A Sensor Fusion Platform for Semantic Segmentation of Outdoor Ice Scenes

by

Richard Rui Jia Duan

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Mechanical and Mechatronics Engineering

Waterloo, Ontario, Canada, 2026

© Richard Rui Jia Duan 2026

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Richard Duan is the sole author of Chapters 6 and 7, which were written under the supervision of Dr. Zhao Pan and Dr. Andrea Scott and were not written for publication.

Components of Chapters 1-5 were submitted to POAC'25 — St. John's as part of [1]. Richard Duan wrote this paper under the supervision of Dr. Zhao Pan and Dr. Andrea Scott, with contributions from Corwin MacMillan from the University of Waterloo and Robert Gash from the Ocean Coastal and River Engineering Research Centre, National Research Council Canada. This work served as a preliminary study with the same motivation, data collection process, and dataset scenes as the thesis. However, the thesis expands on the scope and develops these ideas further, thus only some portions of these chapters overlap with the published work.

Abstract

Due to climate change and global warming, Arctic sea ice coverage has been on a downwards trend that is expected to continue late into this century, opening up new shipping routes through the Arctic. These routes provide opportunities for much shorter transoceanic journeys compared to currently used routes, offering significant savings in transportation time and resource cost. However, travelling through ice-covered waters still provides a constant source of risk from collisions with ice damaging or even sinking the vessel. Thus, modern tools such as ice charts, satellite imagery, and near-field observations from on-board experts are still used to avoid dangerous collisions with ice. In adverse weather and lighting conditions, near-field human observations can be uncertain, causing inaccuracies in determining nearby ice conditions.

This thesis introduces a system composed of an optical RGB camera, a thermal infrared camera, and a polarization camera to aid in such near-field observations for the purposes of determining nearby ice conditions in a variety of outdoor conditions. The novel sensor suite was tested during field trials conducted in February 2025, collecting river ice data across multiple days and at different locations along the Ottawa and St. Lawrence rivers. After image registration and post processing, a first-of-its-kind, fully labelled semantic segmentation dataset of 118 images across 5 scenes was created to test different methods of sensor fusion. For testing, the test set was further split into easy and hard subsets based on perceived clarity to the human eye. Early, middle, and late fusion networks using different combinations of sensor inputs were created based on modifying the widely used fully convolutional neural network U-Net architecture with a pre-trained ResNet-18 backbone.

Experimental results show that early fusion methods performed consistently worse than the baseline case of only using the optical RGB data on both the easy and hard test sets, regardless of the other input sources used, most likely due to the higher levels of noise being introduced by the additional sources. Middle fusion with polarization and thermal managed to outperform the baseline on the hard test set, and late fusion with polarization managed to outperform on the easy test set. While both middle and late fusion methods show improvements over the baseline through extracting useful information even from noisy sources, late fusion with polarization had the highest overall mIoU improvement on the full test set due to an enhanced ability to differentiate between brash ice and water.

Ultimately, sensor fusion shows potential for improving sea ice classification accuracy while being robust to a variety of different environmental and lighting conditions. These preliminary results serve to support continued development of sensor fusion platforms as tools to aid in the tracking and identification of nearby ice conditions.

Acknowledgements

I would like to show my gratitude to all those who made this thesis possible. It has been a much longer journey than I expected with many ups and downs, but I will always remain grateful for all of the knowledge and experiences I have gained throughout the course of completing my master's.

First, I would like to sincerely thank my advisors Dr. Zhao Pan and Dr. Andrea Scott for their endless patience and guidance offered to me. There were many times during this degree where I would have remained lost and stuck, had it not been for their valuable feedback and support that helped me get back on track. I would also like to thank my committee members Dr. Grant Gunn and Dr. Russell Buchanan for their help in refining my thesis.

I greatly appreciate my lab mates Corwin MacMillan and Maksym Veremchuk for their collaboration and insightful discussions. Many parts of this project stands on the foundation they built, learning from their challenges and successes while walking the same path before me.

I would also like to acknowledge the important role that financial support played in this project, with funding from the Natural Sciences and Engineering Research Council of Canada and the University of Waterloo providing me with the resources to pursue this project to its completion. I am grateful to Matthew Garvin and Robert Gash from the National Research Council Canada for their part in overseeing the grant, and for providing exclusive access to reference materials and the NRC-OCRE facilities during preliminary testing.

I owe special thanks to my friends and family for their support during this time. My parents have been a constant source of encouragement, always pushing and believing in me to do the best that I can. Meanwhile, my friends have been a great source of relatability and relief when I needed to vent about the same thing for the n-th time.

Lastly, I would like to thank my dearly departed mischief of Badgey, Jazzy, and Puffy for teaching me an important life lesson — that even seemingly impossibly small holes can be squeezed through with enough will and determination.

Table of Contents

Author's Declaration ii

Statement of Contributions iii

Abstract iv

Acknowledgements v

List of Figures ix

List of Tables xi

1 Introduction 1

1.1 Motivation 1

1.2 Contributions 4

2 Background 5

2.1 Detection and Classification of River and Sea Ice 5

2.2 Related Works 6

2.3 Multi-Sensor Perception Systems 6

2.4 Introduction to U-Net 7

3 Sensor System Design **9**

3.1 Sensor Selection . . . 9

3.2 System Requirements and Constraints . . . 10

3.3 First Iteration . . . 11

3.4 Second Iteration . . . 13

3.5 Final Iteration . . . 17

3.6 Sensor Calibration and Synchronization . . . 20

4 Dataset Creation **24**

4.1 Data Collection . . . 24

4.2 Image Pre-Processing . . . 28

4.3 Ground Truth Annotation . . . 31

4.4 Dataset Exploration . . . 33

5 Experimental Setup **36**

5.1 Fusion Architecture and Datasets . . . 36

5.1.1 Early Fusion . . . 36

5.1.2 Middle Fusion . . . 39

5.1.3 Late Fusion . . . 41

5.2 Training Details . . . 43

6 Results and Discussion **45**

6.1 Performance Results . . . 45

6.1.1 Early Fusion . . . 45

6.1.2 Middle Fusion . . . 49

6.1.3 Late Fusion . . . 53

7 Conclusion **56**

7.1 Summary 56

7.2 Future Work 57

7.2.1 Improved Training Dataset 57

7.2.2 Depth Map Refinement 58

7.2.3 Thermal Data Usage 58

References **59**

List of Figures

1.1 Comparison of current transoceanic routes (solid lines) with proposed routes through the Arctic (dashed lines) using the Northwest Passage (left) and the Northern Sea Route (right) to travel between popular ports (adapted from [2]). 2

3.1 Connection diagram of the first sensor suite iteration. 12

3.2 Close-up of the custom multi-camera unit with 4 linear polarization filtered cameras on the left and an unobstructed baseline camera on the right. . . . 13

3.3 Connection diagram of the second sensor suite iteration. 15

3.4 Setup of the second iteration of the sensor suite mounted on a model ship operating in the ice tank. 17

3.5 Connection diagram of the final sensor suite iteration. 18

3.6 Setup of the final sensor suite. 19

3.7 (a) Full sensor system with each camera labelled. Camera calibration was conducted using a rigid asymmetric circle grid board, with an example frame shown from each of the synchronized views of (b) the thermal camera, (c) the polarization camera, (d) the baseline RGB camera, (e) and (f) the left and right pair of RGB stereoscopic cameras. 21

4.1 Map of the data collection locations in red for each scene, numbered according to the scenes listed in Table 4.1. 26

4.2 Example of inconsistency in the generated disparity maps of consecutive frames. 29

4.3 Progressive labelling of the dataset by incorporating information from each sensor to ensure that details do not get lost. 32

4.4 Example RGB images from the easy and hard subsets. 34
4.5 Histograms of class distributions in (a) the training set, (b) the full test set, (c) the easy test subset, and (d) the hard test subset. 35

5.1 Pairs of example inputs from each sensor and the two false-colour images from the easy (top row) and hard (bottom row) test subsets. 37
5.2 Custom U-Net architecture for early fusion with a ResNet encoder backbone. Note that the initial $c_1+c_2+...+c_n \rightarrow 3$ convolutional layer was not included for the baseline RGB case, which only has 3 channels. 38
5.3 Custom U-Net architecture for middle fusion with a ResNet encoder backbone. Note that although only 2 inputs are shown in this example, additional inputs can be included through adding on additional encoder branches with their respective skip connections. 40
5.4 Custom U-Net architecture for late fusion with a ResNet encoder backbone. Note that although only 2 inputs are shown in this example, additional inputs can be included through adding on additional U-Net branches. . . . 42

6.1 Confusion matrices of different early fusion methods compared to the baseline on the full dataset. 46
6.2 Pairs of example outputs of different early fusion methods on the easy (left) and hard (right) test subsets compared against the baseline. 48
6.3 Confusion matrices of different middle fusion methods compared to the baseline on the full dataset. 50
6.4 Confusion matrices of different middle fusion methods compared to the baseline on the hard dataset. 51
6.5 Pairs of example outputs of different middle fusion methods on the easy (left) and hard (right) test subsets compared against the baseline. 52
6.6 Confusion matrices of different late fusion methods compared to the baseline on the full dataset. 54
6.7 Pairs of example outputs of different late fusion methods on the easy (left) and hard (right) test subsets compared against the baseline. 55

List of Tables

3.1 Sensors Included in the First Iteration 11
3.2 Sensors Included in the Second Iteration 14
3.3 Sensors Included in the Final Iteration 19

4.1 Data Collection Location Details 27
4.2 Scene Image Distribution 33

6.1 Early Fusion Results 45
6.2 Middle Fusion Results 49
6.3 Late Fusion Results 53

Chapter 1

Introduction

1.1 Motivation

With climate change in full effect, Arctic sea ice cover has been trending downwards in terms of both absolute ice coverage as well as the almost complete loss of old multiyear ice [3]. Thanks to the advancement of remote sensing platforms such as the European Space Agency's CryoSat-2 and the National Aeronautics and Space Administration's ICESat making sea ice thickness measurements possible at large scales [4, 5], sea ice volume can be more accurately estimated throughout the year. With shorter ice seasons, larger sections of the Canadian Arctic Archipelago (CAA) are remaining ice free for longer periods of the year, although the existence of choke points where multi-year ice gathers still limits its current usability for shipping [6]. The openings in ice provide opportunities to explore previously impassable routes, such as the Northwest Passage (NWP) as seen in Figure 1.1, for use in commercial shipping up to 4 months of the year by the middle of the 21st century [7, 8]. The use of such routes is of interest both economically and environmentally due to potential savings in distance compared to current routes. For example, a simulation conducted by Somanathan et al. shows that a trip between the port cities St. John's, Newfoundland and Yokohama, Japan can be shortened by 3500 nautical miles — a 33% decrease — by traversing through the NWP [9]. With over 90% of international goods being shipped by sea [10], any significant savings in distance travelled and decreased reliance on the Panama and Suez canals makes these routes an attractive choice for ship operators. Additionally, increased marine traffic through the CAA would stimulate economic activity in remote Northern communities along the path of the ships [11, 12].

The exploration of Arctic waters for potential shipping routes is not a recent devel-

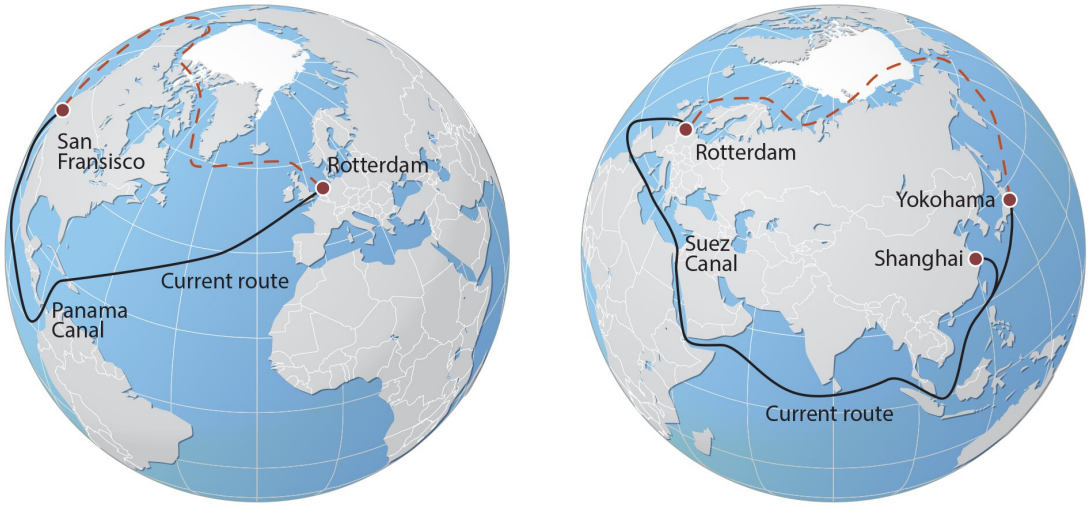


Figure 1.1: Comparison of current transoceanic routes (solid lines) with proposed routes through the Arctic (dashed lines) using the Northwest Passage (left) and the Northern Sea Route (right) to travel between popular ports (adapted from [2]).

opment. A notable early attempt was the Franklin expedition of 1845, which ended in tragedy when both the HMS *Erebus* and HMS *Terror* became icebound in the Canadian Arctic and resulted in the loss of all crew members involved in the expedition. Even with modern advances in technology and marine safety, it is still hazardous to navigate in ice-covered waters [13]. There are constant dangers of hull damage from collisions with ice or becoming beset within a high-pressure ice field. As such, accurate information of upcoming ice conditions is crucial for safe operation in ice-covered waters. Finding an optimal path through ice-covered waters is a difficult task involving navigation efficiency, icebreaking cost, and navigation safety [14–16]. Ships that operate within this environment are classified according to the level of ice strengthening applied to their hull. Although the highest rating of Polar Class 1 theoretically enables year-round operation in all polar waters, very few ships manage to obtain this rating [17]. Thus, even the strongest ships travelling through the Arctic need to avoid certain ice conditions to avoid risking damage or becoming trapped [18], with special care required to cater the route to the specific ship's abilities [**?**].

While modern ice charts and satellite imagery are crucial for general route planning, on-site assessments of ice conditions surrounding a ship from specially trained ice experts are still required for real-time navigation [19]. Typically, detailed on-board observations are

limited to what can be seen ahead by the observer for near-ship ice conditions. However, such observations are somewhat subjective in nature and can become ineffective in low visibility conditions caused by adverse weather or at night. Furthermore, the human eye is limited to seeing only the narrow spectrum of visible light, losing out on potentially important information present in the rest of the spectrum and in the polarization of light that can more clearly identify the upcoming ice field in a wide range of weather conditions.

To supplement on-board human observations, researchers have employed a wide range of sensor modalities to determine nearby ice conditions. The most commonly used are ship-mounted optical sensors that can cover a large field of view, enabling a computer to assess the entire scene at once. In addition, specialized cameras allows for analysis in wavelengths of light not available to the human eye, such as in the near-infrared (NIR), short-wavelength infrared (SWIR), mid-wavelength infrared (MWIR), and long-wavelength infrared (LWIR) ranges. Combined with a relatively high framerate providing dense temporal information, consistency can be enforced from frame to frame. Radar and LIDAR have also been incorporated to directly measure the surface topography of ice features, providing estimates of ice thickness and surface roughness while also tracking the movement of individual floes. More accurate estimates of ice thickness can be obtained through the use of ground-penetrating radar [20] and electromagnetic induction sensors [21], although the use of these sensors is more niche due to their close proximity to the ice surface required for measurements. Additionally, the real-time effect of the surrounding ice conditions on ship dynamics can be measured through recording hull vibrations during collisions with ice using accelerometers to estimate the ramming loads and ice resistance generated to distinguish between different levels of ice coverage [22]. However, limited methods to process and interpret the massive amount of collected data restricts the widespread adoption of using these supplementary observations.

Advancements in computing hardware and machine learning has made it easier to process large amounts of data in close to real-time. When applied to the problem of environmental perception for navigational purposes, sensor systems can lighten the load from human operators by continuously monitoring nearby conditions and keeping track of numerous points of interest at the same time. After post-processing to extract useful feature information, such observations can then be fed into a decision support system to serve as a navigational aid [23], taking into account vehicle interactions with the environment to plot an efficient route to the target destination [24]. However, this typically requires custom-designed system architectures with specially collected and labelled data for training. Compared to autonomous driving, there are much fewer opportunities to collect training data on various types of ice in the near field, especially on board ships travelling through ice-covered waters. Even once collected, expertise is required to differentiate and

label the upcoming ice scene. Thus, there is a lack of high-quality training data available in this niche.

1.2 Contributions

The work presented in this thesis serves as a comprehensive study on the usage of different sensors to determine near field ice conditions for monitoring and navigational purposes. The contributions of this work are as follows:

1. A novel sensor suite is introduced, specialized for collecting optical data to identify different ice conditions using an array of different cameras. Details are included explaining the design decisions that led to the choice of the included sensors and detailing the methodology used in synchronizing, calibrating, and spatially registering the different sensor views in preparation for sensor fusion.

2. A first-of-its-kind, fully labelled multi-sensor semantic segmentaion dataset of river ice in the near field is created. The data was collected during field tests by the introduced sensor suite in a variety of environmental conditions, providing a rare resource in an area with few sources of high-quality publicly available datasets.

3. Different sensor fusion strategies are explored using custom-designed deep learning networks to combine image data from multiple input sources. The strategies are evaluated using various combinations of the collected data, compared against a single sensor as the baseline case.

Together, these contributions establish a fundamental platform for future research in multi-sensor perception of sea ice in the near field, and offer a unique dataset for the advancement of sensor fusion strategies in ice classification.

Chapter 2

Background

2.1 Detection and Classification of River and Sea Ice

Monitoring the formation and movement of river and sea ice is important for water management and secure offshore activities. In rivers, large accumulations of ice can form ice jams and cause flooding to nearby areas when melting [25, 26], whereas finding open water regions in sea ice ensures safe navigation. Additionally, accurate measurements of information such as ice type, drift speed, and concentration can show trends in river and sea ice to be used in climate forecasting.

Typically, images of river and sea ice are collected either via remote sensing through satellites or through relatively small-scale campaigns with airborne vehicles. Ice classification is the task of identifying the presence of ice and categorizing the type of ice through analyzing its features. The World Meterological Organization (WMO) publishes a list of common features for each type of ice, such as surface roughness, thickness, shape, colour, and interactions with surrounding ice to help standardize ice nomenclature [27]. However, ice charting typical only requires a few major categories, with images interpreted by experienced experts either by assigning a single label/concentration to the entire image or applying polygonized labels to large homogenous regions [28, 29]. While it has the potential to provide a greater amount of information, human-based fine pixel-wise labelling of ice through semantic segmentation remains out of reach for most use cases due to the time-consuming nature of the task. Thus, ice segmentation is typically automated through image processing techniques, either through traditional thresholding methods or through deep-learning methods [25].

