three stages in the decoder, i.e. for O, O_1 and O_2, with the final loss being the summation of the three terms:

$$L_{SOD} = L_{SOD}^{O} + L_{SOD}^{O_1} + L_{SOD}^{O_2} \tag{5.7}$$

5.5 Experiments

5.5.1 Dataset and Evaluation Metrics

We validate the proposed model CSDNet on the VDT-2048 dataset as reported by Song et al. [88], which contains 1048 images for training and 1000 images for testing. In the context of SOD evaluation, we incorporate the following five benchmark metrics: mean absolute error (MAE) [76], F-measure (F_m) [1], weighted F-measure (W_F) [66], structure measure (S_m) [27] and E-measure (E_m) [28]:

(1) Mean absolute error (MAE) quantifies the difference between the predicted saliency map O and ground truth GT on a pixel-by-pixel basis, i.e.: $MAE = \sum_{i=1}^{w}\sum_{j=1}^{h}|O(i,j) - GT(i,j)|/(w \times h)$. Where w and h denote the height and width of the images, respectively. The MAE score ranges within [0,1], with 0 indicating a perfect overlap between the saliency map and the ground truth and 1 indicating complete disparity. However, the limitation of MAE lies in its inability to precisely depict the divergence between the saliency map and the ground truth for smaller objects, given that it computes an average value across the entire image, potentially resulting in a lower MAE for smaller objects.

(2) F-Measure (F_m) calculates a weighted harmonic mean of precision and recall. An adaptive thresholding approach is employed to establish the comparison of the binary saliency map with ground truth, wherein the threshold is set dynamically based on the saliency map. $F_\beta = (1+\beta^2) \cdot precision \cdot recall / (\beta^2 \cdot precision + recall)$. Specifically, the threshold is defined as twice the average value of the saliency map.

(3) Weighted F-Measure (W_F) is the F-measure with weighted precision (a measure of exactness) and weighted recall (a measure of completeness): $F_\beta^w = (1+\beta^2) \cdot precision^w \cdot recall^w / (\beta^2 \cdot precision^w + recall^w)$

(4) Structure Measure (S_m) assesses the structural integrity of the detected salient regions by incorporating two distinct components, region-aware (S_r) and object-aware (S_o). Unlike the MAE and F-measure, which compare two images on a pixel-by-pixel basis, the S-measure emphasises the structural similarity between the saliency map and ground truth: $S_\alpha = \alpha \cdot S_o + (1-\alpha) \cdot S_r$

(5) E-Measure (E_m) is the enhanced-alignment measure that combines local pixel values with the image-level mean value, jointly capturing image-level statistics and local pixel-matching information: $E_\xi = \sum_{i=1}^{w} \sum_{j=1}^{h} \varphi(i,j)/(w \times h)$. φ represents the enhanced alignment matrix, which denotes the correlational relationship between the predicted saliency maps and the corresponding ground truth.

5.5.2 Implementation Details

The proposed model CSDNet is built and trained using the Pytorch framework. Optimisation during training is achieved through the application of the Adam optimisation algorithm. The MobileNet-V2 architecture, serving as the backbone of the network, is initialised with pre-trained weights obtained from the ImageNet dataset during the SEP phase, whereas initialisation for the remaining network components is conducted randomly. All experimental training and testing procedures are performed on a machine equipped with an Intel 8C16T Core i7-11700KF at 3.6 GHz × 16 and an

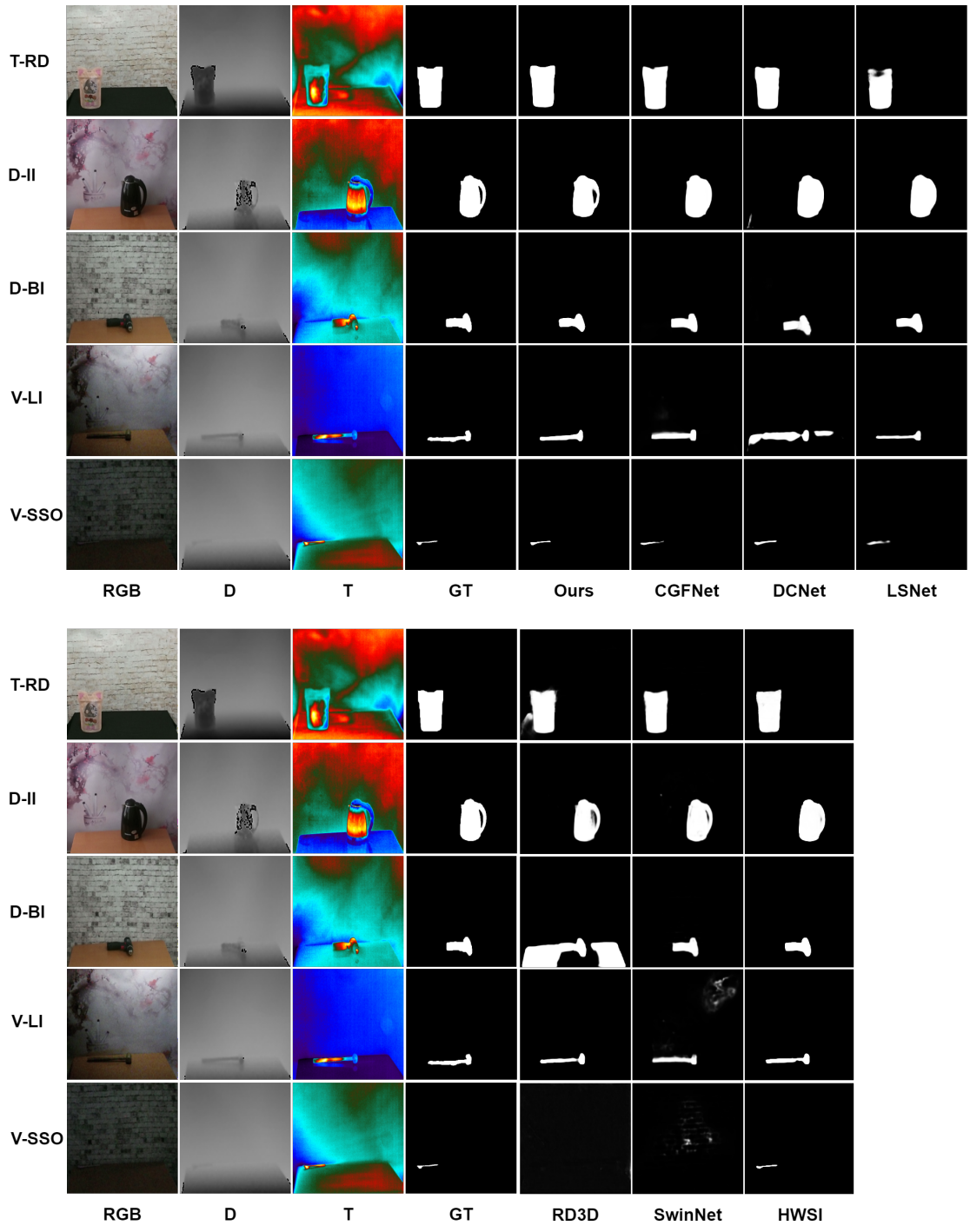

Figure 5.5: Visual Comparison on VDT-2048 dataset

NVIDIA GeForce RTX 3060 graphics card.

