

Advancing Glacier Mapping with Glacier-Seg: A Lightweight Deep Learning Model for Efficient Semantic Segmentation in Multi-Modal Remote Sensing Imagery

by

Meheruba Hasin Alif
22341011

Marzanul Momenine
22301196

Syeda Mahmuda
22301142

A Thesis submitted to the Department of Computer Science and Engineering
in partial fulfilment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
October 2025

© 2025. Brac University
All rights reserved.

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

blankimage.png

Meheruba Hasin Alif
22341011

blankimage.png

Marzanul Momenine
22301196

blankimage.png

Syeda Mahmuda
22301142

Approval

The thesis titled “Advancing Glacier Mapping with Glacier-Seg: A Lightweight Deep Learning Model for Efficient Semantic Segmentation in Multi-Modal Remote Sensing Imagery” submitted by

1. Meheruba Hasin Alif(22341011)
2. Marzanul Momenine(22301196)
3. Syeda Mahmuda(22301142)

Of Summer, 2025 has been accepted as satisfactory in partial fulfilment of the requirement for the degree of B.Sc. in Computer Science on August 23, 2025.

Examining Committee:

Supervisor:
(Member)

blankimage.png

Amitabha Chakrabarty, PhD
Professor
Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)

blankimage.png

Sanjida Tasnim
Adjunct Lecturer
Department of Computer Science and Engineering
Brac University

Program coordinator:
(Member)

blankimage.png

Md. Golam Rabiul Alam, PhD
Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

blankimage.png

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Ethics Statement

This research adheres strictly to ethical standards, using only publicly available and ethically sourced satellite datasets. No personally identifiable information is collected or utilized. Given that inaccurate predictions or misinterpretation of hazard zones could have significant societal impacts, all model outputs are accompanied by uncertainty metrics. Furthermore, model development prioritises transparency, reproducibility, and explainability, ensuring that stakeholders and local communities can make informed decisions based on reliable data.

Table of Contents

Declaration	i
Approval	ii
Ethics Statement	iv
Abstract	viii
Dedication	ix
Acknowledgment	x
List of Figures	xi
List of Tables	xiii
1 Introduction	2
1.1 Background	2
1.2 Rationale of the Study	3
1.3 Problem Statement	4
1.4 Research Objective	6
1.5 Methodology in Brief	7
1.6 Scopes and Challenges	10
1.6.1 Scope of the Study	10
1.6.2 Technical Challenges	11
2 Literature Review	12
2.1 Preliminaries	12
2.2 Review of Existing Research	12
2.2.1 Datasets	12
2.2.2 Traditional Machine Learning Approaches	14
2.2.3 CNN-Based Models for Glacier Segmentation	15
2.2.4 Emerging Transformer and Hybrid Architectures	16
2.3 Summary of Key Finding	17
2.4 Research Gap	23
3 Project Planning and Impacts	25
3.1 Final Specifications and Requirements	25
3.1.1 Data Requirements	25
3.1.2 Computational Resources	25

3.1.3	Model Requirements	25
3.1.4	Training Specifications	26
3.2	Societal Impact	26
3.3	Environmental Impact	27
3.4	Ethical Issues	27
3.5	Project Management Plan	27
3.6	Risk Management	28
3.7	Economic Analysis	28
4	Research Methodology	30
4.1	Methodology Overview	30
4.1.1	Datasets	30
4.1.2	Data Preprocessing	41
4.1.3	Models for Glacier Segmentation	45
4.1.4	Manual Fine-tuning	54
4.1.5	Evaluation	58
4.2	Design (Model) Specification	60
4.2.1	Overall Architecture	60
4.2.2	Key Components	61
4.2.3	Hyperparameters	62
5	Final Design Adjustment	63
5.1	Architecture	65
5.1.1	Input Tensor	65
5.1.2	Involution Patch Embedding	65
5.1.3	Mamba State-Space Mixer	66
5.1.4	Hierarchical MiT-Style Backbone	68
5.1.5	SegFormer Head Decoder	69
5.1.6	Regularisation with DropPath	70
5.2	Analysis of Design Solutions	71
5.2.1	Dataset Ablations on NIRD	71
5.2.2	Blockwise Ablations	74
5.2.3	Layerwise Ablations	75
5.3	Statistical Analysis	76
6	Results and Discussion	81
6.1	Performance Evaluation	81
6.1.1	Quantitative Analysis of Segmentation Model Performance	81
6.1.2	ROC–AUC and Confusion Matrix Analysis:	94
6.1.3	Comparative Evaluation Across Datasets (NIRD and HKH)	96
6.1.4	Runtime and Deployment Efficiency	98
6.2	Discussion	99
6.2.1	Architecture and Dataset Analysis	99
6.2.2	Comparisons and Relationships	105
6.2.3	Temporal Data Analysis	107
6.2.4	Forecasting Pipeline	108
6.2.5	Limitations	109
6.2.6	Future Work	110