2.2 Related Works

With rising interest in incorporating autonomous systems to aid in ship navigation through ice-covered waters, a considerable number of different approaches have been developed recently for both collecting and processing near-field ice data. Several studies have conducted sea-faring trials with custom sensor packages brought on board the ship to collect data in a variety of conditions. Brown et al. introduced a suite of cameras and sensors that were mounted on the CCGS *Henry Larsen* as it travelled through sea ice between Newfoundland and Labrador in March of 2022 [30]. Many other studies collected unique optical data of near-field sea ice to determine statistics about upcoming ice conditions through image processing techniques [31–33]. Furthermore, Panchi and Kim collected optical data of sea ice during harsh weather conditions to create a dataset of weather-degraded images for developing a weather-resilient model [34]. Due to the costs and rarity of sea trials, studies have also used data collected in a simulated environment. For example, Zhou et al. captured optical data by attaching a camera to a model ship in an indoor ice basin [35].

After collecting relevant ice data, processing must be done to extract useful information such as ice concentration, floe size distribution, and the locations of upcoming icebergs. For image-based data, computer vision techniques and deep learning models are used to identify and distinguish different types of sea ice for navigational purposes. Kim et al. introduced an object classification network to distinguish between nine different categories of ice, trained on publicly available close-range imagery of ice [36]. Zhang et al. modified DeepLabv3+, a state-of-the-art semantic segmentation model, to conduct pixel-wise ice image segmentation [37]. The dataset used in the study consisted of manually labelled images captured by an optical camera, mounted on an ice-strengthened cargo ship as it travelled through the Arctic. Similarly, Dowden et al. applied the existing SegNet and PSPNet101 networks to conduct semantic segmentation on images captured on board an ice breaker as it completed an Antarctic expedition [38]. The simulated ice data collected by Zhou et al. was used to train YOLACT, a deep learning network, to perform instance segmentation on the scene [35]. As the goal was to estimate ice concentration and floe size distribution, only three classes (ship, ice, and water) were used for labelling this dataset.

2.3 Multi-Sensor Perception Systems

For a system to outperform human observations, it is clear that no single sensor can return useful information in all environmental conditions. A review by Fung et al. found that in the field of autonomous vehicles, state-of-the-art perception systems all leverage

the power of fusing together information from a complementary assortment of different sensors (such as cameras, GPS units, radar, and LIDAR for autonomous driving) [39]. For example, Rankin and Matthies used a combination of passive sensors composed of visible, multispectral (including NIR, SWIR, MWIR, and LWIR), polarization, and stereo sensors to detect mud hazards for autonomous off-road navigation [40]. Each sensor received different cues depending on the moisture content of the soil, which can then be cross-referenced with each other to obtain a high confidence in the prediction in a wide range of environmental conditions. The previously mentioned study by Brown et al. used a suite of different sensors to collect field data of sea ice on board the CCGS *Henry Larsen* icebreaker — including stereoscopic and 360° cameras, LIDAR, and an electromagnetic sensor [30]. While individual sensor performance on extracting sea ice information was evaluated, fusion of the different sensor data was not explored.

In general, an appropriate data fusion strategy must be chosen to effectively use the multimodal data collected by different sensors. While there are many ways to combine information from multiple sensors, three general locations exist within the data processing pipeline where sensor fusion can occur. Data level or early fusion combines the calibrated data from all sensors and is usually applied when the data modality is similar across the sensors, such as in the case of multiple cameras. Feature level or middle fusion first conducts feature extraction on the calibrated data from each sensor separately, before combining the detected features to be further analyzed. Finally, decision level or late fusion combines the analyzed output from each individual sensor to generate a single decision, and is usually applied to extremely different data formats that are difficult to combine.

2.4 Introduction to U-Net

U-Net is a fully convolutional neural network (CNN) initially developed to process medical imaging [41]. A convolution in the context of image processing is a kernel-based function that takes information from surrounding pixels and passes it through a filter. An important property of convolution is the use of shift-invariant kernels, which applies the same transformation throughout the entire image regardless of its location. Depending on the composition of the kernel, it can serve as a filter to highlight certain image features while ignoring others. CNNs take advantage of this property to pass the original images through a progressive series of different filters, transforming them into feature maps at each layer

in the process. The convolution of a 2D image is defined as:

$$g(x,y) = h(x,y) * f(x,y) = \sum_{i=-a}^{a} \sum_{j=-b}^{b} h(i,j) \cdot f(x-i, y-j) \tag{2.1}$$

where f is the input image, h is a kernel with size $(2a+1) \times (2b+1)$, and g is the output convolution of h and f. To maintain the same image dimensions at the output as the input, images are typically padded at the borders during the process. Compared to a fully connected network where each pixel has a corresponding neuron associated with it, a CNN is much more efficient at processing spatial information, and can be easily adapted to images of different scales and resolutions without needing to make significant changes to the network architecture.

U-Net leverages the power of a CNN through an autoencoder-style architecture, where the input image is first downsampled through a series of encoding convolutional layers, before being upsampled back to its original size through a similar series of decoding deconvolutional layers. Sometimes, an existing neural network can be used in place of the custom U-Net encoder as the feature extraction backbone, taking advantage of generalizability of pretrained networks trained on large datasets [42]. A unique aspect of the U-Net architecture is the inclusion of skip connections between the encoder and decoder. This ensures that spatial information is still retained throughout the network, even as spatial resolution is reduced in favour of increasing feature information in the deeper layers.

Chapter 3

Sensor System Design

3.1 Sensor Selection

RGB cameras are typically included in autonomous perception systems for their low cost and close resemblance to human vision, but image quality can be easily degraded by poor environmental and lighting conditions. Since thermal cameras do not rely on the presence of a visible light source, they are commonly used in combination with RGB cameras to provide better performance in low-light conditions. Another popular combination is RGB with depth information, which can provide more context to RGB data by providing information on the size and location of different objects in view, improving accuracy in tasks involving object tracking and localisation [43]. While LIDAR and radar can be used to obtain depth information, stereoscopic cameras offer a more accessible option requiring only a pair of calibrated cameras. Therefore, taking inspiration from existing multimodal systems in autonomous vehicles, RGB, thermal, and stereoscopic cameras were chosen to be tested in the system.

More specific to ice classification, field studies conducted by Perovich measured the reflection of light from Arctic sea ice at various wavelengths and polarization angles [44]. These measurements showed significant differences in the linear polarization of reflected light across different types of ice and snow, due to features such as surface roughness and internal structure affecting the polarization of incoming light. Additionally, polarization can reduce specular reflections through filtering out specific angles of polarized light, removing distractions and occlusions [45]. With advances in microgrid polarimeters, light at different polarization angles can be captured as a single image using a special polarization

camera [46]. Thus, the advantages and accessibility of polarization imaging motivated the inclusion of a polarization camera into the system.

Beyond various forms of imaging, ship dynamics as a result of colliding with ice can also be used to differentiate nearby ice conditions. The thickness and structure of different types of ice affects the ramming loads generated on the hull as it passes through, which can be measured through the use of accelerometers placed on board the ship [19]. While this data can only provide retrospective information on nearby ice conditions, it could serve as an additional measurement on the previous state of the ice field and adjust future predictions as needed. Due to a tendency for accelerometers to drift over time, GPS units are commonly paired as a way to anchor the system and correct sensor bias [39]. Additionally, the GPS unit provides locational data logs which can be compared to charts and satellite imagery afterwards as an extra source of verification. Therefore, accelerometers and GPS units were also considered for the sensor system.

3.2 System Requirements and Constraints

The vision of the multi-sensor system was for it to be small and self-contained so it could be easily incorporated on the bridge of the ship. This stand-alone system allows for flexibility in the location where it is placed to maximize visibility, without needing to install multiple sensors in different locations and run cables throughout the ship. As such, the system would also have a similar field of view as the crew, acting as an additional set of eyes and enhancing their current view of the upcoming ice conditions.

To validate the performance of such a system and to facilitate the testing of different sensors, a set of constraints and requirements were created for the prototype system. To allow for preliminary field data collection, the system needed to be portable enough that a single person can transport and set the system up on location. As such, the system needed to be self-contained without relying on external sources of power. Additionally, considering the expected environmental conditions required for ice to form in bodies of water, the sensor suite needed to be fully weatherproof to operate in below-freezing temperatures while being protected from water ingression. Initially, a requirement was set that all sensors had to use the same interfacing protocol to ensure ease of synchronization and organization. However, this requirement was later dropped to sacrifice ease of software integration for improved capture quality.

3.3 First Iteration

Initially, a low-budget sensor system was envisioned, centred on simple, low-level camera sensors connected to a modified Raspberry Pi board. The sensors included in this system are shown in Table 3.1.

Table 3.1: Sensors Included in the First Iteration

Sensor Type	**Sensor Model**	**Resolution/Range**	**Sampling/ Capture Rate**
Baseline RGB Camera	1 × Arducam IMX519	4656 × 3496	1 FPS
Stereoscopic Camera	2 × Raspberry Pi HQ Camera IMX477R	4056 × 3040	1 FPS
Polarization Camera	4 × Arducam IMX519	2320 × 1748	1 FPS
IMU	1 × MPU-6050	±2000 °/s (gyro) ±16 g (accel.)	100 Hz
GPS	1 × BN-880	±2 m	1 Hz

The Raspberry Pi Compute Module 4 [47] system-on-a-chip was chosen as the sensor interface due to its versatility and ease of integration with external sensors. Supporting the two Compute Module 4 chips were two StereoPi V2 boards acting as the carrier input/output board. The StereoPi V2 board [48] allows for the connection of two simultaneous Camera Serial Interface-2 (CSI-2) inputs, giving a total of 4 avaliable CSI-2 inputs. On one of the StereoPi V2 boards, one CSI-2 input was dedicated to an Arducam IMX519 camera module, which would serve as the baseline optical RGB view. The other input on that board was further split into four additional IMX519 camera modules through the use of an Arducam multiplexer board [49]. To configure the four cameras as a linear polarization camera, each of the cameras was fitted with a linear polarizing filter oriented at one of four different angles — 0°, 45°, 90°, and 135°. All five cameras were mounted onto a single acrylic plate, arranged as close as possible to reduce the perspective shift between camera views. The other StereoPi V2 board was connected to a pair of Raspberry Pi High Quality IMX477 camera modules that allowed mounting of different lenses, serving as the stereoscopic cameras. Additionally, a MPU-6050 chip containing a combined 3-axis accelerometer and 3-axis gyroscope into a 6-axis inertial measurement unit (IMU)

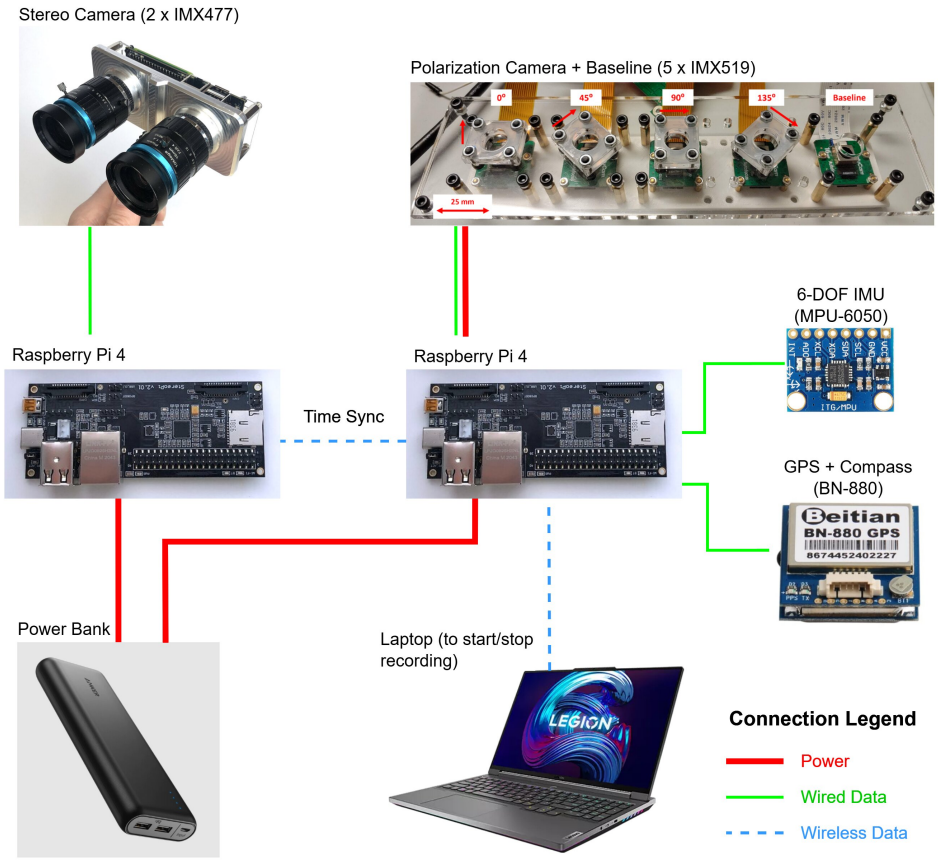


Figure 3.1: Connection diagram of the first sensor suite iteration.

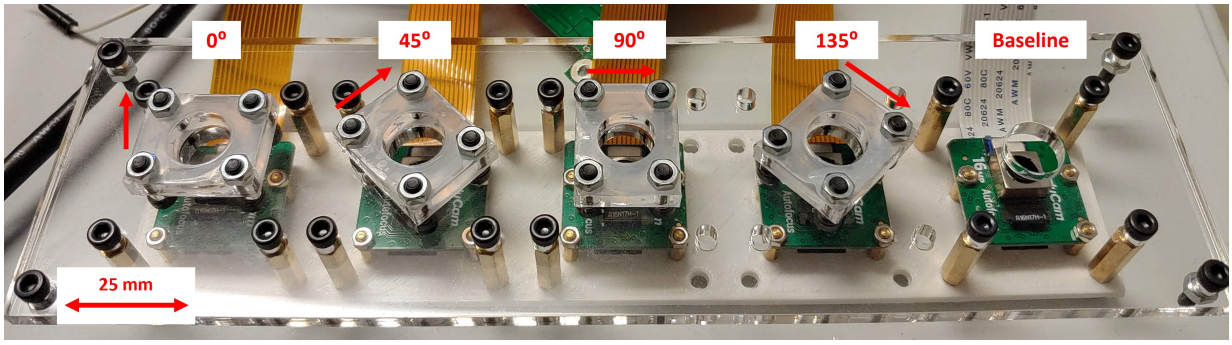


Figure 3.2: Close-up of the custom multi-camera unit with 4 linear polarization filtered cameras on the left and an unobstructed baseline camera on the right.

and a BN-880 GPS module were included. The incorporation of every sensor into a single platform of the Raspberry Pi made it simple to capture and synchronize the data with appropriate triggers and timestamps. As the Raspberry Pi can run on 5 V DC, a common power bank was used to maintain portability of the system.

Although simple and quick to integrate, the system did have a few limitations. The combination of a small sensor size and an inexpensive lens resulted in poor capture quality at a distance of beyond 50 m. Additionally, the exposed board and camera modules made it difficult to weatherproof the entire system for use outdoors, while also not impeding the view of the cameras. Thus, the system was redesigned to remove the dependence on self-contained low-level camera modules.

3.4 Second Iteration

Following the difficulty in trying to use a number of third-party components with limited support, the next iteration of the sensor system focused on incorporating sensors that are better supported and easier to integrate. The availability of a global shutter, different sensor and lens combinations, built-in weatherproofing, and the existence of a self-contained polarization camera led to the integration of industrial machine vision cameras in the system. The set of 4 polarization cameras were replaced with a single LUCID Vision Labs Triton camera that incorporated the Sony IMX264MYR colour polarization sensor. The pair of stereoscopic cameras were likewise also replaced with a pair of LUCID Vision Labs Triton cameras with the Sony IMX428 colour sensor [50]. A GoPro HERO12 Black action camera acted as the new baseline RGB camera due to its robustness in harsh environ-

ments and the high quality captures offered through its built-in image processing pipeline. Additionally, a FLIR MD625 marine-grade thermal camera was included due to the wide usage of thermal infrared (TIR) cameras in marine environments and autonomous driving. By operating in the long-wavelength infrared range of 8-14 µm, TIR cameras offer good performance in low-light conditions to supplement the poor performance of RGB optical cameras [51]. Thus, the second iteration consists of the sensors in Table 3.2.

Table 3.2: Sensors Included in the Second Iteration

Sensor Type	**Sensor Model**	**Resolution/Range**	**Sampling/ Capture Rate**
Baseline RGB Camera	1 × GoPro HERO12 Black	3840 × 2160	29.97 FPS
Stereoscopic Camera	2 × LUCID TRI071S-CC	3208 × 2200	1 FPS
Polarization Camera	1 × LUCID TRI050S1-QC	1224 × 1024	1 FPS
Thermal Camera	1 × FLIR MD625	640 × 480	29.97 FPS
IMU	3 × MPU-6050	±2000 °/s (gyro) ±16 g (accel.)	100 Hz
GPS	1 × BN-880	±2 m	1 Hz

The increase in number of IMUs to 3 is an attempt to work around the inherent noisiness of an IMU, where information from each of the individual sensors can be used to filter out noise and serve as a voting mechanism to disregard input from an unreliable sensor. However, increasing the number of IMUs while introducing a new computer-based data collection system required a redesign of the Raspberry Pi enclosure to ensure access to the sensors while also remaining weatherproof. To do so, the StereoPi V2 board and the 3 IMUs were mounted on a flatly machined block of aluminum, with carefully drilled holes to ensure that all the IMUs aligned. The entire assembly was then placed in a waterproof box rigidly mounted to the camera frame.