Table 5.1: Quantitative Comparison Results of Different Methods on VDT-2048 Dataset. ↑/↓ indicates that a larger/smaller value is better.

Model	Type	MAE↓	F_m ↑	W_F ↑	S_m ↑	E_m ↑
CGFNet [98]	RGB-T	0.0034	0.7777	0.8468	0.9166	0.9299
CSRNet [47]	RGB-T	0.0050	0.7828	0.8159	0.8827	0.9460
DCNet [95]	RGB-T	0.0038	0.8457	0.8284	0.8803	0.9699
LSNet [116]	RGB-T	0.0045	0.7434	0.8046	0.8878	0.9201
MoADNet [51]	RGB-D	0.0126	0.5753	0.5796	0.7697	0.8376
RD3D [19]	RGB-D	0.0047	0.6444	0.7948	0.9090	0.8345
SwinNet [63]	RGB-D	0.0038	0.7287	0.8385	0.9194	0.8962
RFNet [100]	RGB-D	0.0031	0.8252	0.8680	0.9175	0.9635
Ours	D-T	0.0029	**0.8904**	0.8833	0.8794	**0.9806**
HWSI [88]	RGB-D-T	**0.0027**	0.8581	**0.8983**	**0.9324**	0.9765

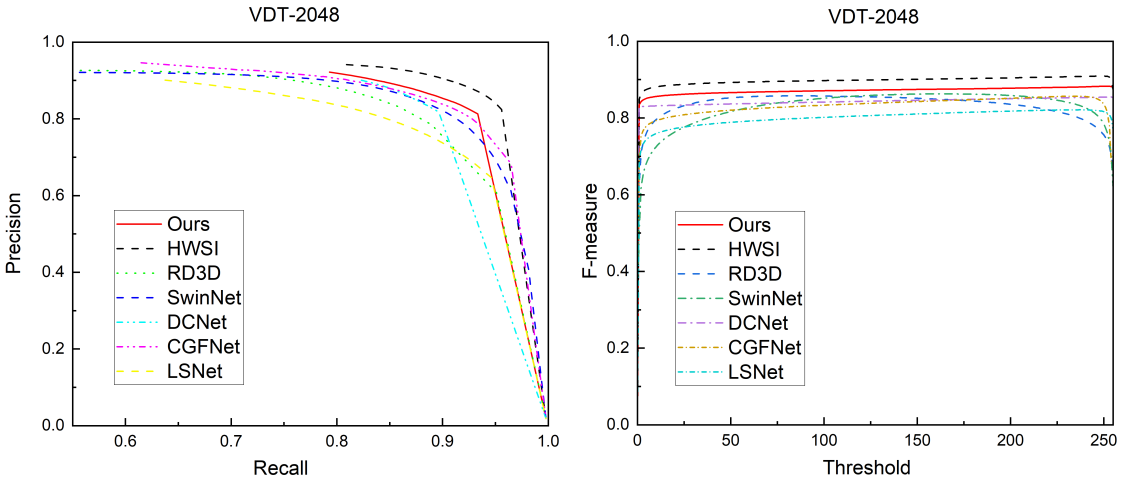

Figure 5.6: Precision-Recall Curve and F_m-Threshold Curve Comparison with Different Methods

5.5.3 Experimental Results

We conduct the quantitative comparison of our proposed method with eight SOTA methods, including four RGB-T methods CGFNet [98], CSRNet [47], DCNet [95], LSNet [116], and four RGB-D methods MoADNet [51], RD3D [19], SwinNet [63] and RFNet [100]. In addition, we include the RGB-D-T benchmark method HWSI [88] for a comprehensive evaluation. To ensure a fair comparison, we use the official

Table 5.2: Quantitative Results in V Challenges. LI, NI, SI, and SSO denote Low Illumination, No Illumination, Side Illumination and Small Salient Object, respectively

Model	V-LI			V-NI			V-SI			V-SSO		
	MAE↓	W_F ↑	E_m ↑	MAE↓	W_F ↑	E_m ↑	MAE↓	W_F ↑	E_m ↑	MAE↓	W_F ↑	E_m ↑
CGFNet	.0046	.8287	.9334	.0035	.7614	.8670	.0043	.8404	.9324	.0012	.7196	.7884
CSRNet	.0058	.8117	.9533	.0043	.7647	.9185	.0078	.7905	.9380	.0014	.7067	.8188
DCNet	.0048	.8198	.9474	.0038	.7291	.9339	.0047	.8411	**.9833**	.0012	.6769	**.9324**
MoADNet	.0153	.5543	.8555	.0143	.3209	.7917	.0142	.5776	.8580	.0049	.4077	.6391
PANet	.1000	.1360	.7603	.1207	.0927	.6804	.1001	.1271	.7526	.1321	.0152	.3773
LSNet	.0057	.7895	.9297	.0053	.6734	.8634	.0064	.7721	.9301	.0018	.6491	.7245
RD3D	.0062	.7583	.8258	.0066	.6110	.6964	.0070	.7521	.8224	.0025	.6646	.5698
SwinNet	.0050	.8059	.8864	.0055	.6718	.7717	.0055	.8075	.8862	.0016	.7039	.6073
RFNet	.0043	.8450	.9633	.0043	.7292	.9161	.0053	.8179	.9513	.0011	.7858	.8816
Ours	**.0031**	**.8927**	.9674	**.0024**	**.8610**	**.9541**	**.0033**	**.8934**	.9740	.0012	.7649	.9254
HWSI	.0038	.8683	**.9695**	.0028	.8453	.9522	.0038	.8757	.9734	**.0008**	**.8402**	.9134