7	Conclusion	113
	References	115
	Appendix A: Diagrams	121
7.1	Architecture Diagrams	122
7.1.1	U-Net	122
7.1.2	ResNet-50	123
7.1.3	EfficientNet	124
7.1.4	VGG-19	125
7.1.5	DeepLabV3	126
7.1.6	SegFormer	127
7.1.7	Vision Mamba	129
7.1.8	Glacier-Seg	131
7.2	Methodology	133

Abstract

The accelerating rise in global temperature is reflected in the rapid retreat of glaciers worldwide. Continuous, accurate glacier mapping is therefore essential for monitoring ice-mass balance, regional water resources and downstream hydrology. Instead of the slow and tedious manual process of assessing and classifying glacial boundaries, computational methods are sought out for automatically detecting and analysing glacier changes from remote sensing data sources. We introduce Glacier-Seg, a lightweight hybrid *Mamba-Transformer-CNN* architecture developed for efficient and accurate semantic segmentation of multi-modal remote sensing imagery. The research addresses major challenges in automated glacier mapping, including sparse ground truth annotations, sensor noise, and high computational complexity. Experiments were conducted using the *CaFFe* dataset (Synthetic Aperture Radar-only) and the *NIRD* dataset (multimodal SAR, optical, and digital elevation data). Glacier-Seg achieved a mean Intersection-over-Union (mIoU) of **0.956** and a Dice coefficient of **0.977** on NIRD with only **0.68M parameters**, outperforming several established architectures including U-Net, ResNet-50, DeepLabV3+, SegFormer, and Vision Mamba. The performance improvements of Glacier-Seg were statistically validated using one-way ANOVA and paired t-tests ($p < 0.001$), confirming the significance of the proposed model enhancements. The limitations of current deep learning models reveal a lack of adaptability when processing multimodal, high-resolution remote sensing imagery, where sensor noise and dataset sparsity often lead to poor model generalisation. As a computationally efficient and scalable framework for cryospheric monitoring, Glacier-Seg can become capable of real-time glacier boundary tracking and change detection. Future extensions will explore temporal forecasting, improved data harmonisation, and evaluation across additional global glacier datasets to enhance generalisability and long-term environmental applicability.

Keywords: Glacier Inventory; Glacier Monitoring; Glacier Mapping; Machine Learning (ML); Cryosphere; Remote Sensing; Classification; Vision Transformers ; Artificial Intelligence; Computer Vision; Deep Learning Algorithms;

Dedication

We dedicate this work to our esteemed supervisor, Dr. Amitabha Chakrabarty, whose unwavering guidance and profound insights illuminated our path through the complexities of cryospheric research, and to our co-supervisor, Sanjida Tasnim, whose meticulous mentorship fostered not only our technical prowess but also motivation to push through till the end.

Furthermore, we would also like to dedicate this work to a few people who have been the source of resilience behind the authors. To the gentle and persistent spirit of Meheruba's mother, Mabilia, whose quiet sacrifices sustained her through endless nights of code and contemplation, and to Safa Amin Hanan for instilling the perseverance in her to pursue truths as vast as the glaciers we study. To Syeda's sister Samira Mahmuda, without whom her undergraduate journey would have been tough and her mother Hasina Momtaz, whose support and prayers has been indispensable for her amidst the tight deadlines.

Their influences, profound and personal alike, have etched this endeavour with grace and gratitude.

Acknowledgement

This research would not have been possible without the continued guidance, encouragement, and intellectual generosity of our supervisor, **Dr. Amitabha Chakrabarty**, whose insightful feedback and patient mentorship have shaped the direction and depth of this work. We are equally grateful to our co-supervisor, **Sanjida Tasnim Ma'am**, for her thoughtful advice, constructive criticism, and constant support which have both refined and inspired the technical aspects of this thesis.

We would also like to extend a heartfelt gratitude to our research assistant, **Azwad Aziz** and former research assistant, **Fahim-Ul-Islam** for their invaluable help with model evaluation techniques, and discussions that strengthened the empirical framework of this study.

Our sincere appreciation also goes to the Department of Computer Science and Engineering at BRAC University for providing the computational resources and academic environment that enabled this research. We also acknowledge the open-source community and research groups behind *PyTorch*, *Earthdata*, *NASA*, and *Google Earth Engine*, whose publicly available tools and datasets formed the backbone of this work. Additionally, we express our gratitude to the creators of the CaFFe, NIRD, and HKH datasets—particularly the PANGAEA team and the GLIMS initiative—for their meticulous curation of cryospheric data, which made our multimodal analysis feasible.

Finally, we owe deep gratitude to our family and close friends for their unwavering support, patience, and faith. Their encouragement has been the quiet strength behind every line of code, every paragraph written, and every challenge overcome throughout this journey.