The machine vision cameras and the thermal camera all required power over Ethernet (PoE), which was accomplished through the use of a PoE network switch. As the network

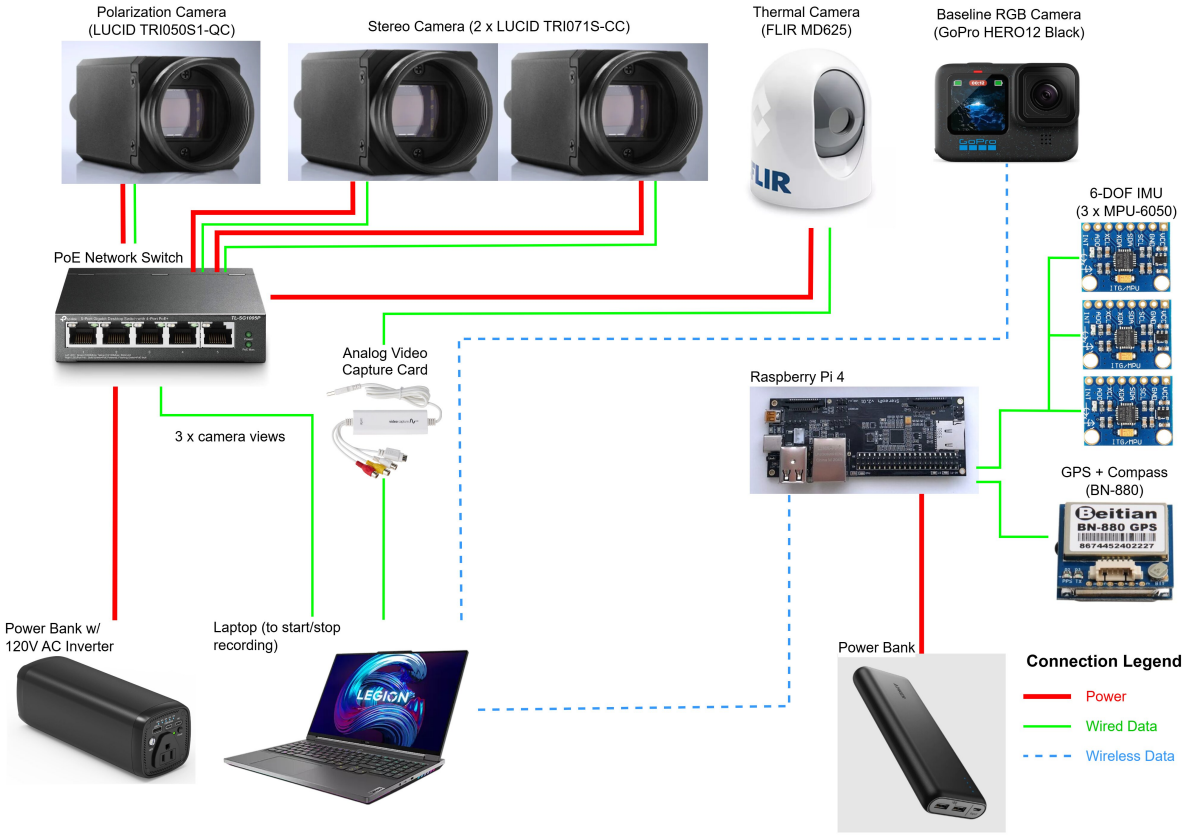


Figure 3.3: Connection diagram of the second sensor suite iteration.

switch needed external 120 V AC power to operate and forward that power to the cameras, a portable power bank with a built-in 120 V inverter was used to provide that power. Weatherproofing of the system was achieved through the built-in ingress protection offered by the respective manufacturer of the cameras. The LUCID Vision Labs cameras and lens housings are rated to IP67, whereas the FLIR MD625 is rated to IP65, both with connection cables included in the rating. Finally, the GoPro HERO12 Black includes built-in waterproofing from the manufacturer. Thus, this means that while the data collection laptop and power supply may not necessarily be rated to withstand a similar environment, only the parts that needed to be exposed (the sensors) are well rated to survive the expected environmental conditions that will be encountered. All of the sensors were mounted to a single piece of aluminum extrusion using custom-machined mounting blocks. This provided the versatility to mount the system either as-is, strapped or bolted to a stable surface, or to attach a tripod for more flexibility.

The shift to using more off-the-shelf components had the consequence of increasing complexity for integration, as now there are multiple different capture systems that needed to be synchronized together. The LUCID Vision Labs cameras sent captures digitally using GigE Vision through a local Ethernet network, connected via the network switch. The FLIR thermal camera only outputs analog video, which was passed through an Elgato Video Capture analogue-to-digital video converter. Finally, the footage from the GoPro camera was captured completely separately through the use of the GoPro Quik smartphone app due to a lack of officially available application programming interface.

However, a few issues were experienced with this iteration of sensors. First, the LUCID Vision Labs cameras and FLIR thermal camera through the analog capture device were recorded using what was believed to be a synchronized signal sent to each device. This was found to later be false, with the analog capture device serving inconsistent framerates compared to the more exact machine vision cameras. Furthermore, although the Raspberry Pi was synchronized to the laptop's time server, it still remained difficult to match up the exact movements of the system with a specific captured frame. The included GPS unit also did not provide useful information as testing was conducted indoors in the National Research Council Canada — Ocean, Coastal, and River Engineering Research Centre's (NRC-OCRE) ice tank. Thus, following testing and experimentation, a third and final iteration of the sensor suite was required.

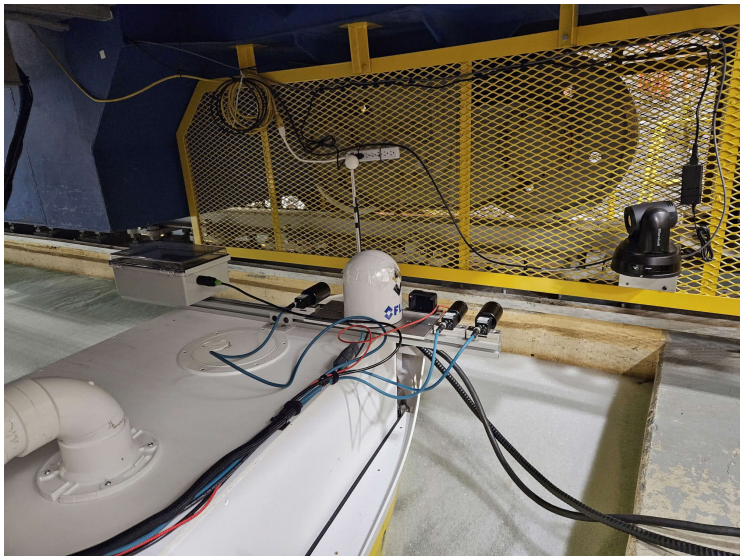


Figure 3.4: Setup of the second iteration of the sensor suite mounted on a model ship operating in the ice tank.

3.5 Final Iteration

The final iteration of the sensor suite followed a much simpler combination of sensors and capture system, prioritizing robustness and ease of deployment. One of the main changes to the system was the removal of the Rapsberry Pi components and its respective sensors entirely. Field trials outdoors and in the ice tank showed that the IMU data was noisy and difficult to correlate with image data, whereas the GPS data was unhelpful if the system did not move great distances during image capture. Additionally, the pair of stereoscopic LUCID Vision Labs cameras were replaced with two more GoPro HERO12 Black action cameras, as the image quality was deemed to be better in a wider range of environments. Thus, the final set of sensors included in the suite and their respective field of view (FOV) are included in Table 3.3. Furthermore, it was found that the baseline distance between the stereoscopic cameras in the second iteration was too large for the expected effective range of the system. Although larger baselines offer more precision in depth measurements for stereoscopic camera setups, the overlapping field of view of the cameras also decrease, resulting in some closer objects being out of frame. Thus, the camera mounting rig was also redesigned following the changes to the sensor lineup, and the cameras were moved closer together to further reduce the perspective shift between each view.

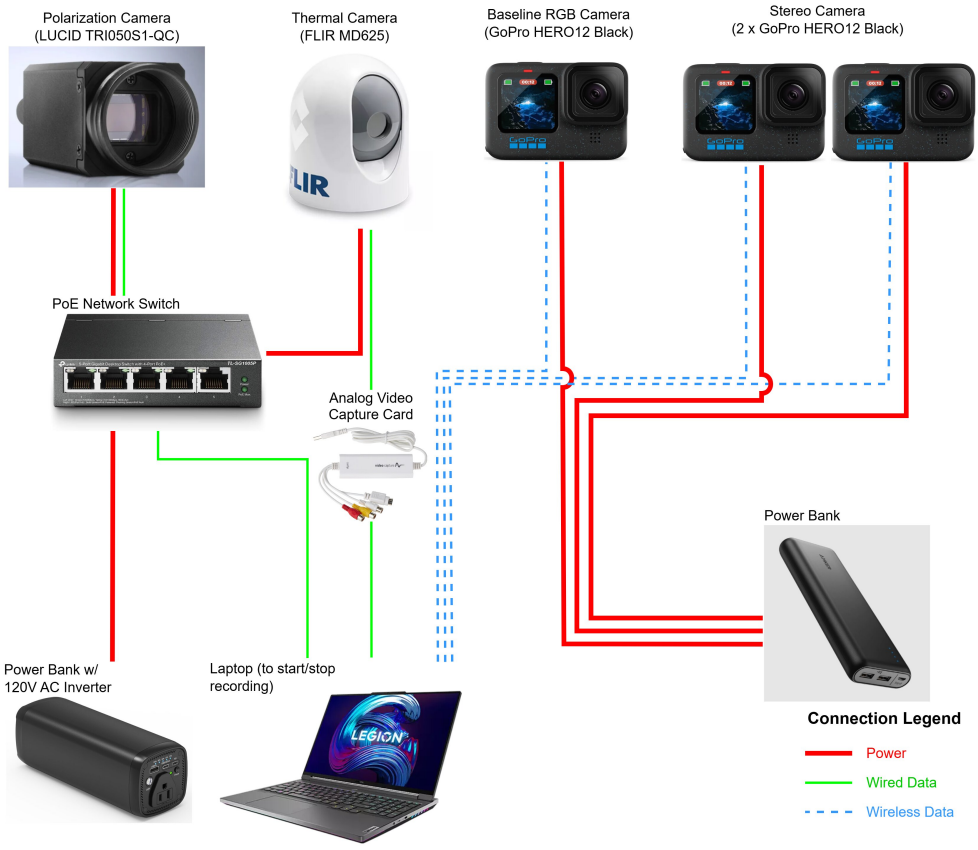


Figure 3.5: Connection diagram of the final sensor suite iteration.

Table 3.3: Sensors Included in the Final Iteration

Sensor Type	**Sensor Model**	**Resolution/FOV**	**Sampling/ Capture Rate**
Baseline RGB Camera	1 × GoPro HERO12 Black	3840 × 2160 92°× 61°	29.97 FPS
Stereoscopic Camera	2 × GoPro HERO12 Black	3840 × 2160 92°× 61°	29.97 FPS
Polarization Camera	1 × LUCID TRI050S1-QC	1224 × 1024 19°× 16°	1 FPS
Thermal Camera	1 × FLIR MD625	640 × 480 25°× 20°	29.97 FPS



Figure 3.6: Setup of the final sensor suite.

3.6 Sensor Calibration and Synchronization

As with all applications involving computer vision, camera calibration is an extremely important aspect to ensure that the objects in view are not being distorted by the optical properties of the sensor and lens [52–54]. This is especially critical with a mix of different sensor sizes and lenses to ensure that all views are brought to a common viewpoint [30, 43, 51, 55]. Camera calibration solves this problem by finding the relation between points in 3D space and their projected 2D points on the camera's image plane.

An ideal pinhole camera can be modelled as a transformation that maps a point in 3D space to a location on the flat image plane. A point in the camera coordinate system is denoted by $\mathbf{x} = [x, y, z]^T$ and its corresponding point in 3D space is denoted by $\mathbf{X} = [X, Y, Z]^T$. Since the image plane is flat, the z coordinate can be set arbitrarily by making the image plane coincident to the x-y plane in the camera coordinate system. Thus, each pixel in an image has a location of (x, y), and the transformation from the world coordinate system to the camera coordinate system can be represented with the augmented vectors $\tilde{\mathbf{x}} = [x, y, 1]^T$ and $\tilde{\mathbf{X}} = [X, Y, Z, 1]^T$ as:

$$w\tilde{\mathbf{x}} = \mathbf{K}[\mathbf{R} \quad \mathbf{t}]\tilde{\mathbf{X}}, \quad \text{with } \mathbf{K} = \begin{bmatrix} f_x & s & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3.1}$$

where w is an arbitrary scale factor, $[\mathbf{R} \quad \mathbf{t}]$ includes the extrinsic camera parameters that relate the rotation and translation from the real-world coordinate system to the camera coordinate system, and $\mathbf{K}$ includes the intrinsic parameters which describe the physical properties of the camera and lens. The intrinsic parameters are (f_x, f_y) as the focal lengths in pixels, (u_0, v_0) as the coordinates of the optical center (principle point) in pixels, and s as the skew coefficient that compensates for non-perpendicular image axes, that map the camera coordinates to the image plane. With 6 degrees of freedom for extrinsic parameters (3 for rotation and 3 for translation) and 5 degrees of freedom for the intrinsic parameters, a minimum of 11 equations are required to solve for camera calibration.

However, this model assumes an ideal pinhole camera that can project points onto the image plane in perfectly straight lines, whereas real camera assemblies include lenses which introduce additional distortion throughout the image, warping the objects projected onto the image plane. In most cases, this can be modelled as a 2-coefficient radial distortion. First, the image coordinates must be normalized as $\hat{\mathbf{x}} = [\hat{u}, \hat{v}]^T$, which are found through translating the pixel locations to be relative to the optical centre and dividing by its

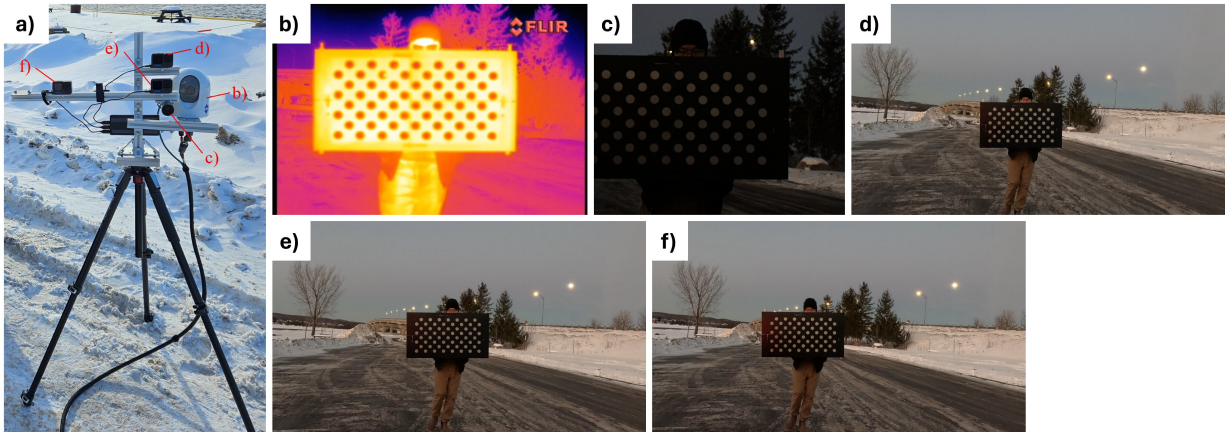


Figure 3.7: (a) Full sensor system with each camera labelled. Camera calibration was conducted using a rigid asymmetric circle grid board, with an example frame shown from each of the synchronized views of (b) the thermal camera, (c) the polarization camera, (d) the baseline RGB camera, (e) and (f) the left and right pair of RGB stereoscopic cameras.

respective focal lengths in pixels (f_x, f_y) as follows:

$$\hat{u} = \frac{x - u_0}{f_x} \tag{3.2}$$

$$\hat{v} = \frac{y - v_0}{f_y} \tag{3.3}$$

where (x, y) is the original location of the pixel as captured by the camera. The radial distortion is then modelled as:

$$\hat{u} = u(1 + k_1 r^2 + k_2 r^4) \tag{3.4}$$

$$\hat{v} = v(1 + k_1 r^2 + k_2 r^4) \tag{3.5}$$

where (u, v) is the theoretical undistorted location of the pixel, k_1, k_2 are radial distortion coefficients, and $r^2 = u^2 + v^2$ is the radial distance from the optical centre. For a single camera system with a camera-centred coordinate system, the goal of calibration is to find the unknown intrinsic parameters in $\mathbf{K}$ and radial distortion coefficients k_1, k_2 to map from the captured image coordinates (x, y) to the undistorted coordinates (u, v). However, a multi-camera system also requires finding the extrinsic camera parameters $[\mathbf{R} \quad \mathbf{t}]$ for each individual camera to warp the captured image points to a common coordinate system.

Including the 11 degrees of freedom from the pinhole camera model and the 2 radial distortion coefficients, this gives a minimum of 13 equations required to solve for the

unknown parameters in the camera and lens models. To find these camera parameters, a known planar pattern is typically used to locate known points in 3D space and refer them to their respective pixel locations.

Usually, camera calibration is conducted through the use of a printed known planar pattern (most often a checkerboard) where the dimensions of the features are known. This pattern is then shown in a variety of locations and orientations throughout the entire field of view of the camera. With this, it is possible to map how the pattern changes as it moves in front of the camera. However, the inclusion of a thermal camera requires more than just a printed pattern, as thermal cameras cannot differentiate between visible colours. Thus, a custom calibration board had to be developed that can be seen by both the optical and thermal cameras at the same time to ensure proper calibration and synchronization between the views.

The solution to the printed pattern problem was to instead design a composite calibration board made up of materials with contrasting properties in the thermal infrared (TIR) range that are also visibly different. The two materials used were clean aluminum with a high reflectivity in TIR and has a gray-ish colour, and black acrylic which mostly absorbs TIR. Furthermore, rather than using a checkerboard pattern, an asymmetrical circle pattern was used. This was due to the fact that a checkerboard pattern requires exact focus to ensure that the intersection points at the corners of each checkerboard was sharp, which is difficult to do considering the thermal camera does not offer a way of adjusting the focus of the lens, in addition to the fact that adjusting the focus of the lens specifically for the calibration and then re-adjusting it later to look further away defeats the purpose of camera calibration. Thus, a circle pattern allows for blurry edges while still accurately locating the centres of each circle.

Rather than trying to time the captures with each other for a synchronized frame, the relatively high capture rate of the cameras were exploited using the pigeonhole principle. Seeing as how it was close to impossible to synchronize the thermal camera's feed through the analog capture device due to inherent delays and inconsistencies in the analog to digital conversion process, the synchronization of different sensors were instead left to the data processing step. With this, all cameras except the LUCID polarization camera were set to capture at a higher framerate of around 30 frames per second (FPS), with the LUCID camera capturing at a constant 1 FPS. This ensures an upper bound of 17 ms to the temporal difference Δt_{frame} between the sensors for each capture, calculated as:

$$\Delta t_{frame} = \frac{\text{FPS}^{-1}}{2} \tag{3.6}$$

For synchronization, a clapping motion was conducted in view of each camera at the start

of capture, similar to how a clapper board is used in forms of video media. The frames were then extracted and aligned manually to find the time offset in frames to match each camera up.

Chapter 4

Dataset Creation

4.1 Data Collection

Initially, field tests were conducted with the second iteration of the sensor suite. Outdoor tests were conducted in March 2024 of lake ice on Lake Huron near Sudbury, ON at the coordinates 46°01'00.1"N, 81°47'08.7"W. However, as this was at the end of a warm winter, the data collected was only composed of semi-melted stationary ice. Additional data was then collected indoors in June 2024 mounted on a model ship travelling through simulated sea ice sheets using the indoor ice tank at NRC-OCRE. These preliminary trials served as a chance to test the sensors and data collection system in a field environment, and provided valuable information to enhance the sensor suite for the final iteration.