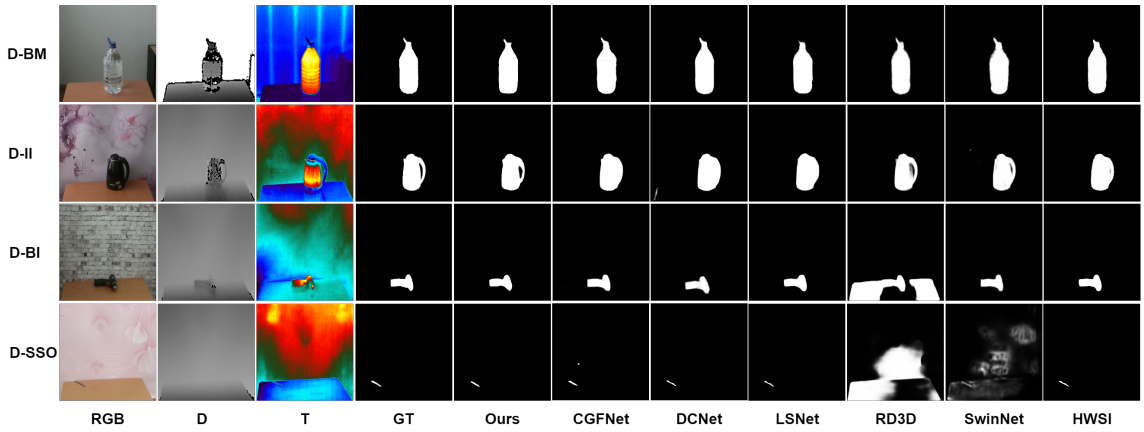

Figure 5.7: Visual Comparison in D Challenges

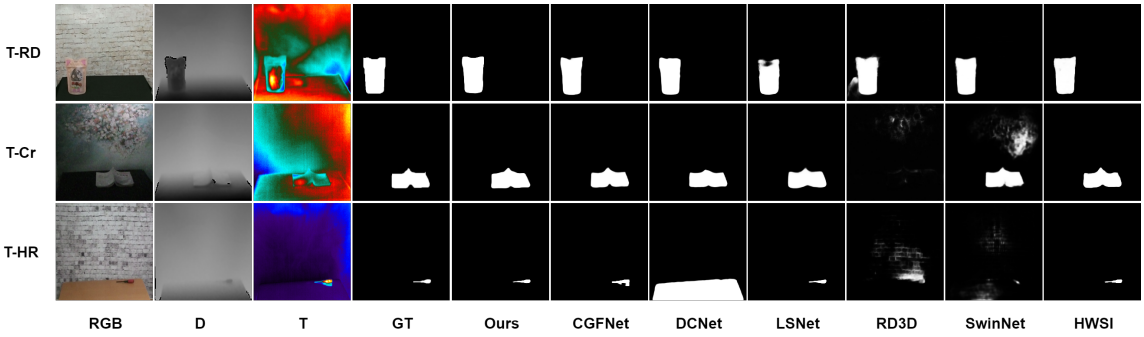

Figure 5.8: Visual Comparison in T Challenges

Table 5.3: Quantitative Results in V Challenges Cont. BSO, MSO, and SA denote Big Salient Object, Multiple Salient Object, and Similar Appearance, respectively

Model	V-BSO			V-MSO			V-SA		
	MAE↓	W_F ↑	E_m ↑	MAE↓	W_F ↑	E_m ↑	MAE↓	W_F ↑	E_m ↑
CGFNet [98]	.0071	.9433	.9917	.0049	.8542	.9456	.0030	.8421	.9262
CSRNet [47]	.0139	.8755	.9618	.0089	.8000	.9389	.0047	.7598	.9484
DCNet [95]	.0082	.9382	.9908	.0055	.8401	.9801	.0032	.8338	**.9807**
MoADNet [51]	.0439	.5349	.8073	.0137	.6750	.8684	.0067	.6975	.8929
PANet [48]	.0966	.3852	.9881	.1075	.1505	.8349	.1007	.0885	.7448
LSNet [116]	.0101	.9123	.9889	.0061	.8180	.9364	.0041	.8052	.9289
RD3D [19]	.0091	.9240	.9858	.0058	.8188	.8871	.0038	.8140	.8705
SwinNet [63]	.0072	.9434	.9918	.0050	.8414	.9195	.0030	.8524	.9142
RFNet [100]	.0067	.9484	.9920	.0040	.8708	.9661	.0024	**.8892**	.9704
Ours	**.0060**	**.9562**	.9878	**.0040**	.8885	.9712	.0028	.8644	.9716
HWSI [88]	.0061	.9517	**.9927**	.0041	**.8915**	**.9716**	**.0024**	.8803	.9748

Table 5.4: Quantitative Comparison with HSWI on Different Modality Combinations

Model	Modality Setting	MAE↓	F_m ↑	W_F ↑	S_m ↑	E_m ↑
HWSI	RGB-D	0.0048	0.8454	0.8179	0.8398	0.9548
	RGB-T	0.0034	0.8888	0.8708	0.8709	0.9653
	D-T	0.0050	0.8343	0.8088	0.8516	0.9580
Ours	D-T	0.0029	0.8904	0.8833	0.8793	0.9806

code released by the authors. The quantitative results are presented in Table 5.1, where the red colour highlights the best results among all double-modality methods, and bold text highlights the top-performing results across all types of methods. The proposed method demonstrates significant advantages over all other dual-modality approaches, regardless of whether they are RGB-D or RGB-T, and it achieves comparable results to the triple-modality benchmark method HWSI on the VDT-2048 dataset. Specifically, our method outperforms the other dual-modality approaches up to 0.97%, 31.51%, 30.37%, 10.97% and 14.61% in terms of MAE, F_m, W_F, S_m and E_m, respectively. Compared to the triple-modality method, the proposed method exhibits a mere 0.02% disparity in MAE and gains 3.23%, 0.41% advantages in F_m and E_m. For the remaining two metrics, W_F and S_m, our method also achieves comparable results. The visual comparison is shown in Figure 5.5. Furthermore, we

analyse the triple-modality benchmark method HWSI using different bimodal inputs, demonstrating the comprehensive advantages of our method over the HSWI across various dual-modality combinations. The proposed method exhibits advantages up to 0.21%, 5.61%, 7.45%, 3.95% and 2.58% in terms of the five SOD metrics, and the numerical comparison is detailed in Table 5.4. In Figure 5.6, our method is positioned in the upper-right corner among the PR curves and towards the top of the F-measure against the threshold diagram, demonstrating its superior performance.

Moreover, to demonstrate the effectiveness and robustness of the proposed method, we also present the numerical results comparison with other methods on the challenges proposed in the VDT-2048 dataset. We assess the V-challenges encompassing low illumination (LI), no illumination (NI), side illumination (SI), small salient objects (SSO) in Table 5.2 and big salient object (BSO), multiple salient object (MSO) and similar appearance (SA) in Table 5.3, D-challenges including background interference (BI), background messy (BM), information incomplete (II) and small salient object (SSO) in Table 5.5, and T-challenges including thermal crossover (Cr), heat reflection (HR), and radiation dispersion (RD) in Table 5.6. Since the proposed method is designed for indoor privacy-preserving applications and mobile platforms in search and rescue operations, it does not rely on the RGB visible light data. Thus, our approach exhibits significant advantages in all challenging illumination scenarios. However, in the SSO challenge, our method trails the HWSI by 0.04%, 7.53%, 1.2% in terms of MAE, W_F, E_m respectively, since the triple-modality does have a much richer texture that can benefit the segmentation in small targets. Despite this, our method, which emphasises the synergy between depth and thermal data by introducing CSP and ICAN, outperforms most dual-modality methods that primarily use visible light data and achieve comparable results to the trimodal HWSI. In the D-Challenges, our method consistently outperforms all other dual-modality approaches across all metrics. While the triple-modality HWSI method achieves slightly lower MAE values in some cases (e.g., D-BI and D-BM), our method demonstrates comparable or better

Table 5.5: Quantitative Results in D-Challenges. BI, BM, II and SSO Denote Background Interference, Background Messy, Information Incomplete and Small Salient Object Respectively

Model	D-BI			D-BM		
	MAE↓	F_m ↑	E_m ↑	MAE↓	F_m ↑	E_m ↑
MoADNet [51]	.0107	.5750	.8353	.0147	.4674	.8017
PANet [48]	.1084	.4528	.6916	.1062	.5024	.7368
RD3D [19]	.0046	.6161	.8167	.0044	.6383	.8289
SwinNet [63]	.0036	.6858	.8650	.0044	.6958	.8717
RFNet [100]	.0029	.8124	.9596	.0032	.8130	.9614
Ours	.0027	**.8808**	**.9789**	.0031	**.8909**	**.9807**
HWSI [88]	**.0024**	.8501	.9748	**.0029**	.8460	.9727

Model	D-II			D-SSO		
	MAE↓	F_m ↑	E_m ↑	MAE↓	F_m ↑	E_m ↑
MoADNet [51]	.0183	.5788	.8461	.0049	.3251	.6391
PANet [48]	.1013	.6426	.8271	.1321	.1119	.3773
RD3D [19]	.0052	.7316	.8901	.0025	.3303	.5698
SwinNet [63]	.0047	.7845	.9205	.0016	.3720	.6073
RFNet [100]	.0037	.8652	.9752	.0011	.6544	.8816
Ours	**.0035**	**.9200**	**.9859**	.0011	**.7700**	**.9254**
HWSI [88]	**.0035**	.8829	.9813	**.0008**	.7130	.9134

performance in terms of F_m and E_m. The visual comparison is visualized in Figure 5.7. In the T-Challenges, our method also shows better performance compared to other dual-modality approaches in terms of MAE and F_m. For T-HR and T-RD, our method achieves the best MAE and F_m among dual-modality methods. However, the heat reflection and radiation dispersion destroy more alignment information in thermal data. Although HWSI achieves marginally better results in some metrics (e.g., MAE for T-Cr and E_m for T-RD), our method consistently performs competitively and outperforms HWSI in other key metrics. Visualized results are shown in Figure 5.8.

Table 5.7 list the quantitative comparison of the proposed method against the other SOTA methods in terms of running time, number of parameters and FLOPs. The proposed model achieves comparable performance with HWSI while utilising only 0.06

Table 5.6: Quantitative Results in T-Challenges. Cr, HR, RD Represent Crossover, Heat Reflection and Radiation Dispersion Respectively

Model	T-Cr			T-HR			T-RD		
	MAE↓	F_m ↑	E_m ↑	MAE↓	F_m ↑	E_m ↑	MAE↓	F_m ↑	E_m ↑
CCGFNet [98]	.0034	.7350	.9075	.0029	.8329	.9668	.0046	.8522	.9750
CSRNet [47]	.0050	.7304	.9234	.0034	.8495	.9792	.0061	.8453	.9769
LSNet [116]	.0042	.7119	.8986	.0043	.7826	.9489	.0065	.8083	.9647
DCNet [95]	.0039	.8116	.9548	.0032	.8928	**.9893**	.0051	.8845	.9845
Ours	.0032	**.8622**	.9558	**.0022**	**.9459**	.9780	**.0037**	**.9202**	.9732
HWSI [88]	**.0025**	.8323	**.9670**	.0027	.8818	.9878	.0041	.8854	**.9888**

Table 5.7: Comparison in terms of running time, model parameters and FLOPs

Model	CGFNet	DCNet	LSNet	RD3D	SwinNet	HWSI	Ours
Runtime (FPS)	3.59	6.15	12.83	9.03	6.10	1.72	10.27
Params (M)	66.38	24.06	4.56	46.90	198.78	100.77	5.96
FLOPs (G)	345.54	207.21	1.23	50.86	124.72	357.93	1.30

times the number of parameters and requiring 0.0036 times fewer FLOPs. In terms of processing speed, our model is approximately 5.97 times faster than HWSI. These results demonstrate that our model is more suited for deployment on edge devices or mobile platforms.

5.5.4 Ablation Analysis

To demonstrate the effectiveness of each module in our CSDNet (i.e., CSP, ICAN and SEP), we conduct the ablation experiments incrementally incorporating the cross shallow and deep scheme (i.e. CSP+ICAN) as well as the SEP framework, the numerical results are presented in Table 5.8. The results demonstrate that each module contributes to improving the model performance. Where CSP and ICAN improve the MAE, F_m, W_F, and S_m by 0.059%, 1.872%, 2.759% and 0.982% by exchanging and enhancing the scene interpretation. SEP improves the MAE, F_m, W_F, S_m and E_m by 0.02%, 0.386%, 0.612%, 0.112% and 0.14%. When all the modules are enabled,

the model achieves the best performance, which indicates that there is a synergistic effect among the modules and the proposed method achieves the effective integration of depth and thermal modalities based on the intrinsic characteristics of the modality. Table 5.9 demonstrates that our model performs best on the D-T modality, as the proposed method is designed to maximize the utilization of information differences between the low coherence modality. Figure 5.9 visualizes that our approach introduces more detailed edge and shape information in the extracted features effectively.

Table 5.8: Ablation Study on the Effectiveness of Modules

Settings			MAE↓	$F_m \uparrow$	$W_F \uparrow$	$S_m \uparrow$	$E_m \uparrow$
CSP	ICAN	SEP					
×	×	×	0.00366	0.87046	0.85372	0.86939	0.98170
✓	✓	×	0.00307	0.88918	0.88131	0.87921	0.97833
×	×	✓	0.00346	0.87432	0.85984	0.87051	0.98310
✓	✓	✓	0.00291	0.89039	0.88332	0.87934	0.98061

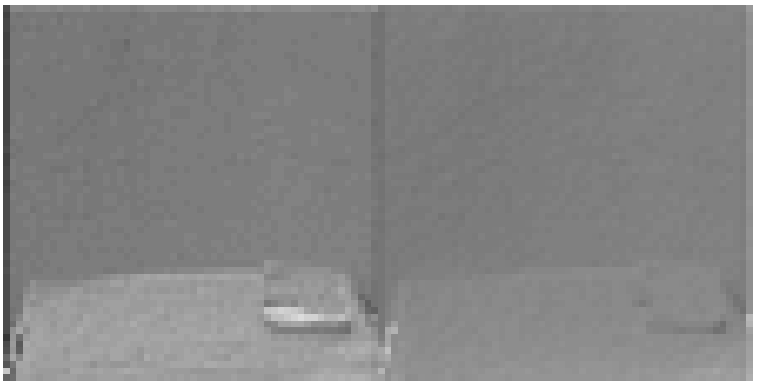

Figure 5.9: Feature difference (left) incorporating the cross shallow and deep scheme; (right) without the cross shallow and deep scheme.

Table 5.9: Different Modality Setting on Our Method

Settings	MAE↓	$F_m \uparrow$	$W_F \uparrow$	$S_m \uparrow$	$E_m \uparrow$
RGB-D	0.00403	0.86006	0.83918	0.85730	0.97664
RGB-T	0.00338	0.88890	0.87153	0.87590	0.97899
D-T	0.00291	0.89039	0.88332	0.87934	0.98061

5.6 Conclusion

In this study, we introduce CSDNet, a novel lightweight cross shallow and deep perception network designed to effectively integrate low-coherence modalities. The proposed method is implemented and assessed on salient object detection task with depth and thermal imagery. CSDNet outperforms current state-of-the-art RGB-D and RGB-T methods and achieves comparable results to RGB-D-T on the VDT-2048 dataset, running 5.96 times faster and demanding 0.0036 times fewer FLOPs. This makes it suitable for edge device applications with privacy concerns, such as home care or mobile platforms in challenging lighting conditions, like search and rescue robots. Extensive experiments under various conditions—including difficult illuminations, small objects, background interferences, and various thermal interferences—demonstrate the robustness of our approach. CSDNet effectively integrates low-coherence depth-thermal modalities and shows great potential for generalization to other low-coherence modalities.