While the system mostly worked as expected, there were a few issues found with the system. The main issue noticed was the relatively poor capture quality of the machine vision cameras compared to that of even a smartphone camera, leading to blurry and non-ideal images. Additionally, the IMU and GPS combination from the Raspberry Pi did not provide much useful information, having a low signal-to-noise ratio even with the multiple IMU setup due to the relatively small movements experienced. Regardless, the outdoor data showed the capabilities of the polarization camera in differentiating between water and ice, even with the presence of glare, whereas the indoor ice tank data served as a good testbed for experimenting with ways to register each view to that of the baseline.

Following a redesign of the system to the final iteration, land-based data of river ice was collected in February 2025 at various locations along the Ottawa and St. Lawrence rivers, as shown in Figure 4.1. This data was collected in lieu of sea ice due to difficulties

in securing opportunities to conduct sea-faring trails. However, river ice composition and dynamics are different compared to sea ice [56, 57], especially since sections of the St. Lawrence Seaway are managed waterways as opposed to free-flowing water [58]. Events such as the rapid formation and breakup of ice jams and hanging dams in rivers can result in dramatic changes in ice types and flow dynamics [59], which are typically not seen in larger bodies of water. The shift from sea ice to land-based river ice data also results in different types and scale of the ice encountered, as sea ice includes large floes, older multiyear ice, and icebergs that are not seen in rivers. Additionally, ships usually aim for areas of lower ice concentration to reduce risk, thus data collected on board will likely see more open water compared to data collected from land. Nonetheless, these field studies provided an opportunity to test the sensor suite on similar-looking data in a variety of environmental conditions. Time spent recording at each location ranged from 5-30 minutes, providing hours of raw video footage. Data was collected across multiple days at a total of six locations, referred to hereafter as scenes. Camera calibration was conducted at the start and end of the whole trip with consistent results, ensuring that the camera parameters did not change throughout the acquisition process.

The different views were first synchronized by manually matching up the synchronizing action across each recording frame-by-frame, which was then extrapolated throughout the remainder of the recording. As 1 Hz is still fairly quick for the relatively slow-moving ice, every 5^{th} (for fast-flowing ice scenes) or 10^{th} (for slower-flowing ice) frame was extracted to further reduce the dataset size. Once the synchronized data was reduced to a manageable amount, the subset of data first underwent a manual review to remove all captures that were either out of synchronization or of poor quality.

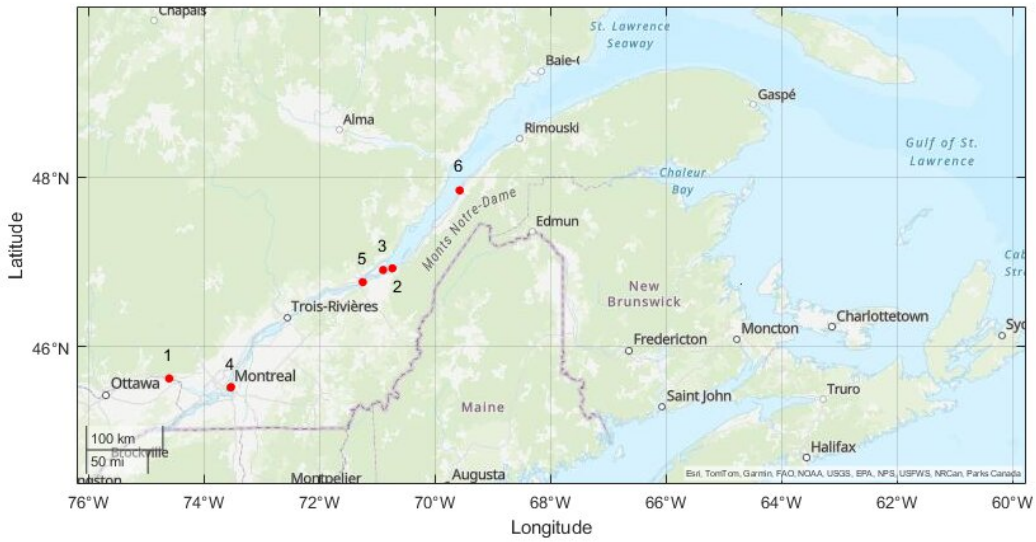


Figure 4.1: Map of the data collection locations in red for each scene, numbered according to the scenes listed in Table 4.1.

Table 4.1: Data Collection Location Details

Scene	**Date/ Time***	**Location**	**Length of Capture**	**Lighting/ Weather Conditions**	**Notes**
1	2025-02-11 16:30	Hawkesbury, ON 45°37'05.1"N 74°35'54.1"W	20 min.	Good/Clear	Stationary ice, tested different camera angles looking down from bridge
2	2025-02-12 14:00	Berthier-sur-Mer, QC 46°56'06.8"N 70°44'08.8"W	30 min.	Good/Clear	Relatively fast-flowing brash ice
3	2025-02-12 18:15	Saint-Jean-de-l'Île-d'Orléans, QC 46°54'56.0"N 70°53'46.8"W	5 min.	Poor/Clear	After sunset, fast-flowing brash ice difficult to see in most views
4	2025-02-13 07:15	Montréal, QC 45°30'46.7"N 73°31'43.0"W	5 min.	Medium/Snow	At dawn with overcast sky and heavy snowfall, mostly small clumps of thin ice
5	2025-02-15 07:00	Québec, QC 46°46'13.7"N 71°14'54.3"W	25 min.	Poor/Clear	At sunrise with prominent reflection of the sun, relatively fast-moving brash ice
6	2025-02-15 15:15	Rivière-du-Loup, QC 47°50'46.4"N 69°34'18.0"W	10 min.	Good/Clear	Large swaths of stationary ice in foreground with fast-moving ice about 400 m from shore

*All times are in Eastern Standard Time.

4.2 Image Pre-Processing

To prepare the dataset, the videos recorded from each camera were initially manually synchronized by temporally aligning the synchronization action across each view. Given the known framerate of each video, the synchronized frame times can then be propagated throughout the entire video for that scene. The synchronized frames were then extracted at 1 frame per second, matching the framerate of the polarization camera. The dataset size was further reduced by taking every 5^{th} or 10^{th} frames depending on the scene, as there was little change from frame-to-frame due to the relatively slow-moving ice. Each frame was then undistorted according to the calibration parameters found using the Computer Vision Toolbox in Matlab [60].

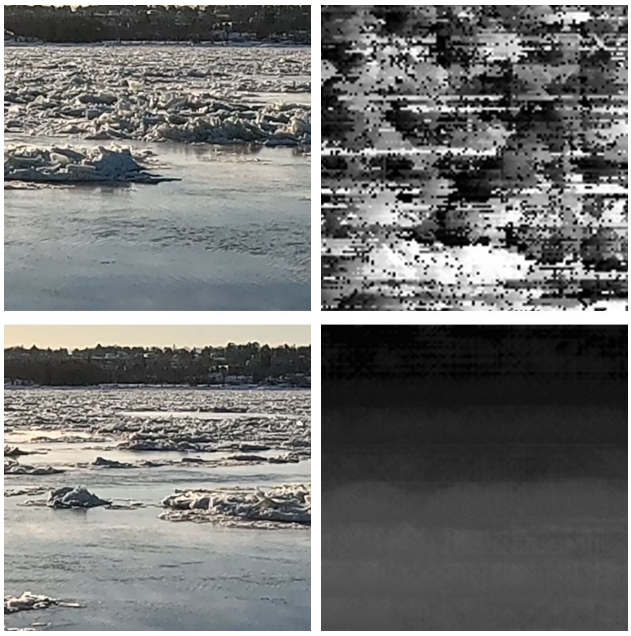
For the stereoscopic cameras, the captured images were first undistorted with the intrinsic parameters $\mathbf{K}$ and radial lens distortion coefficients k_1, k_2. As the left camera was set as the fixed view and the right camera as the moving view, the extrinsic parameters $[\mathbf{R} \quad \mathbf{t}]$ were used to apply a projective transform to the right camera to rectify the two views. Rectification applies the epipolar geometric constraint on the two views, where any given point (the epipole) in one view can be found at the same horizontal level (the epipolar line) in the other view. By matching similar features and calculating the horizontal displacement in pixel locations (also known as the disparity) between the two views, a disparity map can be created. Objects closer in view have higher disparity as they take up a larger portion of the camera's FOV, and thus the perspective shift is more obvious, whereas objects further have lower disparity for the opposite reason. By combining the disparity values with intrinsic properties from the camera, the depth Z of a point in the scene can be directly calculated as follows:

$$Z = \frac{f_x b}{u_L - u_R} \tag{4.1}$$

where u_L, u_R are the horizontal pixel locations in the left and right view respectively, and b is the physical baseline distance between the optical centres of the two cameras [61].

Repeating this process for every pixel in the view generates a depth map, which can then be used as an additional source to rule out discrepancies, such as objects that appear below the ground surface after reconstruction due to reflections from bodies water [40]. However, accurately matching features between the two views remains an outstanding problem in dense stereo, with traditional algorithms reliant on global or local matching still struggling with textureless areas [62]. As finding the disparity between two views requires a single unique match for each pixel, similar-looking areas at different locations throughout the image can cause large jumps and inconsistencies in the generated disparity

map [63]. Since most of the scenes contain a large portion of relatively textureless ice or water scattered throughout the image, this led to highly inconsistent depth maps that would change drastically from frame to frame, as seen in Figure 4.2. Thus, the depth map was not included as an input for most of the datasets.



(a) Baseline RGB (b) Disparity Map

Figure 4.2: Example of inconsistency in the generated disparity maps of consecutive frames.

The polarization camera required some additional processing to obtain linear polarization information. Using images captured at each of the four polarization angles (0°, 45°, 90°, 135°), three Stokes parameters can be calculated as follows:

$$S_0 = I_0 + I_{90} \tag{4.2}$$
$$S_1 = I_0 - I_{90} \tag{4.3}$$
$$S_2 = I_{45} - I_{135} \tag{4.4}$$

where $I_0, I_{45}, I_{90}, I_{135}$ are the pixel intensities at each of the four polarization angles, and S_0, S_1, S_2 are the three Stokes parameters that describe the linear polarization state of that pixel. Thus, the degree of linear polarization (DoLP) and angle of linear polarization

(AoLP) can be calculated as follows:

$$p = \frac{\sqrt{S_1^2 + S_2^2}}{S_0} \tag{4.5}$$

$$2\psi = \tan^{-1}\left(\frac{S_2}{S_1}\right) \tag{4.6}$$

where p is the DoLP and ψ is the AoLP in radians. The resultant polarization images contain a significant amount of salt-and-pepper noise due to the sensitivity in measuring polarizing light, so a median filter was applied to remove some of the noise. Additionally, histogram equalization was applied to each frame to enhance contrast and make details more visible, at the expense of losing the ability to extract physical meaning from the intensity values.

For the temperature of an object to be accurately determined based on its thermal radiation, variables such as ambient air conditions and target emissivity must be known or estimated [64]. However, with unknown objects in view, the ability of a thermal camera to accurately measure temperature is limited. While the absolute temperatures of objects cannot be determined, relative temperature differences, especially between water and ice, can still be used as an additional source of information. The images were captured with the white-hot option, which is a greyscale mode where higher pixel intensities correspond to higher temperatures. Due to the focus on commercial marine use cases forcing automatic exposure adjustment, the temperature range was continuously re-scaled during capturing. To ensure consistency from frame to frame, the thermal images were normalized by using control points in the interface between the water and ice to be a fixed intensity, assuming that the true temperature in those areas are close to the freezing point of water.

Since the thermal camera outputs analogue video, some random noise is expected in the captured images. It was assumed that the noise followed a normal distribution, and therefore a Gaussian filter was applied to denoise and smooth the images out. Although there were still some minor visual artifacts from the analogue to digital conversion, applying a discrete Fourier transform did not isolate any dominant frequencies to filter out. Thus, the artifacts were considered to be statistically insignificant. Additionally, some vignetting was observed near the corners of the image, likely caused by uneven heat distribution as the camera itself warms up during operation. However, the images were later scaled and cropped to only include a section of the centre, which removed the need to compensate for this effect.

Due to the physical distances between the cameras, there is a small shift in perspective between each of the captured views. Corresponding control points were manually selected

for each scene to calculate a unique projective transform for each camera to warp the view to that of the baseline camera. Finally, the frames were cropped and scaled to a resolution of 512×512, containing only those pixels which are present across all cameras after warping.

4.3 Ground Truth Annotation

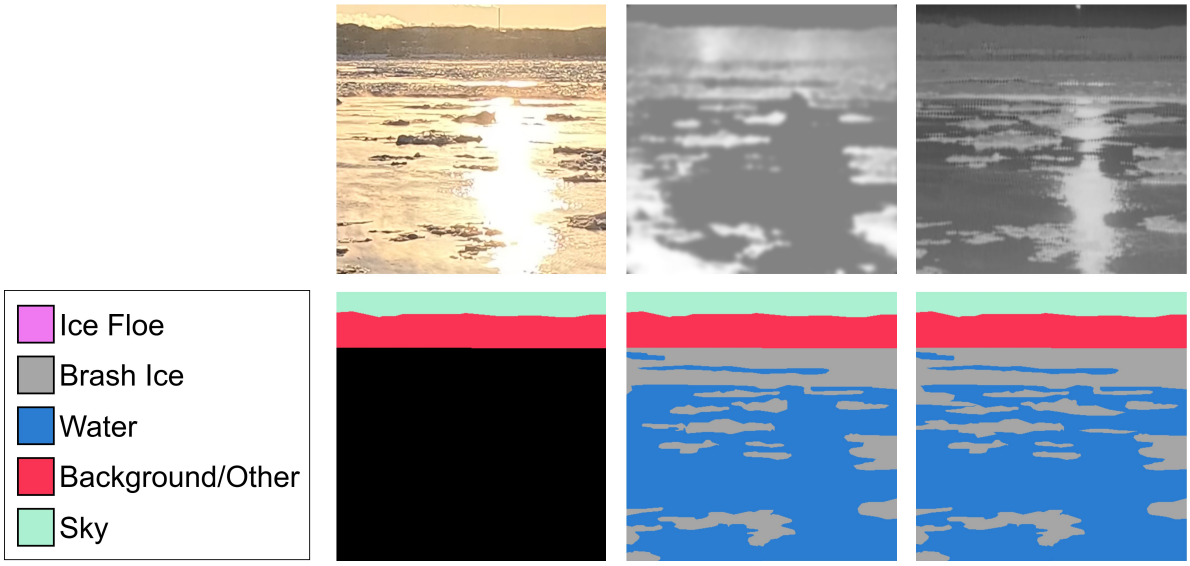
In total, there were over 600 synchronized and cropped frames across all scenes. Of those, every 5^{th} frame was picked out to be manually labelled, giving a total of 133 fully labelled images across the six scenes. Combined with the earlier dataset reduction methods, this resulted in a maximum capture period of 50 s between subsequent frames at a certain scene. Using the horizontal angle of view (HAOV) of the polarization camera from Table 3.3 as it is the most narrow out of all of the cameras, the sensitivity to the speed at which a piece of ice passes across the sensor can be calculated. With simple geometry, the horizontal field of view (HFOV) of the camera can be calculated as:

$$\text{HFOV} = 2 \times d \times \tan\left(\frac{\text{HAOV}}{2}\right) \tag{4.7}$$

where d is the distance from the camera to the target object. At a distance 20 m which is about how far away the camera was from the flowing water, Equation 4.7 gives a HFOV of 6.7 m. Using an average ice velocity of 0.1 m/s [65], a piece of ice will travel 5.0 m in front of the camera's HFOV in the 50 s between subsequent images, which is about 3/4 of the way across the sensor. If each frame was used instead with a capture period of 10 s, the same piece of ice will only travel 0.5 m, less than 1/10 of the way across the sensor. With floes further away, the distance they move between subsequent images is even less. Thus, the further reduction of the dataset helped to reduce the amount of similar-looking data and repeated objects in the dataset.

The dataset was manually labelled from the cropped view of the baseline camera using the Computer Vision Annotation Tool (CVAT) online [66]. Large, relatively simple features such as the sky and background were manually labelled and propagated with small changes throughout the appropriate frames. More complex features, such as small pieces of brash ice, were labelled semi-automatically using the built-in interface to the Segment Anything Model (SAM) introduced by Kirillov et al. [67] with given prompts. The label classes were based on sea ice nomenclature as described by the World Meteorological Organization [27] and adapted to river ice. The following five classes used in labelling were defined as such:

- **Ice Floe** – Large, flat pieces of floating ice with relatively smooth texture
- **Brash Ice** – Congregations of many different small fragments of floating ice, with rough and jagged texture
- **Water** – Open, ice-free water
- **Background/Other** – Items in frame that are not ice or water, such as trees or hills in the background
- **Sky** – Areas above the horizon that are not obstructed by other objects or the background



(a) RGB (b) DoLP (c) Thermal

Figure 4.3: Progressive labelling of the dataset by incorporating information from each sensor to ensure that details do not get lost.

In general, the sky and background/other classes took up the top third of the frame, with the rest of the frame being filled with a combination of the remaining classes. As it is hard to distinguish between ice floes and brash ice in the far field due to the small fragments of ice blending together, the ice in the far field is more likely to be labelled as floes rather than as brash ice.

4.4 Dataset Exploration

While all six scenes were labelled, the night-time condition scene 3 remained unused due to a lack of confidence in accurate data labels as a result of low visibility conditions. Thus, this left a total of 118 images to be used for training and testing. From this, a close to 70-15-15 training-validation-test split was used, with 83, 16, and 19 images included for the training, validation, and test sets respectively. Due to the different locations and ice conditions observed, the distribution of class labels differed significantly from scene to scene. As such, at least one image from each scene was included in the validation and test sets to ensure full representation of the data. The relative representation from each scene in the different sets are shown in Table 4.2.