Chapter 6

Conclusion and Future Work

6.1 Summary of Contributions

This thesis has presented three novel approaches to address three challenges in intelligent perception for mobile robotic systems. By focusing on semantic awareness in view planning, hybrid quantum-classical optimisation for viewpoint selection, and low-coherence multi-modal integration, this research has significantly advanced the state-of-the-art in autonomous perception under complex environmental conditions.

The first contribution introduced a semantic-aware Next-Best-View (S-NBV) framework that fundamentally reimagines robotic exploration by incorporating semantic information alongside traditional visibility metrics. By formulating a unified information gain function that balances visibility and semantic gain, the approach enables mobile robots to perform purposeful exploration with targeted search-and-acquisition manoeuvres. Experimental evaluations demonstrated substantial improvements over conventional approaches, achieving up to 27.46% improvement in region-of-interest reconstruction and dramatically improved perspective directivity when exploring environments with objects of interest. This semantic-guided exploration paradigm represents an important step toward more intelligent, goal-directed robotic perception

systems.

The second contribution addressed the inherent limitations of classical optimisation methods in viewpoint selection through the development of a Hybrid Quantum-Classical Next-Best-View (HQC-NBV) framework. By leveraging the unique properties of quantum computation—specifically superposition and entanglement—this approach enables more effective exploration of the complex, high-dimensional solution space characteristic of view planning problems. The novel formulation includes a multi-component Hamiltonian and parameter-centric variational ansatz with bidirectional alternating entanglement patterns that capture hierarchical dependencies between viewpoint parameters. Experimental results demonstrated that the quantum-specific components provide measurable advantages, with the approach achieving up to 49.2% higher exploration efficiency compared to classical methods. This work establishes a pioneering connection between quantum computing and robotic perception, offering a new computational paradigm for solving complex optimisation problems in robotics.

The third contribution focused on robust perception in challenging lighting conditions through the development of the Cross Shallow and Deep Perception Network (CSDNet). This lightweight architecture efficiently integrates depth and thermal modalities—two sensing approaches with low coherence but complementary information content. By implementing spatial information prescreening and implicit coherence navigation across network layers, CSDNet achieves state-of-the-art performance while reducing computational requirements by orders of magnitude compared to triple-modality methods. The approach was further enhanced through a Segment Anything Model (SAM)-assisted encoder pre-training framework that effectively guides feature mapping to a generalised feature space. These innovations enable robust perception in extreme lighting conditions while offering inherent privacy advantages due to the non-RGB nature of the sensing modalities.

Collectively, these contributions represent significant advancements in intelligent per-

ception for mobile robotic systems. While developed as distinct solutions addressing different aspects of the perception challenge, they share a common goal of enabling more effective, efficient, and robust environmental understanding for autonomous systems operating in complex real-world environments.

6.2 Limitations and Future Work

Despite the significant contributions presented in this thesis, several limitations and promising avenues for future research remain.

6.2.1 Semantic-Aware Next-Best-View Planning

The S-NBV framework, while demonstrating substantial improvements in targeted exploration, could benefit from several extensions:

1. **Multi-Agent Collaborative Exploration**: Extending the semantic-aware planning to multi-robot systems could significantly enhance exploration efficiency. This would require addressing additional challenges in distributed semantic knowledge representation, consensus-building, and coordinated exploration strategies.

2. **Dynamic Environment Adaptation**: The current framework assumes a static environment during the manoeuvre. Extending the approach to dynamic environments would require incorporating motion removal and developing strategies for rapid re-planning as the environment changes.

6.2.2 Hybrid Quantum-Classical View Planning

The HQC-NBV framework, while pioneering the application of quantum computing to robotic perception, has several limitations that future work could address:

1. **Long-Horizon Planning**: The current approach employs a selection strategy considering only the next single step. Investigating longer-horizon planning that considers sequences of viewpoints could further improve exploration efficiency, potentially through incorporating model predictive control frameworks with the optimisation strategy.

2. **Scalability to Larger State Spaces**: The current implementation is limited by the number of available qubits. Future work could explore techniques such as problem decomposition, quantum-inspired classical algorithms, or hybrid approaches that selectively apply quantum computation to the most challenging subproblems.

3. **Hardware Implementation and Noise Resilience**: Current experiments were primarily conducted on quantum simulators. Deploying the approach on actual quantum hardware would require addressing challenges related to quantum noise, decoherence, and limited qubit connectivity.

6.2.3 Cross Shallow and Deep Perception Network

The CSDNet, while effective for depth-thermal integration, could be enhanced through several future directions:

1. **Temporal Information Integration**: The current architecture processes individual frames independently. Incorporating temporal information through recurrent or transformer-based architectures could enhance performance in dynamic scenes.

2. **Adaptive Modal Weighting**: Different environmental conditions might warrant different reliance on each modality. Developing mechanisms to dynamically adjust the influence of each modality based on environmental conditions could further improve robustness.

3. **Extension to Additional Modalities**: While the current work focuses on depth and thermal integration, the approach could be extended to incorporate additional privacy-preserving modalities such as radar, event cameras, or audio sensors.

6.2.4 Integration and Broader Implications

Looking beyond individual contributions, several promising directions for future research emerge at their intersection:

1. **Sequential Information-Aware View Planning**: Developing view planning systems that leverage temporal coherence and sequential observations to improve exploration efficiency. This could involve incorporating spatio-temporal prediction models, utilising motion patterns for dynamic objects, and exploiting the inherent correlation between consecutive viewpoints to reduce uncertainty accumulation and enhance mapping accuracy.

2. **Intelligent Quantum-Classical Resource Allocation**: Developing adaptive scheduling frameworks that dynamically allocate computational tasks between classical and quantum processors based on problem characteristics, resource availability, and task urgency. This would involve creating quantum-classical workload classifiers, real-time resource monitoring, and optimisation strategies that maximise overall system performance while managing quantum coherence time constraints and classical processing capabilities.

In conclusion, this thesis has made several contributions to intelligent perception for mobile robotic systems, advancing the state-of-the-art in semantic-aware planning, quantum-enabled view planning, and low-coherence multi-modal perception. While each contribution addresses specific challenges, together they pave the way for more capable, efficient, and robust autonomous systems that can effectively perceive and interact with complex real-world environments. The limitations identified and future research directions proposed offer promising pathways to further enhance these capabilities, ultimately bringing us closer to the vision of truly intelligent and adaptable robotic perception systems.

References

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *2009 IEEE conference on computer vision and pattern recognition*, pages 1597–1604. IEEE, 2009.

[2] Federica Arrigoni, Willi Menapace, Marcel Seelbach Benkner, Elisa Ricci, and Vladislav Golyanik. Quantum motion segmentation. In *European Conference on Computer Vision*, pages 506–523. Springer, 2022.

[3] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[4] Joseph E Banta, LR Wong, Christophe Dumont, and Mongi A Abidi. A next-best-view system for autonomous 3-d object reconstruction. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 30(5):589–598, 2000.

[5] Ana Batinovic, Antun Ivanovic, Tamara Petrovic, and Stjepan Bogdan. A shadowcasting-based next-best-view planner for autonomous 3d exploration. *IEEE Robotics and Automation Letters*, 7(2):2969–2976, 2022.