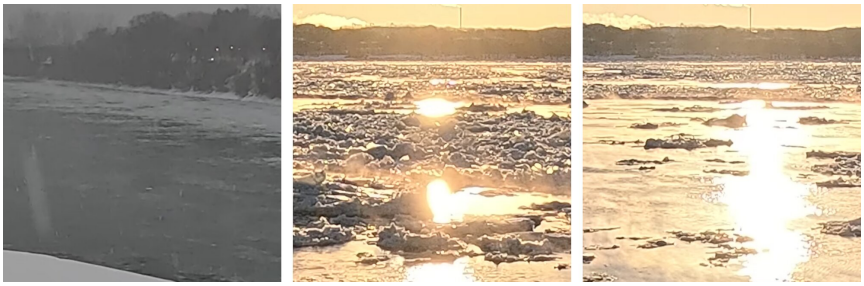
Table 4.2: Scene Image Distribution

Scene	**Training Set**	**Validation Set**	**Test Set**
1 (Easy)	7 (8.4%)	2 (12.5%)	6 (31.6%)
2 (Easy)	31 (37.3%)	7 (43.8%)	5 (26.3%)
4 (Hard)	11 (13.3%)	1 (6.3%)	2 (10.5%)
5 (Hard)	17 (20.5%)	4 (25.0%)	4 (21.1%)
6 (Easy)	17 (20.5%)	2 (12.5%)	2 (10.5%)
Total	83	16	19

Furthermore, the test set was split into an easy and hard subset depending on the lighting and environmental conditions witnessed in the scene as recorded in Table 4.1. The easy subset (scenes 1, 2, and 6) consists of images acquired in close to ideal conditions, with clear skies in the afternoon providing good lighting without glare and little to no physical obstructions. These images were hypothesized to perform well even with only a single view from the baseline camera, due to the rich colour and texture information available in the visible spectrum. On the other hand, the hard subset (scenes 4 and 5) consists of images taken in adverse conditions near dawn, resulting in non-ideal lighting conditions before the sun rises or with significant sun glare. These images were hypothesized to be more challenging for the baseline camera to accurately segment, either due low-light and low-contrast conditions, or due to the added obstructions from falling snow and sun glare



(a) Easy Subset

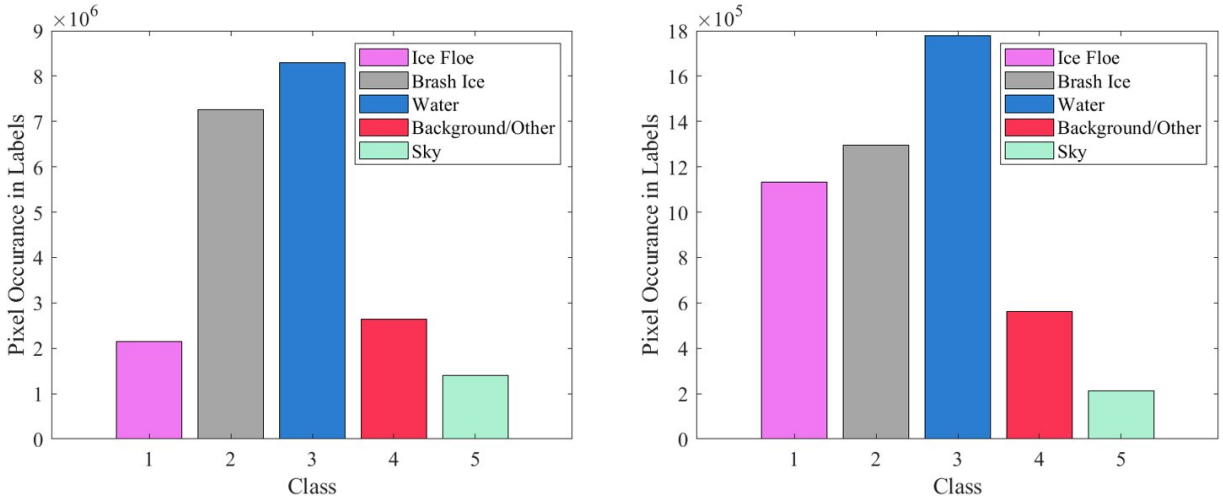


(b) Hard Subset

Figure 4.4: Example RGB images from the easy and hard subsets.

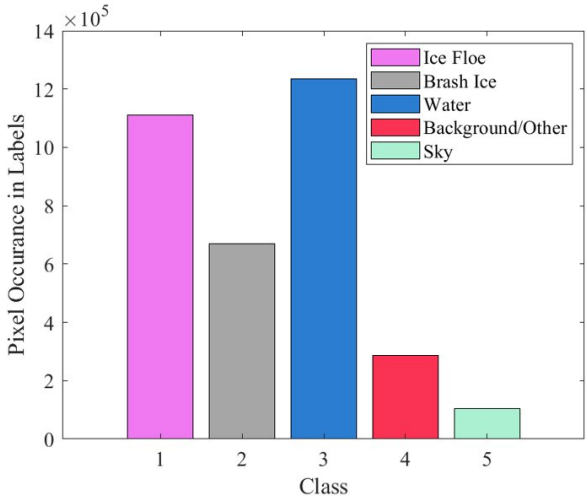
blocking sections of the image. The differences in conditions between the two subsets can be seen in Figure 4.4. Following the split, the original 19 test images were separated into 13 images in the easy subset and 6 images in the hard subset. The class label distributions for the full test set as well as the easy and hard subsets are shown in Figure 4.5.

From the histograms, it can be seen that there are large changes in class distributions between the different conditions. This causes the easy subset to be more biased towards ice floes, whereas the hard subset is more biased towards brash ice. Unlike the test sets, the dataset used for training was not partitioned based on scenes, and thus models were trained with full knowledge of each scene. The entire training dataset has a roughly 2:1 ratio of easy to hard scenes, which artificially places more importance on performance in ideal conditions during the training process. The variability in conditions highlights the importance of finding a model that is robust yet adaptable to do well in both ideal and adverse conditions for real-world applications.

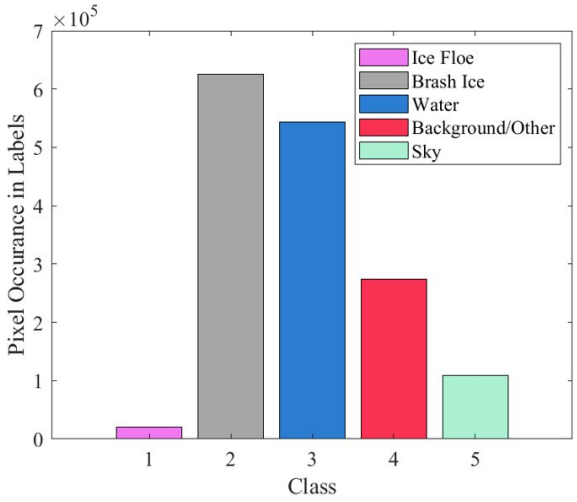


(a) Training Set

(b) Full Test Set



(c) Test Subset — Easy



(d) Test Subset — Hard

Figure 4.5: Histograms of class distributions in (a) the training set, (b) the full test set, (c) the easy test subset, and (d) the hard test subset.

Chapter 5

Experimental Setup

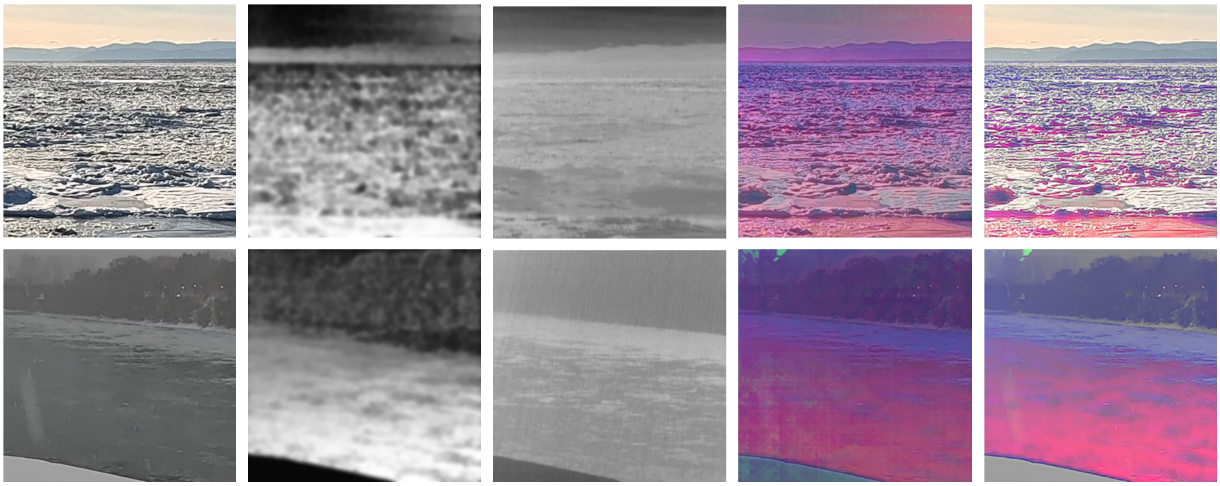
5.1 Fusion Architecture and Datasets

Multiple datasets were created from the available data using various combinations of sensor inputs to test the performance of different fusion methods using additional inputs against the baseline single RGB view. The methods were split into three main categories based on where the fusion occurred in the data processing pipeline. Early fusion methods combined different sources at the input data level, creating a new artificial data format for the network. Middle fusion methods combined the features of different sources to generate higher dimensional feature maps. Finally, late fusion methods had separate branches for each source, only combining outputs at the decision level.

5.1.1 Early Fusion

Initial early fusion methods involved pre-fusion of the inputs prior to introducing them to the network. Classical image processing techniques were used to blend together multiple input channels into a single bespoke 3-channel image in false colour, similar to how multi-spectral images are sometimes processed to be more easily interpreted [68]. This was done to leverage the pre-trained ResNet backbone, which expects a typical 3-channel RGB image as the input, without modification. Although the colour information at each pixel may not be relevant due to the unique false colour construction, the preserved spatial information may help with general feature and object recognition.

All four input sources were used at this stage, offering a total of 6 input channels: baseline (3-channel RGB), thermal (1-channel greyscale), DoLP (1-channel greyscale), and



(a) Baseline RGB (b) DoLP (c) Thermal (d) Alpha Blending (e) Wavelet

Figure 5.1: Pairs of example inputs from each sensor and the two false-colour images from the easy (top row) and hard (bottom row) test subsets.

depth (1-channel greyscale). Two methods were used to blend together the different inputs, with each of the original red, green, and blue channels from the baseline view being blended with the DoLP, depth, and thermal views respectively, effectively providing each channel with two different inputs. The first method (dubbed `alpha-blending`) used alpha blending at a fixed 50% transparency set for each of the channels. This effectively halved the bandwidth from each input before adding the two inputs together as a single channel. Examples of 3-channel images generated by `alpha-blending` are shown in Figure 5.1d.

The other method (dubbed `wavelet`) used the discrete 2D wavelet transform [69] to first deconstruct each channel into different wavelet components and their respective coefficients. The discrete wavelet transform is thus similar to a discrete Fourier transform regarding its usage of different frequencies to represent signals. However, it has the added advantage of also including spatial information in addition to frequency, allowing for knowledge of specific frequencies at a certain location. Dominant features in an image can be obtained through comparing the coefficients for each wavelet, with larger coefficients indicating a dominant wavelet, and thus a dominant feature. Images from different sources can be combined through using this property. By comparing and keeping only the largest coefficients for each wavelet, this ensures that only the most dominant features from each source is kept in the final combined image. An inverse wavelet transform is then performed to transform the selected wavelets back to the channel. This was then repeated for each

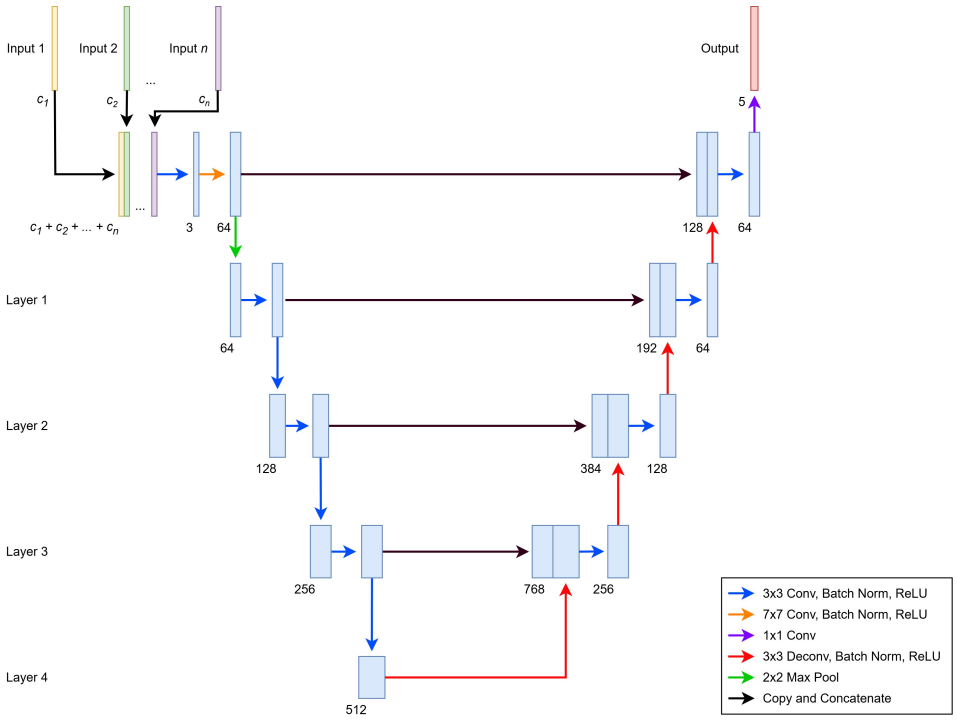


Figure 5.2: Custom U-Net architecture for early fusion with a ResNet encoder backbone. Note that the initial $c_1 + c_2 + ... + c_n \rightarrow 3$ convolutional layer was not included for the baseline RGB case, which only has 3 channels.

of the three channels with their respective inputs as was done with `alpha-blending` to create a wavelet-fused false-colour image. Examples of sensor inputs and of the false-colour 3-channel images generated by `wavelet` are shown in Figure 5.1e.

However, this introduced the problem of a loss of information as the additional sources are added, due to the less dominant features from one input being overwritten by those from the other input in each channel. Thus, to allow for the inclusion of additional sources without a loss of information, the sources were simply concatenated along the channel dimension to create a pseudo-multi-spectral image, testing different combinations of the views. This resulted in the creation of three additional early fusion datasets — `concat-dolp` (4-channel), `concat-thermal` (4-channel), and `concat-dolp-thermal` (5 channel) indicating the addition of DoLP, thermal, and DoLP + thermal channels to the baseline 3-channel RGB image respectively. The addition of extra channels required a slight modification to the stock ResNet encoder, as the there are now more than the expected 3 input channels. To solve this, a simple convolutional layer was added to compress the numerous channels back down to 3 channels before feeding it into the rest of the encoder, as seen in Figure 5.2.

5.1.2 Middle Fusion

Similar to the early fusion concatenation cases, different combinations of the views were tested using middle fusion. The three middle fusion datasets were thus dubbed `midfusion-dolp` (RGB + DoLP), `midfusion-thermal` (RGB + thermal), and `midfusion-dolp-thermal` (RGB + DoLP + thermal). Since middle fusion requires the individual inputs to undergo some level of feature extraction on their own first, the inputs had to remain separated until the fusion step. This was accomplished through the inclusion of separate ResNet encoders for each of the inputs during the feature extraction step, as seen in Figure 5.3. The features were then concatenated after the final encoding layer before being passed to a unified decoder for upsampling, with skip connections including a copy from each encoder branch.

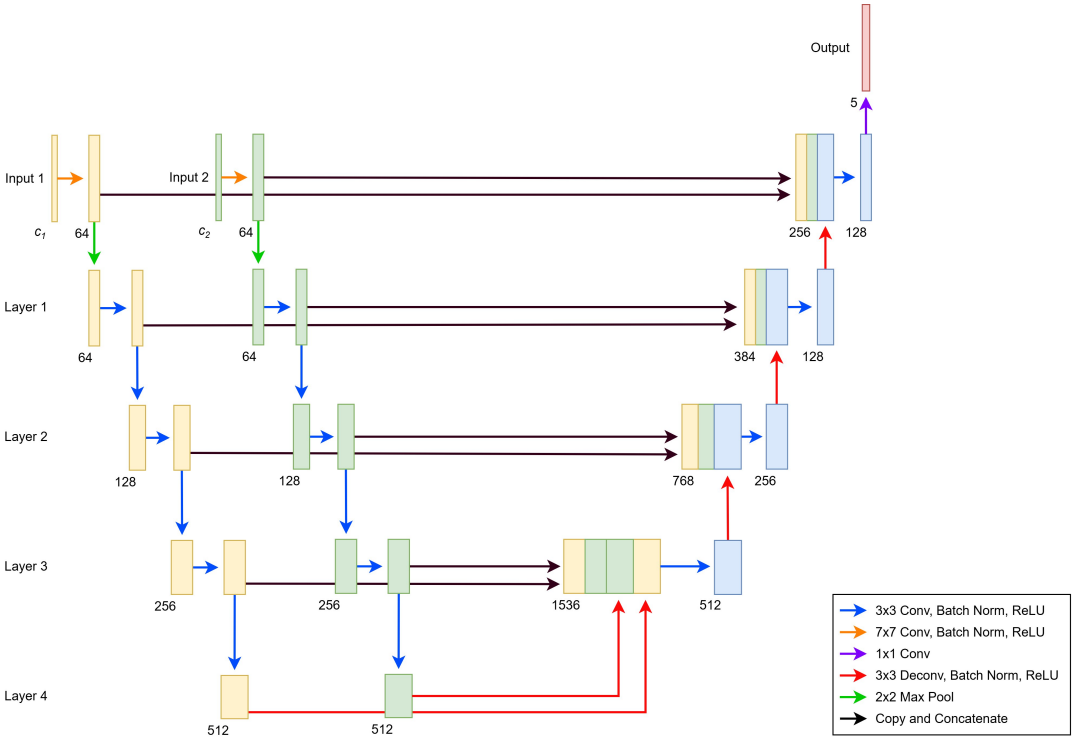


Figure 5.3: Custom U-Net architecture for middle fusion with a ResNet encoder backbone. Note that although only 2 inputs are shown in this example, additional inputs can be included through adding on additional encoder branches with their respective skip connections.

5.1.3 Late Fusion

In comparison, late fusion kept the different views completely separated until the very end of the network, as seen in Figure 5.4. Each input was passed through its own dedicated U-Net branch, allowing for full feature extraction and decoding within the individual branches before the outputs were concatenated together. The fused outputs were then passed through a final convolutional layer to generate the prediction. As with the earlier fusion experiments, different combinations of views were tested, and the resulting datasets were named `latefusion-dolp` (RGB + DoLP), `latefusion-thermal` (RGB + thermal), and `latefusion-dolp-thermal` (RGB + DoLP + thermal).