[6] Ana Batinovic, Tamara Petrovic, Antun Ivanovic, Frano Petric, and Stjepan Bogdan. A multi-resolution frontier-based planner for autonomous 3d exploration. *IEEE Robotics and Automation Letters*, 6(3):4528–4535, 2021.

[7] Marcel Seelbach Benkner, Vladislav Golyanik, Christian Theobalt, and Michael Moeller. Adiabatic quantum graph matching with permutation matrix constraints. In *2020 International conference on 3D vision (3DV)*, pages 583–592. IEEE, 2020.

[8] Marcel Seelbach Benkner, Zorah Lähner, Vladislav Golyanik, Christof Wunderlich, Christian Theobalt, and Michael Moeller. Q-match: Iterative shape matching via quantum annealing. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7586–7596, 2021.

[9] Harshil Bhatia, Edith Tretschk, Zorah Lähner, Marcel Seelbach Benkner, Michael Möller, Christian Theobalt, and Vladislav Golyanik. Ccuantumm: Cycle-consistent quantum-hybrid matching of multiple shapes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[10] Hongbo Bi, Ranwan Wu, Ziqi Liu, Huihui Zhu, Cong Zhang, and Tian-Zhu Xiang. Cross-modal hierarchical interaction network for rgb-d salient object detection. *Pattern Recognition*, 136:109194, 2023.

[11] Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. Quantum machine learning. *Nature*, 549(7671):195–202, 2017.

[12] Andreas Bircher, Mina Kamel, Kostas Alexis, Helen Oleynikova, and Roland Siegwart. Receding horizon" next-best-view" planner for 3d exploration. In *2016 IEEE international conference on robotics and automation (ICRA)*, pages 1462–1468. IEEE, 2016.

[13] Andreas Bircher, Mina Kamel, Kostas Alexis, Helen Oleynikova, and Roland Siegwart. Receding horizon path planning for 3d exploration and surface inspection. *Autonomous Robots*, 42:291–306, 2018.

[14] Federica Bogo, Javier Romero, Matthew Loper, and Michael J Black. Faust: Dataset and evaluation for 3d mesh registration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3794–3801, 2014.

[15] Nizar Bouhlel and Stéphane Méric. Maximum-likelihood parameter estimation of the product model for multilook polarimetric sar data. *IEEE Transactions on Geoscience and Remote Sensing*, 57(3):1596–1611, 2018.

[16] Antonio Chella, Salvatore Gaglio, Maria Mannone, Giovanni Pilato, Valeria Seidita, Filippo Vella, and Salvatore Zammuto. Quantum planning for swarm robotics. *Robotics and Autonomous Systems*, 161:104362, 2023.

[17] Cheng Chen, Juzheng Miao, Dufan Wu, Zhiling Yan, Sekeun Kim, Jiang Hu, Aoxiao Zhong, Zhengliang Liu, Lichao Sun, Xiang Li, et al. Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation. *arXiv preprint arXiv:2309.08842*, 2023.

[18] Gang Chen, Feng Shao, Xiongli Chai, Hangwei Chen, Qiuping Jiang, Xiangchao Meng, and Yo-Sung Ho. Cgmdrnet: Cross-guided modality difference reduction network for rgb-t salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(9):6308–6323, 2022.

[19] Qian Chen, Zhenxi Zhang, Yanye Lu, Keren Fu, and Qijun Zhao. 3-d convolutional neural networks for rgb-d salient object detection and beyond. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[20] SY Chen and YF Li. Vision sensor planning for 3-d model acquisition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 35(5):894–904, 2005.

[21] Titus Cieslewski, Elia Kaufmann, and Davide Scaramuzza. Rapid exploration with multi-rotors: A frontier selection method for high speed flight. In *2017*

IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 2135–2142. IEEE, 2017.

[22] Cl Connolly. The determination of next best views. In *Proceedings. 1985 IEEE international conference on robotics and automation*, volume 2, pages 432–435. IEEE, 1985.

[23] Micah Corah, Cormac O'Meadhra, Kshitij Goel, and Nathan Michael. Communication-efficient planning and mapping for multi-robot exploration in large environments. *IEEE Robotics and Automation Letters*, 4(2):1715–1721, 2019.

[24] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 303–312, 1996.

[25] Jeffrey Delmerico, Stefan Isler, Reza Sabzevari, and Davide Scaramuzza. A comparison of volumetric information gain metrics for active 3d object reconstruction. *Autonomous Robots*, 42(2):197–208, 2018.

[26] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022.

[27] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE international conference on computer vision*, pages 4548–4557, 2017.

[28] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. *arXiv preprint arXiv:1805.10421*, 2018.

[29] Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. A quantum approximate optimization algorithm. *arXiv preprint arXiv:1411.4028*, 2014.

[30] Edward Farhi and Hartmut Neven. Classification with quantum neural networks on near term processors. *arXiv preprint arXiv:1802.06002*, 2018.

[31] Matteo Farina, Luca Magri, Willi Menapace, Elisa Ricci, Vladislav Golyanik, and Federica Arrigoni. Quantum multi-model fitting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13640–13649, 2023.

[32] Mengyang Feng, Huchuan Lu, and Errui Ding. Attentive feedback network for boundary-aware salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1623–1632, 2019.

[33] Fadri Furrer, Michael Burri, Markus Achtelik, and Roland Siegwart. Rotors—a modular gazebo mav simulator framework. *Robot Operating System (ROS) The Complete Reference (Volume 1)*, pages 595–625, 2016.

[34] Lina Gao, Bing Liu, Ping Fu, and Mingzhu Xu. Depth-aware inverted refinement network for rgb-d salient object detection. *Neurocomputing*, 518:507–522, 2023.

[35] Georgios Georgakis, Bernadette Bucher, Anton Arapin, Karl Schmeckpeper, Nikolai Matni, and Kostas Daniilidis. Uncertainty-driven planner for exploration and navigation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 11295–11302. IEEE, 2022.

[36] Shizhan Gong, Yuan Zhong, Wenao Ma, Jinpeng Li, Zhao Wang, Jingyang Zhang, Pheng-Ann Heng, and Qi Dou. 3dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable medical image segmentation. *arXiv preprint arXiv:2306.13465*, 2023.

[37] Margarita Grinvald, Fadri Furrer, Tonci Novkovic, Jen Jen Chung, Cesar Cadena, Roland Siegwart, and Juan Nieto. Volumetric instance-aware semantic mapping and 3d object discovery. *IEEE Robotics and Automation Letters*, 4(3):3037–3044, 2019.

[38] Lov K Grover. A fast quantum mechanical algorithm for database search. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pages 212–219, 1996.

[39] Slawomir Grzonka, Giorgio Grisetti, and Wolfram Burgard. Towards a navigation system for autonomous indoor flying. In *2009 IEEE international conference on Robotics and Automation*, pages 2878–2883. IEEE, 2009.

[40] Haiyang He, Jing Wang, Xiaolin Li, Minglin Hong, Shiguo Huang, and Tao Zhou. Eaf-net: an enhancement and aggregation–feedback network for rgb-t salient object detection. *Machine Vision and Applications*, 33(4):64, 2022.

[41] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[42] Geoffrey A Hollinger and Gaurav S Sukhatme. Sampling-based robotic information gathering algorithms. *The International Journal of Robotics Research*, 33(9):1271–1287, 2014.