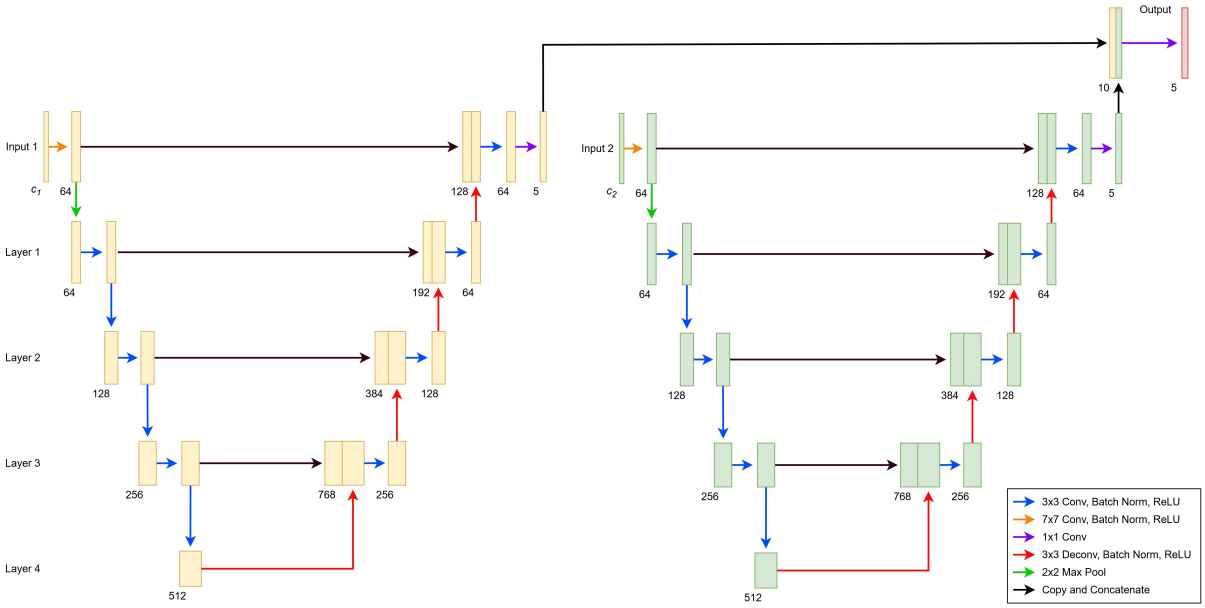


Figure 5.4: Custom U-Net architecture for late fusion with a ResNet encoder backbone. Note that although only 2 inputs are shown in this example, additional inputs can be included through adding on additional U-Net branches.

5.2 Training Details

Due to the small size of the datasets, data augmentation was employed to increase the variety of training data encountered by the model. The following transformations were applied to the training data before being passed to the model:

- Random cropping from the original 512×512 to 256×256
- Random rotation within the range of $[-20°, 20°]$
- Random colour jitter of the brightness and saturation [70], both within the range of $[0.75, 1.25]$

These specific transformations were chosen as they were representative of the different conditions that might be faced when collecting outdoor data. The random rotation simulates movement of the sensor system, whereas the random colour jitter simulates variable lighting conditions.

All models were trained for a maximum of 400 epochs with a batch size of 8. Training was stopped early if the validation loss did not improve for 40 epochs, saving the best performing model up to that point for evaluation. All training conducted in this study was performed on an NVIDIA RTX 3070 Ti Laptop GPU. An AdamW optimizer [71] was used with a weight decay coefficient of 1×10^{-7} combined with a cosine annealing learning rate scheduler [72] with a minimum learning rate of 1×10^{-9}. The class-balanced softmax cross entropy loss [73] was used for training, which is given as:

$$L_{CE} = -\sum_{n=1}^{N} w_{t_n} \log \left(\frac{\exp(z_{n,t_n})}{\sum_{c=1}^{C} \exp(z_{n,c})} \right) \tag{5.1}$$

where z is the model output, t is the target, w is the class weight, C is the number of classes, and N is the number of samples in the minibatch.

The popular ResNet family of networks introduced in [74] was used as the encoder backbone for the U-Net architecture, which is pre-trained on the widely used and adaptable ImageNet [75] dataset. Initially, the ResNet-50 network with a depth of 50 layers was used. However, the network proved to be too large, resulting in overfitting on the training data. Therefore, a smaller ResNet-18 was chosen to be used for the experiments. To avoid training the encoder from scratch, pre-trained weights for ResNet-18 trained on ImageNet were loaded in with the weights from layers 1-3 being frozen during training. This was

done to maintain the earlier pre-trained feature extraction layers from the encoder, only fine-tuning the deepest layer to modify it for the purpose of detecting ice. This also helped with reducing the number of trainable parameters, speeding up the training process.

Image segmentation models are typically evaluated using two metrics that measure the performance of the prediction. Pixel accuracy and intersection over union (IoU) are two different ways of quantifying how well the predicted segmentation aligns with the ground truth. These two metrics are defined as follows:

$$Accuracy = \frac{TP}{TP + FN} \tag{5.2}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{5.3}$$

where TP is the number of true positive pixels, FP is the number of false positive pixels, and FN is the number of false negative pixels. Accuracy simply measures the ratio of correctly predicted pixels to the total number of pixels for that class. For multi-class segmentation, accuracy and IoU scores are calculated for each individual class and averaged across all classes to produce a single mean pixel accuracy (mAcc) and a single mean intersection over union (mIoU) performance score. Although mAcc highlights the detection ability per each class, it does not take into account false positives, which can lead to good scores even with over-segmentation. In contrast, mIoU accounts for and penalizes false positives, providing a more robust method of measuring how well the predicted regions align with the ground truth.

Chapter 6

Results and Discussion

6.1 Performance Results

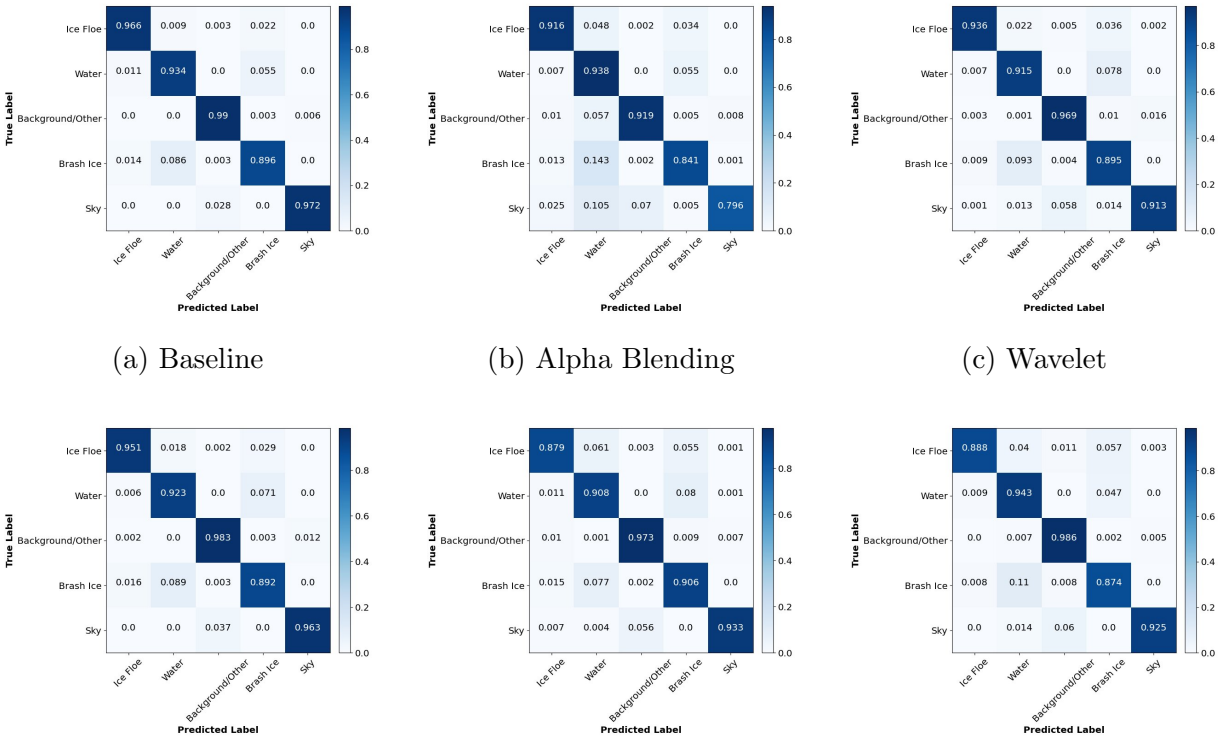
6.1.1 Early Fusion

Table 6.1: Early Fusion Results

	Full		**Easy**		**Hard**	
	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc
Baseline	**0.909**	**0.939**	**0.919**	**0.951**	**0.857**	**0.915**
Alpha Blending	0.822	0.900	0.891	0.931	0.679	0.832
Wavelet Fusion	0.868	0.921	0.889	0.936	0.828	0.888
Concat – RGB + DoLP	*0.893*	*0.930*	*0.906*	*0.941*	0.847	0.905
Concat – RGB + Thermal	0.859	0.909	0.860	0.912	0.840	0.904
Concat – RGB + DoLP + Thermal	0.865	0.917	0.864	0.919	*0.847*	*0.912*

Note: Best performance in **bold**, second best in *italics*.

For early fusion, the baseline case performed the best across all datasets, as seen in the results in Table 6.1. The two methods of creating false-colour 3-channel images, `alpha-blending` and `wavelet`, performed the worst on the full and hard subset. This is likely due

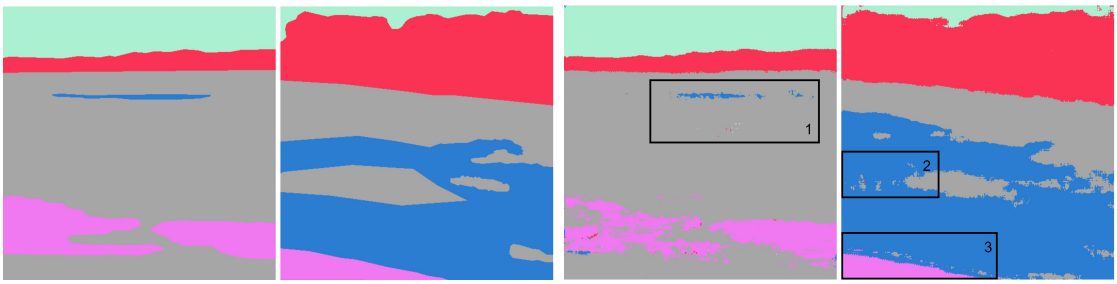


(a) Baseline (b) Alpha Blending (c) Wavelet

(d) Concat − RGB + DoLP (e) Concat − RGB + Thermal (f) Concat − RGB + DoLP + Thermal

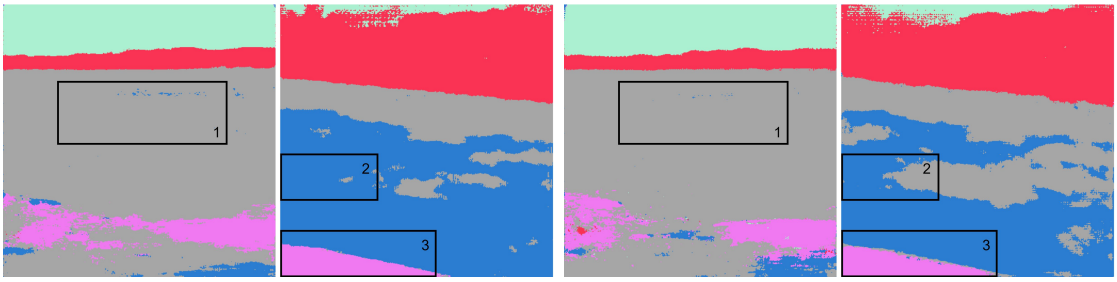
Figure 6.1: Confusion matrices of different early fusion methods compared to the baseline on the full dataset.

to the fact that there is too much valuable information lost this early on by trying to reduce the number of channels for the input. However, on the easy subset, `concat-thermal` and `concat-dolp-thermal` were the worst performers. Referring to the confusion matrices in Figure 6.1, this suggests that the inclusion of thermal data through early concatenation in optically ideal environments negatively impacts the ability of the network's ability to differentiate between ice floes and brash ice. This can be seen clearly in box 1 in Figure 6.2, where `concat-thermal` is completely missing the strip of water in the far field which is more correctly predicted by the baseline and `concat-dolp`. However, `alpha-blending`, `wavelet`, and `concat-dolp` all seem to suffer from over-segmentation of water compared to the baseline in the ideal case. However, the hard subset proved to be more challenging for the baseline, as it is missing a significant amount of brash ice as highlighted by box 2, which is present in both `concat-dolp` and `concat-thermal`. Additionally, the boundary between water and ice floe in box 3 is the most accurate in `alpha-blending`, followed by `concat-dolp`, with both performing better than the baseline. However, although each form of early fusion appears to have advantages in certain areas, each also includes significant drawbacks which prevent any of the early fusion methods from outperforming the baseline.



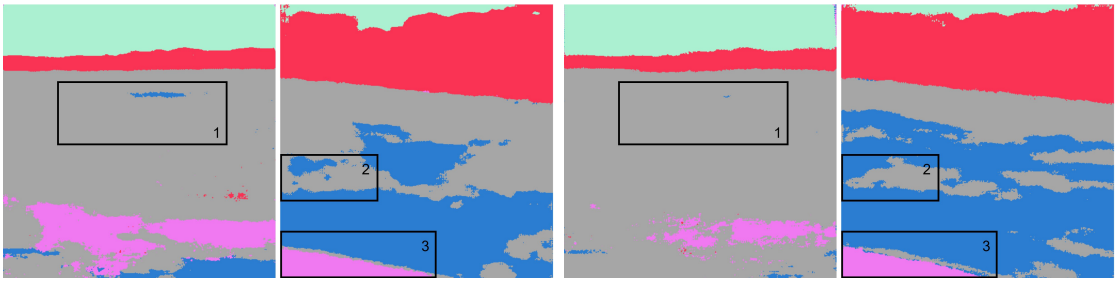
(a) Ground Truth

(b) Baseline



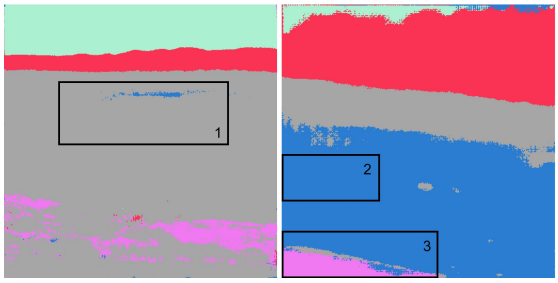
(c) Alpha Blending

(d) Wavelet



(e) Concat − RGB + DoLP

(f) Concat − RGB + Thermal



(g) Concat − RGB + DoLP + Thermal

Figure 6.2: Pairs of example outputs of different early fusion methods on the easy (left) and hard (right) test subsets compared against the baseline.

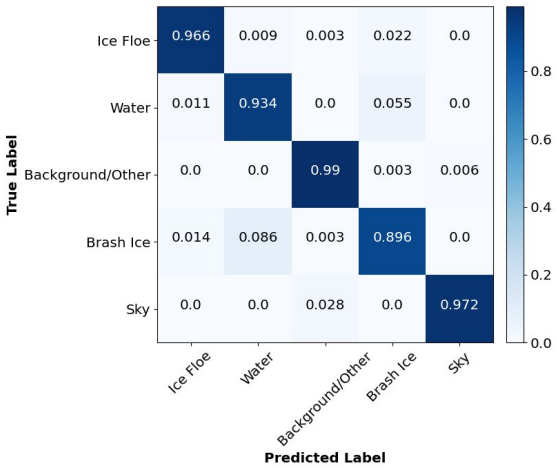
6.1.2 Middle Fusion

Table 6.2: Middle Fusion Results

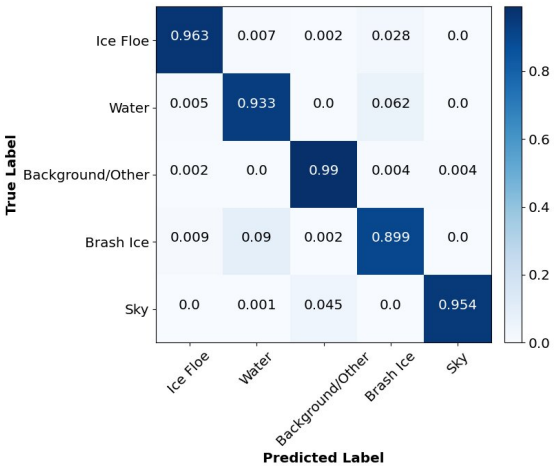
	Full		**Easy**		**Hard**	
	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc
Baseline	*0.909*	*0.939*	**0.919**	*0.951*	0.857	0.915
RGB + DoLP	0.907	0.938	*0.918*	0.950	*0.900*	0.912
RGB + Thermal	0.906	0.938	0.913	0.948	0.866	*0.916*
RGB + DoLP + Thermal	**0.913**	**0.944**	0.917	**0.953**	**0.903**	**0.925**

Note: Best performance in **bold**, second best in *italics*.

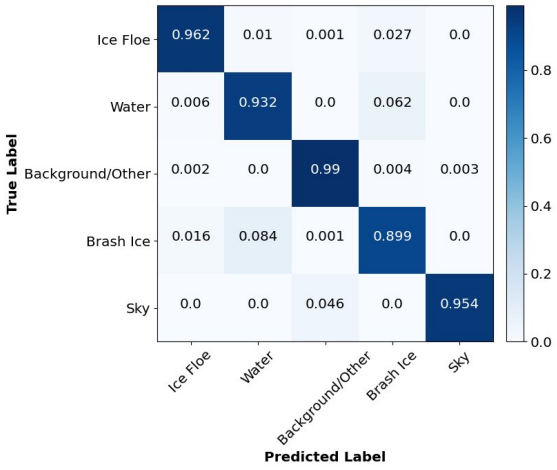
The results for middle fusion are shown in Table 6.2, with corresponding confusion matrices on the full dataset in Figure 6.3. The experiments with middle fusion show a slight advantage of fusion over the baseline for most cases. On the full dataset, `midfusion-dolp-thermal` achieves the highest overall performance, with slight improvements in both mIoU and mAcc over the baseline. On the hard subset, this improvement over the baseline is even more apparent, as seen in the corresponding confusion matrices in Figure 6.4. The results show that whereas the baseline struggles with differentiating between ice floe and water in difficult scenes, the inclusion of DoLP and thermal data together is complementary in most cases and helps to stabilize performance from degradation. This can be seen clearly in Figure 6.5, where the missing brash ice in box 2 from the baseline and `midfusion-dolp` are present in `midfusion-thermal` and `midfusion-dolp-thermal`. On the easy subset, the baseline performs similarly well compared to the fusion approaches, with the highest mIoU score and second-highest mAcc score. This is likely caused by the added noise introduced by the DoLP source resulting in a slight over-segmentation of water in clear but busy scenes. This shows that although RGB alone can perform sufficiently in ideal conditions, the addition of extra sources of input through middle fusion does not degrade performance as much as it does through early fusion. As such, middle fusion seems to be more robust in filtering out unhelpful and unnecessary information.