[43] Armin Hornung, Kai M Wurm, Maren Bennewitz, Cyrill Stachniss, and Wolfram Burgard. Octomap: An efficient probabilistic 3d mapping framework based on octrees. *Autonomous robots*, 34:189–206, 2013.

[44] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[45] Nianchang Huang, Qiang Jiao, Qiang Zhang, and Jungong Han. Middle-level feature fusion for lightweight rgb-d salient object detection. *IEEE Transactions on Image Processing*, 31:6621–6634, 2022.

[46] Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems*, 34:10944–10956, 2021.

[47] Fushuo Huo, Xuegui Zhu, Lei Zhang, Qifeng Liu, and Yu Shu. Efficient context-guided stacked refinement network for rgb-t salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(5):3111–3124, 2021.

[48] Tanveer Hussain, Abbas Anwar, Saeed Anwar, Lars Petersson, and Sung Wook Baik. Pyramidal attention for saliency detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2877–2887. IEEE, 2022.

[49] IBM. Qiskit. `https://www.ibm.com/quantum/qiskit`.

[50] Liren Jin, Xieyuanli Chen, Julius Rückin, and Marija Popović. Neu-nbv: Next best view planning using uncertainty estimation in image-based neural rendering. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11305–11312. IEEE, 2023.

[51] Xiao Jin, Kang Yi, and Jing Xu. Moadnet: Mobile asymmetric dual-stream networks for real-time and lightweight rgb-d salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7632–7645, 2022.

[52] Mina Kamel, Thomas Stastny, Kostas Alexis, and Roland Siegwart. Model predictive control for trajectory tracking of unmanned aerial vehicles using robot

operating system. *Robot Operating System (ROS) The Complete Reference (Volume 2)*, pages 3–39, 2017.

[53] Sertac Karaman and Emilio Frazzoli. Sampling-based algorithms for optimal motion planning. *The international journal of robotics research*, 30(7):846–894, 2011.

[54] Sebastian A Kay, Simon Julier, and Vijay M Pawar. Semantically informed next best view planning for autonomous aerial 3d reconstruction. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3125–3130. IEEE, 2021.

[55] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[56] Michael Krainin, Brian Curless, and Dieter Fox. Autonomous generation of complete 3d object models using next best view manipulation planning. In *2011 IEEE international conference on robotics and automation*, pages 5031–5037. IEEE, 2011.

[57] Simon Kriegel, Christian Rink, Tim Bodenmüller, Alexander Narr, Michael Suppa, and Gerd Hirzinger. Next-best-scan planning for autonomous 3d modeling. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2850–2856. IEEE, 2012.

[58] Simon Kriegel, Christian Rink, Tim Bodenmüller, and Michael Suppa. Efficient next-best-scan planning for autonomous 3d surface reconstruction of unknown objects. *Journal of Real-Time Image Processing*, 10(4):611–631, 2015.

[59] Jonathan Wei Zhong Lau, Kian Hwee Lim, Harshank Shrotriya, and Leong Chuan Kwek. Nisq computing: where are we and where do we go? *AAPPS bulletin*, 32(1):27, 2022.

[60] Steven M LaValle et al. Rapidly-exploring random trees: A new tool for path planning. 1998.

[61] Nam H Le, Milan Sonka, and Fatima Toor. A quantum optimization method for geometric constrained image segmentation. *arXiv preprint arXiv:2310.20154*, 2023.

[62] Zhengyi Liu, Xiaoshen Huang, Guanghui Zhang, Xianyong Fang, Linbo Wang, and Bin Tang. Scribble-supervised rgb-t salient object detection. *arXiv preprint arXiv:2303.09733*, 2023.

[63] Zhengyi Liu, Yacheng Tan, Qian He, and Yun Xiao. Swinnet: Swin transformer drives edge-aware rgb-d and rgb-t salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7):4486–4497, 2021.

[64] Seth Lloyd, Masoud Mohseni, and Patrick Rebentrost. Quantum principal component analysis. *Nature physics*, 10(9):631–633, 2014.

[65] Jiebo Luo, Chang Wen Chen, and Kevin J Parker. Applications of gibbs random field in image processing: from segmentation to enhancement. In *Visual Communications and Image Processing'94*, volume 2308, pages 1289–1300. SPIE, 1994.

[66] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 248–255, 2014.

[67] Jasna Maver and Ruzena Bajcsy. Occlusions as a guide for planning the next view. *IEEE transactions on pattern analysis and machine intelligence*, 15(5):417–433, 1993.

[68] Zehui Meng, Hailong Qin, Ziyue Chen, Xudong Chen, Hao Sun, Feng Lin, and Marcelo H Ang. A two-stage optimized next-view planning framework for

3-d unknown environment exploration, and structural reconstruction. *IEEE Robotics and Automation Letters*, 2(3):1680–1687, 2017.

[69] Rohit Menon, Tobias Zaenker, Nils Dengler, and Maren Bennewitz. Nbv-sc: Next best view planning based on shape completion for fruit mapping and reconstruction. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4197–4203. IEEE, 2023.

[70] John P Morgan Jr and Richard L Tutwiler. Real-time reconstruction of depth sequences using signed distance functions. In *Signal Processing, Sensor/Information Fusion, and Target Recognition XXIII*, volume 9091, pages 388–399. SPIE, 2014.

[71] Menaka Naazare, Francisco Garcia Rosas, and Dirk Schulz. Online next-best-view planner for 3d-exploration and inspection with a mobile manipulator robot. *IEEE Robotics and Automation Letters*, 7(2):3779–3786, 2022.

[72] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE international symposium on mixed and augmented reality*, pages 127–136. Ieee, 2011.

[73] Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Marc Stamminger. Real-time 3d reconstruction at scale using voxel hashing. *ACM Transactions on Graphics (ToG)*, 32(6):1–11, 2013.

[74] Helen Oleynikova, Zachary Taylor, Marius Fehr, Roland Siegwart, and Juan Nieto. Voxblox: Incremental 3d euclidean signed distance fields for on-board mav planning. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1366–1373. IEEE, 2017.

[75] Open Robotics. Gazebo models. `https://github.com/osrf/gazebo_models`, 2021.

[76] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *2012 IEEE conference on computer vision and pattern recognition*, pages 733–740. IEEE, 2012.

[77] Joelle Pineau, Geoff Gordon, Sebastian Thrun, et al. Point-based value iteration: An anytime algorithm for pomdps. In *Ijcai*, volume 3, pages 1025–1032, 2003.

[78] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, Andrew Y Ng, et al. Ros: an open-source robot operating system. In *ICRA workshop on open source software*, volume 3, page 5. Kobe, Japan, 2009.

[79] Dhanesh Ramachandram and Graham W Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine*, 34(6):96–108, 2017.

[80] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[81] Patrick Rebentrost, Masoud Mohseni, and Seth Lloyd. Quantum support vector machine for big data classification. *Physical review letters*, 113(13):130503, 2014.

[82] Victor Massagué Respall, Dmitry Devitt, Roman Fedorenko, and Alexandr Klimchik. Fast sampling-based next-best-view exploration algorithm for a mav.

In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 89–95. IEEE, 2021.

[83] Szymon Rusinkiewicz and Marc Levoy. Efficient variants of the icp algorithm. In *Proceedings third international conference on 3-D digital imaging and modeling*, pages 145–152. IEEE, 2001.

[84] Lukas Schmid, Michael Pantic, Raghav Khanna, Lionel Ott, Roland Siegwart, and Juan Nieto. An efficient sampling-based method for online informative path planning in unknown environments. *IEEE Robotics and Automation Letters*, 5(2):1500–1507, 2020.

[85] William R Scott, Gerhard Roth, and Jean-François Rivest. View planning for automated three-dimensional object reconstruction and inspection. *ACM Computing Surveys (CSUR)*, 35(1):64–96, 2003.

[86] Magnus Selin, Mattias Tiger, Daniel Duberg, Fredrik Heintz, and Patric Jensfelt. Efficient autonomous exploration planning of large-scale 3-d environments. *IEEE Robotics and Automation Letters*, 4(2):1699–1706, 2019.

[87] David Silver and Joel Veness. Monte-carlo planning in large pomdps. *Advances in neural information processing systems*, 23, 2010.

[88] Kechen Song, Jie Wang, Yanqi Bao, Liming Huang, and Yunhui Yan. A novel visible-depth-thermal image dataset of salient object detection for robotic visual perception. *IEEE/ASME Transactions on Mechatronics*, 2022.

[89] Mengke Song, Wenfeng Song, Guowei Yang, and Chenglizhao Chen. Improving rgb-d salient object detection via modality-aware decoder. *IEEE Transactions on Image Processing*, 31:6124–6138, 2022.

[90] Soohwan Song and Sungho Jo. Online inspection path planning for autonomous 3d modeling using a micro-aerial vehicle. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6217–6224. IEEE, 2017.

[91] Soohwan Song, Daekyum Kim, and Sungho Jo. Online coverage and inspection planning for 3d modeling. *Autonomous Robots*, 44(8):1431–1450, 2020.

[92] Marco Steinbrink, Philipp Koch, Bernhard Jung, and Stefan May. Rapidly-exploring random graph next-best view exploration for ground vehicles. In *2021 European Conference on Mobile Robots (ECMR)*, pages 1–7. IEEE, 2021.

[93] Congying Sui, Kejing He, Congyi Lyu, and Yun-Hui Liu. Accurate 3d reconstruction of dynamic objects by spatial-temporal multiplexing and motion-induced error elimination. *IEEE Transactions on Image Processing*, 31:2106–2121, 2022.

[94] Sebastian Thrun. Probabilistic robotics. *Communications of the ACM*, 45(3):52–57, 2002.

[95] Zhengzheng Tu, Zhun Li, Chenglong Li, and Jin Tang. Weakly alignment-free rgbt salient object detection with deep correlation network. *IEEE Transactions on Image Processing*, 31:3752–3764, 2022.

[96] J Irving Vasquez-Gomez, L Enrique Sucar, Rafael Murrieta-Cid, and Efrain Lopez-Damian. Volumetric next-best-view planning for 3d object reconstruction with positioning error. *International Journal of Advanced Robotic Systems*, 11(10):159, 2014.

[97] Bin Wan, Xiaofei Zhou, Yaoqi Sun, Tingyu Wang, Chengtao Lv, Shuai Wang, Haibing Yin, and Chenggang Yan. Mffnet: Multi-modal feature fusion network for vdt salient object detection. *IEEE Transactions on Multimedia*, 2023.

[98] Jie Wang, Kechen Song, Yanqi Bao, Liming Huang, and Yunhui Yan. Cgfnet: Cross-guided fusion network for rgb-t salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(5):2949–2961, 2021.

[99] Junde Wu, Rao Fu, Huihui Fang, Yuanpei Liu, Zhaowei Wang, Yanwu Xu, Yueming Jin, and Tal Arbel. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023.

[100] Zongwei Wu, Shriarulmozhivarman Gobichettipalayam, Brahim Tamadazte, Guillaume Allibert, Danda Pani Paudel, and Cédric Demonceaux. Robust rgb-d fusion for saliency detection. In *2022 International Conference on 3D Vision (3DV)*, pages 403–413. IEEE, 2022.

[101] Dan Xiang, Hanxi Lin, Jian Ouyang, and Dan Huang. Combined improved a* and greedy algorithm for path planning of multi-objective mobile robot. *Scientific Reports*, 12(1):13273, 2022.

[102] Zhengxuan Xie, Feng Shao, Gang Chen, Hangwei Chen, Qiuping Jiang, Xiangchao Meng, and Yo-Sung Ho. Cross-modality double bidirectional interaction and fusion network for rgb-t salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[103] Zhefan Xu, Di Deng, and Kenji Shimada. Autonomous uav exploration of dynamic environments via incremental sampling and probabilistic roadmap. *IEEE Robotics and Automation Letters*, 6(2):2729–2736, 2021.

[104] Brian Yamauchi. A frontier-based approach for autonomous exploration. In *Proceedings 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA'97.'Towards New Computational Principles for Robotics and Automation'*, pages 146–151. IEEE, 1997.

[105] Luke Yoder and Sebastian Scherer. Autonomous exploration for infrastructure modeling with a micro aerial vehicle. In *Field and Service Robotics: Results of the 10th International Conference*, pages 427–440. Springer, 2016.

[106] Jan-Nico Zaech, Alexander Liniger, Martin Danelljan, Dengxin Dai, and Luc Van Gool. Adiabatic quantum computing for multi object tracking. In *Proceed-*

ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8811–8822, 2022.

[107] Tobias Zaenker, Julius Rückin, Rohit Menon, Marija Popović, and Maren Bennewitz. Graph-based view motion planning for fruit detection. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4219–4225. IEEE, 2023.

[108] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023.

[109] Chaoning Zhang, Dongshen Han, Sheng Zheng, Jinwoo Choi, Tae-Ho Kim, and Choong Seon Hong. Mobilesamv2: Faster segment anything to everything. *arXiv preprint arXiv:2312.09579*, 2023.

[110] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Xin Yu, Yiran Zhong, Nick Barnes, and Ling Shao. Rgb-d saliency detection via cascaded mutual information minimization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4338–4347, 2021.

[111] Qiang Zhang, Qi Qin, Yang Yang, Qiang Jiao, and Jungong Han. Feature calibrating and fusing network for rgb-d salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[112] Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu, and Xiang Ruan. Self-supervised pretraining for rgb-d salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3463–3471, 2022.

[113] Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. Fast segment anything. *arXiv preprint arXiv:2306.12156*, 2023.

[114] Heng Zhou, Chunna Tian, Zhenxi Zhang, Chengyang Li, Yuxuan Ding, Yongqiang Xie, and Zhongbo Li. Position-aware relation learning for rgb-thermal salient object detection. *IEEE Transactions on Image Processing*, 2023.

[115] Tao Zhou, Huazhu Fu, Geng Chen, Yi Zhou, Deng-Ping Fan, and Ling Shao. Specificity-preserving rgb-d saliency detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4681–4691, 2021.

[116] Wujie Zhou, Yun Zhu, Jingsheng Lei, Rongwang Yang, and Lu Yu. Lsnet: Lightweight spatial boosting network for detecting salient objects in rgb-thermal images. *IEEE Transactions on Image Processing*, 32:1329–1340, 2023.