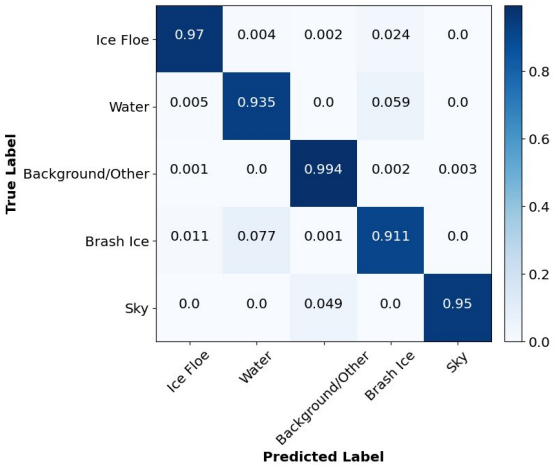
(a) Baseline



(b) RGB + DoLP

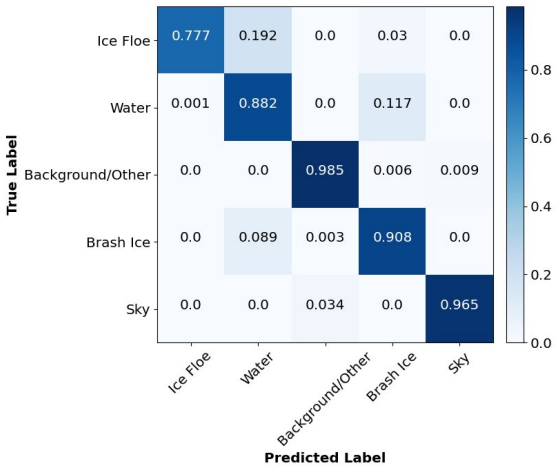


(c) RGB + Thermal

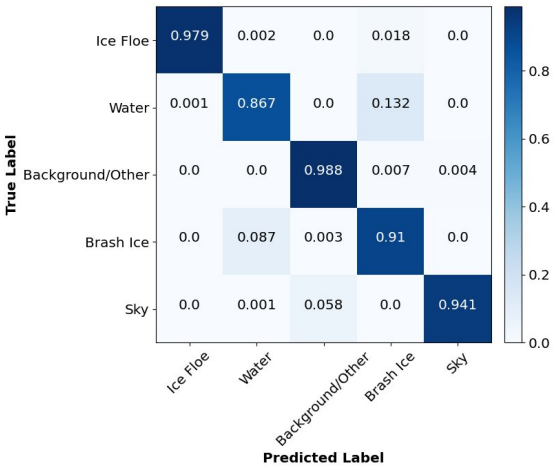


(d) RGB + DoLP + Thermal

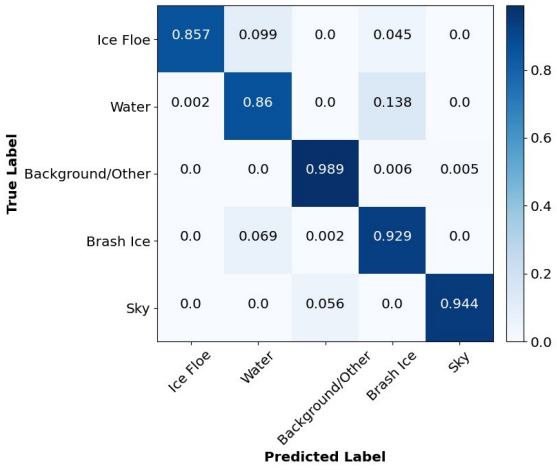
Figure 6.3: Confusion matrices of different middle fusion methods compared to the baseline on the full dataset.



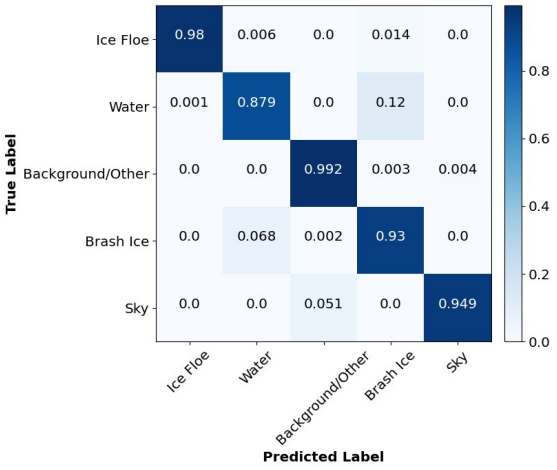
(a) Baseline



(b) RGB + DoLP

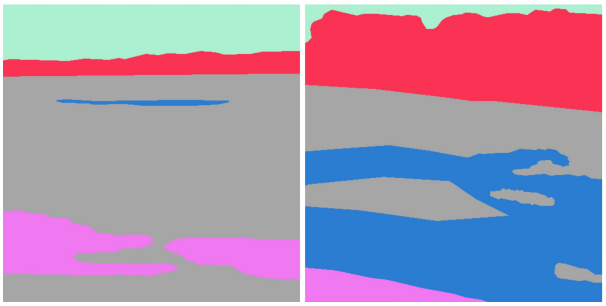


(c) RGB + Thermal

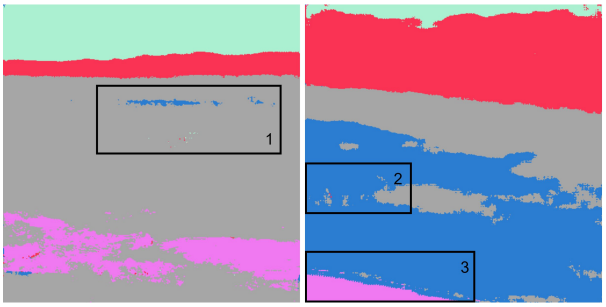


(d) RGB + DoLP + Thermal

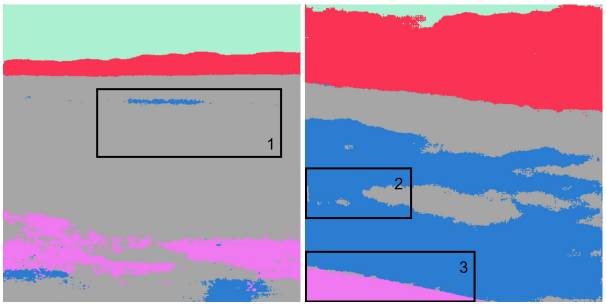
Figure 6.4: Confusion matrices of different middle fusion methods compared to the baseline on the hard dataset.



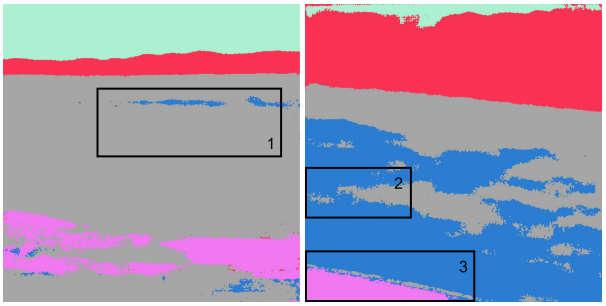
(a) Ground Truth



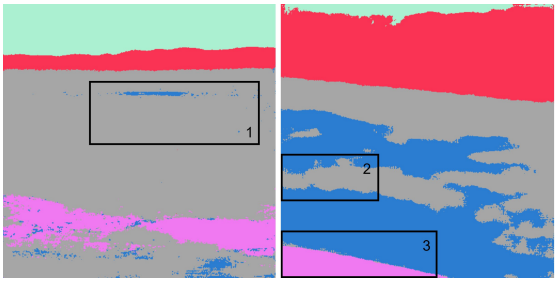
(b) Baseline



(c) Middle Fusion – RGB + DoLP



(d) Middle Fusion – RGB + Thermal



(e) Middle Fusion – RGB + DoLP + Thermal

Figure 6.5: Pairs of example outputs of different middle fusion methods on the easy (left) and hard (right) test subsets compared against the baseline.

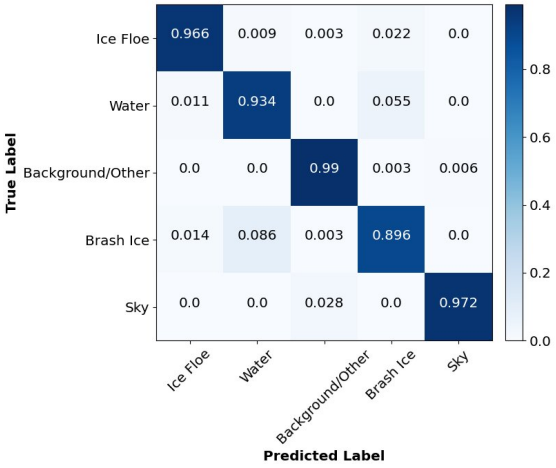
6.1.3 Late Fusion

Table 6.3: Late Fusion Results

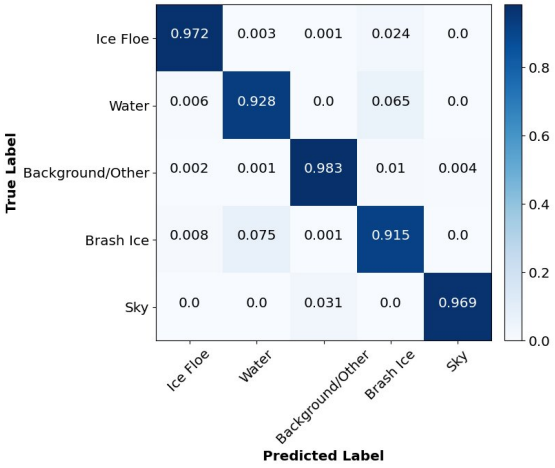
	Full		**Easy**		**Hard**	
	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc
Baseline	0.909	0.939	0.919	0.951	*0.857*	*0.915*
RGB + DoLP	**0.915**	**0.943**	**0.932**	**0.958**	0.851	0.909
RGB + Thermal	*0.915*	*0.942*	*0.925*	*0.955*	0.839	0.914
RGB + DoLP + Thermal	0.909	0.940	0.918	0.951	**0.873**	**0.916**

Note: Best performance in **bold**, second best in *italics*.

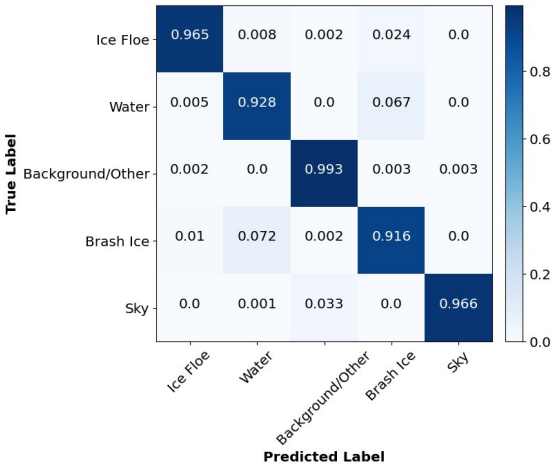
Finally, the experiments with late fusion show a more significant increase in fusion performance over the baseline, as seen in the results in Table 6.3. For the easy subset, `latefusion-dolp` achieves the highest mIoU and mAcc scores, performing marginally better than `latefusion-thermal`. In contrast, for the hard subset, `latefusion-dolp-thermal` has the highest mIoU and mAcc scores, followed by the baseline case. This indicates that like in the middle fusion case, the inclusion of all three input sources can provide complementary information in challenging scenes. This can be seen in box 2 in Figure 6.7, where all three late fusion methods were able to correctly include the section of brash ice which was missing from the baseline. Additionally, in box 3, the boundary between ice and water from `latefusion-dolp-thermal` is much better defined as compared to the baseline. However for simpler scenes, two input sources seem to be sufficient, with the inclusion of a third being detrimental to the model's ability to seperate out the different classes. As seen in box 1, `latefusion-dolp-thermal` is more consistent than the baseline which predicted random specs of background and sky in the brash ice, but the water feature in the far distance was not as well defined as either `latefusion-dolp` or `latefusion-thermal` on their own. Overall on the full dataset, `latefusion-dolp` performed the best followed by `latefusion-thermal`, showing a more clear advantage of late fusion over the baseline as compared to more subtle advantage provided by middle fusion.



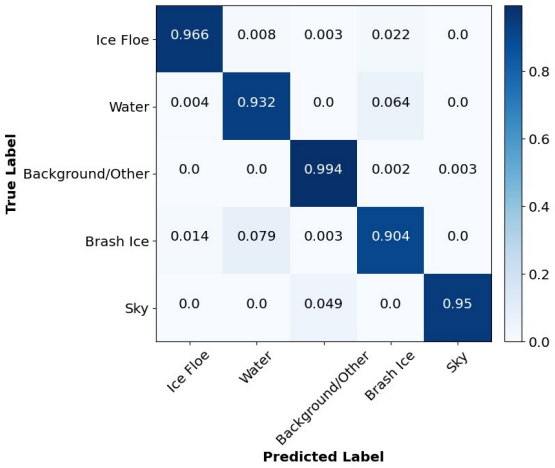
(a) Baseline



(b) RGB + DoLP

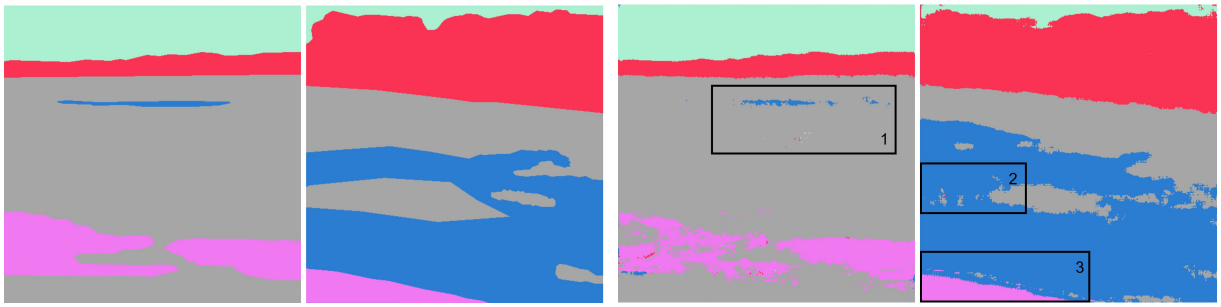


(c) RGB + Thermal



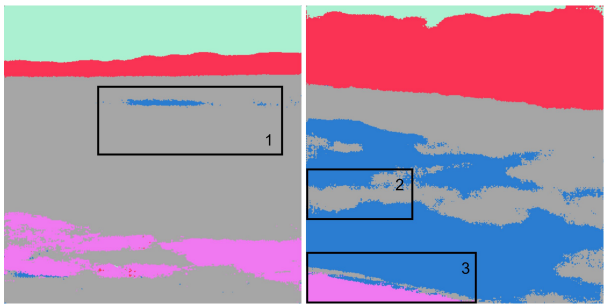
(d) RGB + DoLP + Thermal

Figure 6.6: Confusion matrices of different late fusion methods compared to the baseline on the full dataset.

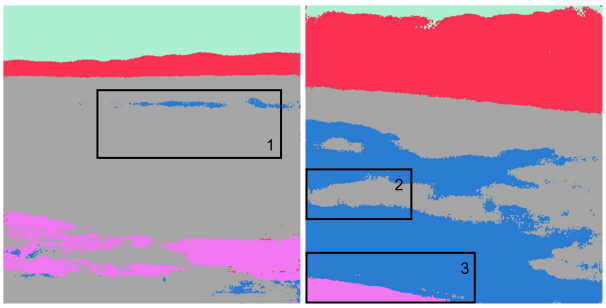


(a) Ground Truth

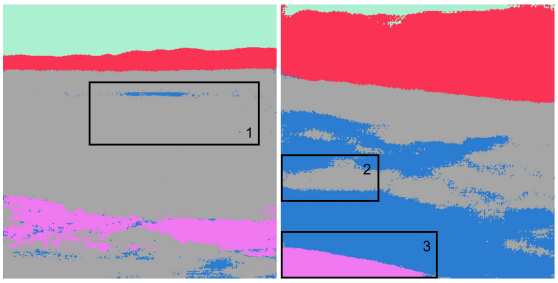
(b) Baseline



(c) Late Fusion − RGB + DoLP



(d) Late Fusion − RGB + Thermal



(e) Late Fusion − RGB + DoLP + Thermal

Figure 6.7: Pairs of example outputs of different late fusion methods on the easy (left) and hard (right) test subsets compared against the baseline.

Chapter 7

Conclusion

7.1 Summary

This thesis explored the usage of a multi-sensor system to improve the ability to segment outdoor ice scenes. A novel sensor platform consisting of optical RGB, polarization, thermal, and stereoscopic cameras was prototyped over multiple iterations and field tested in February 2025. During these tests, river ice data was collected along the shores of the Ottawa and St. Lawrence rivers in a variety of lighting and weather conditions. Techniques were developed to calibrate and synchronize the different cameras together, while post-processing was conducted to co-register each camera and warp the captures to a common viewpoint. A subset of the collected data was manually labelled, creating a dataset of 133 sets of images for conducting data fusion experiments. With this dataset, different combinations of sensor inputs were tested using a variety of early, middle, and late fusion methods.

Semantic segmentation performance on each fusion method was assessed using a modified U-Net architecture with a ResNet-18 encoder backbone. Results were compared against the baseline RGB case, which achieved a mIoU of 0.9094 and a mAcc of 0.9393 on the full test dataset. Despite the already strong performance, the baseline RGB case struggles with differentiating between brash ice and water, with performance significantly degrading in the presence of artifacts such as snow or reflections of sunlight.

Early fusion methods consistently performed worse than the baseline. Although each method showed some localized improvements, drawbacks involving the loss of information and the introduction of noisy features limited the performance of the network when

presented with the additional input sources. Middle fusion methods showed some improvements over the baseline, with `midfusion-dolp-thermal` achieving a mIoU of 0.9134 and a mAcc of 0.9439 on the full test dataset. Although the thermal data is typically more noisy than the optical RGB data, this shows the ability of middle fusion to be robust to noisy sensor inputs while still extracting useful feature information from the additional sources. Finally, late fusion methods showed the largest improvement over the baseline, with `latefusion-dolp` achieving a mIoU of 0.9153 and a mAcc of 0.9427 on the full test dataset. The inclusion of polarization data at the decision level was able to enhance the network's ability to differentiate between brash ice and water, while the inclusion of the noisy thermal data at this late stage proved to have a detrimental effect on the performance of the model.

These results highlight the advantages offered by the inclusion of additional input sources in a fusion setting. Intuitively, late fusion was able to extract and integrate the most amount of useful information due to putting more emphasis on individual sensor performance. Meanwhile, middle fusion was more robust at integrating additional sensor inputs with some complementary information without degrading performance.

7.2 Future Work

7.2.1 Improved Training Dataset

A major obstacle in this study was the lack of data variety available for sensor fusion experiments. Due to the unique nature of the multi-sensor system, all experimental data had to be collected in the field. With access to only a few types of ice conditions accessible from land-based locations and a limited number of different environmental conditions, this severely restricted the variety of the dataset. Due to the bias towards ideal conditions and a large focus on brash ice/water discrimination, certain sensors may have appeared to be less effective than when used on a more balanced dataset. Considering the intended use case on board ships travelling through ice-covered waters, bringing the system to collect data during such a trip would allow for a more accurate test of the capabilities of each sensor. Sea ice features such as icebergs and multi-year ice which can look visibly similar from afar may reveal strengths and weaknesses in a specific sensor's ability to classify them. Meanwhile, conditions not experienced during the trials conducted in this study such as storms and rough waters can challenge the sensitivity of the sensor to non-ideal conditions. Furthermore, increasing the amount of data available for training and testing the model will reduce the chance of an outlier image significantly influencing performance metrics.

7.2.2 Depth Map Refinement

Although stereoscopic cameras were included in the sensor system, its use was limited due to poor results with dense stereo reconstruction. Depth information can be extremely useful for ruling out inconsistencies in the scene and providing information on the relative size of objects in view. However, the lack of resolution and texture in the images introduces difficulties in stereo matching for disparity map estimation, even with fully calibrated and rectified cameras. An alternative method to obtain depth information is to instead use a more active approach by using LIDAR. Recent advances in low-cost solid-state LIDAR technology enables integration of these sensors into portable systems, providing accurate depth information even in poor lighting conditions. Regardless of the method of obtaining depth information, introducing an additional modality could increase the model's ability to enforce consistent predictions across all input sources.

7.2.3 Thermal Data Usage

While the thermal data proved to be a helpful addition in some cases, changes in the data collection and post-processing pipeline can increase the significance of this data. In theory, the different emissivities of water and ice in the TIR range combined with their different surface temperatures (typically around 0 °C for water surrounded by ice, and significantly below that for non-melting ice) should allow for a thermal imaging camera to clearly differentiate between the two. However, reflections and automatic exposure adjustment of the sensor made it difficult to accurately determine the true temperature ranges of the surfaces in view, thus relying on context clues and manual temperature normalization to determine the boundaries between water and ice. By having a reference object with a known temperature in view of the frame, surface temperatures can be more consistently extracted and normalized in preparation for creating the dataset. Combined with a more refined depth map allowing for possibly highly reflective areas in view to be masked out, higher quality thermal data can be provided to the network for training and testing.

References

[1] R. R. J. Duan, K. A. Scott, C. G. J. MacMillan, R. M. Gash, and Z. Pan, "A Sensor Fusion Platform for Semantic Segmentation of Sea Ice Imagery," in *Proceedings of the 28th International Conference on Port and Ocean Engineering under Arctic Conditions (POAC)*, 2025.

[2] L. Dicks, R. Almond, and A. McIvor, *Arctic Climate Issues 2011: Changes in Arctic Snow, Water, Ice and Permafrost*. Oslo, Norway: Arctic Monitoring and Assessment Programme, 2013.

[3] W. N. Meier and J. Stroeve, "An Updated Assessment of the Changing Arctic Sea Ice Cover," *Oceanography*, vol. 35, no. 3/4, pp. 10–19, 2022.

[4] R. Ricker, S. Hendricks, L. Kaleschke, X. Tian-Kunze, J. King, and C. Haas, "A weekly Arctic sea-ice thickness data record from merged CryoSat-2 and SMOS satellite data," *The Cryosphere*, vol. 11, no. 4, pp. 1607–1623, 2017.

[5] J. M. Kim, B. J. Sohn, S. M. Lee, R. T. Tonboe, E. J. Kang, and H. C. Kim, "Differences Between ICESat and CryoSat-2 Sea Ice Thicknesses Over the Arctic: Consequences for Analyzing the Ice Volume Trend," *Journal of Geophysical Research: Atmospheres*, vol. 125, no. 22, p. e2020JD033103, 2020.

[6] A. J. Cook, J. Dawson, S. E. L. Howell, J. E. Holloway, and M. Brady, "Sea ice choke points reduce the length of the shipping season in the Northwest Passage," *Communications Earth & Environment*, vol. 5, p. 362, 2024.

[7] N. Melia, K. Haines, and E. Hawkins, "Sea ice decline and 21st century trans-Arctic shipping routes," *Geophysical Research Letters*, vol. 43, no. 18, pp. 9720–9728, 2016.

[8] Y. Cao, S. Liang, L. Sun, J. Liu, X. Cheng, D. Wang, Y. Chen, M. Yu, and K. Feng, "Trans-Arctic shipping routes expanding faster than the model projections," *Global Environmental Change*, vol. 73, p. 102488, 2022.

[9] S. Somanathan, P. Flynn, and J. Szymanski, "The Northwest Passage: A Simulation," *Transportation Research Part A: Policy and Practice*, vol. 43, no. 2, pp. 127–135, 2009.

[10] L. R. Mudryk, J. Dawson, S. E. L. Howell, C. Derksen, T. A. Zagon, and M. Brady, "Impact of 1, 2 and 4 °C of global warming on ship navigation in the Canadian Arctic," *Nature Climate Change*, vol. 11, pp. 673–679, August 2021.

[11] M. R. Brooks and J. D. Frost, "Providing freight services to remote arctic communities: Are there lessons for practitioners from services to Greenland and Canada's northeast?," *Research in Transportation Business & Management*, vol. 4, pp. 69–78, 2012.

[12] H. P. Huntington, J. Olsen, E. Zdor, A. Zagorskiy, H. C. Shin, O. Romanenko, B. Kaltenborn, J. Dawson, J. Davies, and E. Abou-Abbsi, "Effects of Arctic commercial shipping on environments and communities: context, governance, priorities," *Transportation Research Part D: Transport and Environment*, vol. 118, p. 103731, 2023.

[13] V. V. Lehtola, J. Montewka, and J. Salokannel, "Sea Captains' Views on Automated Ship Route Optimization in Ice-covered Waters," *Journal of Navigation*, vol. 73, no. 2, p. 364–383, 2020.

[14] Y. Shu, Y. Zhu, F. Xu, L. Gan, P. T.-W. Lee, J. Yin, and J. Chen, "Path planning for ships assisted by the icebreaker in ice-covered waters in the Northern Sea Route based on optimal control," *Ocean Engineering*, vol. 267, p. 113182, 2023.

[15] K. Kulkarni, P. Kujala, M. Musharraf, and I. Rainio, "Simulation Tool for Winter Navigation Decision Support in the Baltic Sea," *Applied Sciences*, vol. 12, no. 15, p. 7568, 2022.

[16] V. Kotovirta, R. Jalonen, L. Axell, K. Riska, and R. Berglund, "A system for route optimization in ice-covered waters," *Cold Regions Science and Technology*, vol. 55, no. 1, pp. 52–62, 2009.

[17] J. Dawson, A. Cook, J. Holloway, and L. Copland, "Analysis of Changing Levels of Ice Strengthening (Ice Class) among Vessels Operating in the Canadian Arctic over the Past 30 Years," *ARCTIC*, vol. 75, p. 413–430, December 2022.

[18] J. Li, S. Fu, and Y. Zhang, "Feature Analysis of Icebreaker Escort Operations in Ice-Covered Arctic Waters," in *2023 7th International Conference on Transportation Information and Safety (ICTIS)*, pp. 219–223, 2023.

[19] M. Garvin, "Review of Technologies for Real-Time Shipboard Ice Severity Sensing," in *Smart Ship Technology*, pp. 53–62, 2020.

[20] R. Galley, M. Trachtenberg, A. Langlois, D. Barber, and L. Shafai, "Observations of geophysical and dielectric properties and ground penetrating radar signatures for discrimination of snow, sea ice and freshwater ice thickness," *Cold Regions Science and Technology*, vol. 57, no. 1, pp. 29–38, 2009.

[21] C. Haas, J. Lobach, S. Hendricks, L. Rabenstein, and A. Pfaffling, "Helicopter-borne measurements of sea ice thickness, using a small and lightweight, digital EM system," *Journal of Applied Geophysics*, vol. 67, no. 3, pp. 234–241, 2009.

[22] H.-M. Heyn, M. Blanke, and R. Skjetne, "Ice Condition Assessment Using Onboard Accelerometers and Statistical Change Detection," *IEEE Journal of Oceanic Engineering*, vol. 45, no. 3, pp. 898–914, 2020.

[23] T. Browne, R. Taylor, B. Veitch, P. Kujala, F. Khan, and D. Smith, "A Framework for Integrating Life-Safety and Environmental Consequences into Conventional Arctic Shipping Risk Models," *Applied Sciences*, vol. 10, no. 8, 2020.

[24] L. E. B. Eriksson, Y. Ye, L. Jonasson, W. Qazi, W. Mao, H. Wang, J. Möller, K. Lemmens, and S. Dokken, "EONav - Copernicus Data in Support of Maritime Route Optimization," in *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, pp. 5606–5609, 2018.

[25] X. Zhang, J. Jin, Z. Lan, C. Li, M. Fan, Y. Wang, X. Yu, and Y. Zhang, "ICENET: A Semantic Segmentation Deep Network for River Ice by Fusing Positional and Channel-Wise Attentive Features," *Remote Sensing*, vol. 12, no. 2, 2020.

[26] F. Hicks, "An overview of river ice problems: CRIPE07 guest editorial," *Cold Regions Science and Technology*, vol. 55, no. 2, pp. 175–185, 2009.

[27] World Meteorological Organization, "WMO Sea-Ice Nomenclature," Nomenclature WMO-No. 259, Geneva, 2014.

[28] H. Boulze, A. Korosov, and J. Brajard, "Classification of Sea Ice Types in Sentinel-1 SAR Data Using Convolutional Neural Networks," *Remote Sensing*, vol. 12, no. 13, 2020.

[29] N. Zakhvatkina, V. Smirnov, and I. Bychkova, "Satellite SAR Data-based Sea Ice Classification: An Overview," *Geosciences*, vol. 9, no. 4, 2019.

[30] J. Brown, M. King, R. Briggs, and R. Yulmetov, "Evaluation of Near-Ship Ice Conditions from Ship Bourne Sensors," in *Proceedings of the 33rd International Ocean and Polar Engineering Conference*, 2023.

[31] R. J. Hall, N. Hughes, and P. Wadhams, "A systematic method of obtaining ice concentration measurements from ship-based observations," *Cold Regions Science and Technology*, vol. 34, no. 2, pp. 97–102, 2002.

[32] Q. Zhang and R. Skjetne, "Image Processing for Identification of Sea-Ice Floes and the Floe Size Distributions," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 5, pp. 2913–2924, 2015.

[33] N. Panchi, E. Kim, and A. Bhattacharyya, "Supplementing Remote Sensing of Ice: Deep Learning-Based Image Segmentation System for Automatic Detection and Localization of Sea-ice Formations From Close-Range Optical Images," *IEEE Sensors Journal*, vol. 21, no. 16, pp. 18004–18019, 2021.

[34] N. Panchi and E. Kim, "Deep learning strategies for analysis of weather-degraded optical sea ice images," *IEEE Sensors Journal*, vol. 24, no. 9, pp. 15252–15272, 2024.

[35] L. Zhou, J. Cai, and S. Ding, "The Identification of Ice Floes and Calculation of Sea Ice Concentration Based on a Deep Learning Method," *Remote Sensing*, vol. 15, no. 10, p. 2663, 2023.

[36] E. Kim, G. S. Dahiya, S. Løset, and R. Skjetne, "Can a computer see what an ice expert sees? Multilabel ice objects classification with convolutional neural networks," *Results in Engineering*, vol. 4, p. 100036, 2019.

[37] C. Zhang, X. Chen, and S. Ji, "Semantic image segmentation for sea ice parameters recognition using deep convolutional neural networks," *International Journal of Applied Earth Observation and Geoinformation*, vol. 112, p. 102885, 2022.

[38] B. Dowden, O. De Silva, W. Huang, and D. Oldford, "Sea Ice Classification via Deep Neural Network Semantic Segmentation," *IEEE Sensors Journal*, vol. 21, no. 10, pp. 11879–11888, 2021.

[39] M. L. Fung, M. Z. Q. Chen, and Y. H. Chen, "Sensor fusion: A review of methods and applications," in *Proceedings of the 29th Chinese Control and Decision Conference (CCDC)*, pp. 3853–3860, 2017.

[40] A. L. Rankin and L. H. Matthies, "Passive sensor evaluation for unmanned ground vehicle mud detection," *Journal of Field Robotics*, vol. 27, no. 4, pp. 473–490, 2010.

[41] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

[42] R. Azad, E. K. Aghdam, A. Rauland, Y. Jia, A. H. Avval, A. Bozorgpour, S. Karimijafarbigloo, J. P. Cohen, E. Adeli, and D. Merhof, "Medical Image Segmentation Review: The Success of U-Net," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 10076–10095, 2024.

[43] M. Brenner, N. H. Reyes, T. Susnjak, and A. L. C. Barczak, "RGB-D and Thermal Sensor Fusion: A Systematic Literature Review," *IEEE Access*, vol. 11, pp. 82410–82442, 2023.

[44] D. K. Perovich, "Light reflection from sea ice during the onset of melt," *Journal of Geophysical Research: Oceans*, vol. 99, no. C2, pp. 3351–3359, 1994.

[45] S. Nayar, X. Fang, and T. Boult, "Removal of Specularities Using Color and Polarization," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 583–590, 1993.

[46] J. S. Tyo, C. F. LaCasse, and B. M. Ratliff, "Total elimination of sampling errors in polarization imagery obtained with integrated microgrid polarimeters," *Optics letters*, vol. 34, no. 20, pp. 3187–3189, 2009.

[47] "Raspberry Pi Compute Module 4." https://www.raspberrypi.com/products/compute-module-4. (Accessed April 22, 2026).

[48] "StereoPi." https://stereopi.com/. (Accessed April 22, 2026).

[49] Arducam, "Quick Start Guide for Multi-Camera Adapter Board." https://docs.arducam.com/Raspberry-Pi-Camera/Multi-Camera-CamArray/Quick-Start-Guide-for-Multi-Adapter-Board. (Accessed April 22, 2026).

[50] LUCID Vision Labs, "Triton Industrial Camera." https://thinklucid.com/triton-gige-machine-vision. (Accessed April 22, 2026).

[51] K. Xiao, X. Kang, H. Liu, and P. Duan, "MOFA: A novel dataset for Multi-modal Image Fusion Applications," *Information Fusion*, vol. 96, pp. 144–155, 2023.

[52] R. M. Gash, K. A. Murrant, J. W. Mills, and D. E. Millan, "Machine Vision Techniques for Situational Awareness and Path Planning in Model Test Basin Ice-Covered Waters," in *Global Oceans 2020: Singapore – U.S. Gulf Coast*, pp. 1–8, IEEE, 2020.

[53] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330–1334, 2002.

[54] A. ElSheikh, B. A. Abu-Nabah, M. O. Hamdan, and G.-Y. Tian, "Infrared Camera Geometric Calibration: A Review and a Precise Thermal Radiation Checkerboard Target," *Sensors*, vol. 23, no. 7, p. 3479, 2023.

[55] M. Smith and J. Thomson, "Pancake sea ice kinematics and dynamics using shipboard stereo video," *Annals of Glaciology*, vol. 61, no. 82, pp. 1–11, 2020.

[56] G. D. Ashton, "River and lake ice thickening, thinning, and snow ice formation," *Cold Regions Science and Technology*, vol. 68, no. 1, pp. 3–19, 2011.

[57] L. Polyak, R. B. Alley, J. T. Andrews, J. Brigham-Grette, T. M. Cronin, D. A. Darby, A. S. Dyke, J. J. Fitzpatrick, S. Funder, M. Holland, A. E. Jennings, G. H. Miller, M. O'Regan, J. Savelle, M. Serreze, K. St. John, J. W. White, and E. Wolff, "History of sea ice in the arctic," *Quaternary Science Reviews*, vol. 29, no. 15, pp. 1757–1778, 2010. Special Theme: Arctic Palaeoclimate Synthesis (pp. 1674-1790).

[58] D. Perlman, J. Stanford, and E. Wallischeck, "St. Lawrence Seaway: Overview of Safety, Efficiency, Operational, and Environmental Issues," Research Paper, John A. Volpe National Transportation Systems Center, April 2017. DOT-VNTSC-FWA-17-16.

[59] J. Blackburn and Y. She, "A comprehensive public-domain river ice process model and its application to a complex natural river," *Cold Regions Science and Technology*, vol. 163, pp. 44–58, 2019.

[60] "MATLAB." https://www.mathworks.com/products/matlab.html. (Accessed April 22, 2026).

[61] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2nd ed., 2004.

[62] M. S. Hamid, N. Abd Manap, R. A. Hamzah, and A. F. Kadmin, "Stereo matching algorithm based on deep learning: A survey," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 5, pp. 1663–1673, 2022.

[63] M. V. Rohith, G. Somanath, C. Kambhamettu, C. Geiger, and D. Finnegan, "Modified Region Growing for Stereo of Slant and Textureless Surfaces," in *Advances in Visual Computing*, pp. 666–677, Springer, Berlin, Heidelberg, 2010.

[64] G. C. Holst, *Testing and Evaluation of Infrared Imaging Systems*. Winter Park, Florida: JCD Publishing, 3rd ed., 2020.

[65] R. Chave, D. Lemon, D. Fissel, L. Dupuis, and S. Dumont, "Real-time measurements of ice draft and velocity in the St. Lawrence River," in *Oceans '04 MTS/IEEE Techno-Ocean '04*, vol. 3, pp. 1629–1633, 2004.

[66] "About CVAT: Leading Open-Source Data Annotation Platform." https://www.cvat.ai/about. (Accessed April 22, 2026).

[67] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick, "Segment Anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, 2023.

[68] A. Toet, "Natural colour mapping for multiband nightvision imagery," *Information Fusion*, vol. 4, no. 3, pp. 155–166, 2003.

[69] K. Amolins, Y. Zhang, and P. Dare, "Wavelet based image fusion techniques — An introduction, review and comparison," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 62, no. 4, pp. 249–263, 2007.

[70] A. G. Howard, "Some Improvements on Deep Convolutional Neural Network Based Image Classification," in *2nd International Conference on Learning Representations (ICLR)*, 2014.

[71] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *7th International Conference on Learning Representations (ICLR)*, 2019.

[72] I. Loshchilov and F. Hutter, "SGDR: Stochastic Gradient Descent with Warm Restarts," in *5th International Conference on Learning Representations (ICLR)*, 2017.

[73] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-Balanced Loss Based on Effective Number of Samples," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9260–9269, 2019.

[74] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[75] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, p. 84–90, May 2017.