List of Figures

1.1	Retreat of the Upsala Glacier in Argentina between 2003 to 2024. . .	3
1.2	Glacier Remote Sensing Using Sentinel-2 and Mapping Glacier Ex- tents and Surface Facies, and Comparison to Landsat 8 [15]	5
1.3	Overview of Methodology	9
2.1	Landsat 8 data of central Karakoram from [18]	13
2.2	Sample image retrieved from an SAR dataset from [19]	13
4.1	Overview of the methodology for comparison and our model finalisation.	30
4.2	Sample SAR tiles from Sjörgen and Crane glaciers [11]	32
4.3	Modalities of the NIRD Dataset.	33
4.4	Landsat Sample from HKH dataset.	35
4.5	Sample images from the HKH Glacier Mapping Dataset and their preprocessed patches.	36
4.6	Sample of images from Cityscapes dataset.	37
4.7	Sample images from the CDW-Seg dataset.	38
4.8	Geometric augmentations performed on the dataset 1.	42
4.9	Geometric augmentations performed on the dataset 2.	43
4.10	Rendered images of each channel from the common modalities.	45
4.11	Architecture for Unet Segmentation	46
4.12	Architecture for Resnet-50 Segmentation	47
4.13	Architecture for EfficientNet Segmentation	48
4.14	Architecture for VGG-19 Segmentation	49
4.15	Architecture for DeeplabV3 Segmentation	50
4.16	Architecture for Segformer Segmentation	51
4.17	Overall architecture of Vision Mamba (ViM), showing patch embed- ding, stacked Mamba blocks, [CLS] token with positional embed- dings, and classification head. Adapted from [55].	52
4.18	Simple Architecture of Glacier-Seg	60
4.19	High-level Architecture of Glacier-Seg	61
5.1	GLOPS Comparison of all models.	64
5.2	Structure of the input tensor in NCHW format used for glacier im- agery segmentation.	65
5.3	Involution-based overlapping patch embedding for adaptive feature extraction.	66
5.4	Mamba state-space block for efficient sequence mixing.	67
5.5	Hierarchical MiT-style backbone (four stages).	69

5.6	SegFormer decoder head with projection, fusion, and classification operations.	70
5.7	DropPath regularisation for robust training.	70
5.8	NIRD Dataset Ablation mIOU Lineplot	72
5.9	NIRD Dataset Region Ablation mIOU Lineplot	73
6.1	Comparison of Model Performance on Dataset 1	83
6.2	Predicted Mask by Glacier-Seg on CaFFe Dataset.	84
6.3	Comparison of Model images Performance on Dataset 2	85
6.4	Predicted Output of Glacier-Seg on NIRD Dataset.	87
6.5	HKH Dataset Multiclass Metrics Overview	88
6.6	Predicted mask by Glacier-Seg on HKH Dataset.	89
6.7	CityScape Dataset mIOU & mDice	91
6.8	ROC–AUC curve for the Glacier–Seg model on the HKH dataset. . .	94
6.9	Normalized confusion matrix of the Glacier–Seg model on the HKH dataset.	94
6.10	ROC–AUC curve for the Glacier–Seg model on the NIRD dataset. . .	95
6.11	Normalised confusion matrix of the Glacier–Seg model on the NIRD dataset.	95
6.12	ROC–AUC curve for the Glacier–Seg model on the NIRD dataset. . .	96
6.13	Normalized confusion matrix of the Glacier–Seg model on the NIRD dataset.	96
6.14	HKH Dataset Vs NIRD Dataset Metrics Overview	96
6.15	Grad-Cam of UNet on dataset 2.	100
6.16	Side-by-side comparison of VGG19	101
6.17	Side-by-side comparison of EfficientNet	102
6.18	Output of EfficientNet on dataset 2	103
6.19	Grad-CAM and Saliency Map of DeepLabV3.	103
6.20	Side-by-side comparison of SegFormer	104
6.21	Crane Retreat Trend from CaFFe.	107
6.22	Vertical glacier-retreat forecasting pipeline integrating segmentation, pixel and edge extraction, temporal tracking, and forecasting.	109
7.1	Architecture for Unet Segmentation	122
7.2	Architecture for Resnet-50 Segmentation	123
7.3	Architecture for EfficientNet Segmentation	124
7.4	Architecture for VGG-19 Segmentation	125
7.5	Architecture for DeeplabV3 Segmentation	126
7.6	Architecture for Segformer Segmentation	127
7.7	Overall architecture of Vision Mamba (ViM), showing patch embedding, stacked Mamba blocks, [CLS] token with positional embeddings, and classification head. Adapted from [55].	129
7.8	High-level Architecture of Glacier-Seg	131
7.9	Overview of the methodology for comparison and our model finalisation.	133

List of Tables

1.1	Key challenges encountered and mitigation strategies.	11
2.1	Summary of Glacier Monitoring Studies	23
3.1	Proposed 12-month research timeline for Project P2.	28
4.1	Summary of Datasets Used in This Study	40
4.2	Comparison of Deep Learning Models for Glacier Segmentation . . .	54
4.3	Hyperparameter Modifications for Dataset 1	56
4.4	Hyperparameter Modifications for Dataset 2	56
4.5	Hyperparameter Modifications for Dataset 3	57
4.6	Hyperparameter Modifications for Dataset 4	57
4.7	Hyperparameter Modifications for Dataset 5	57
4.8	Hyperparameter Configuration of Glacier-Seg Model	62
5.1	Complexity comparison of custom hybrid models and baseline seg- mentation architectures (input size: 256×256 , batch size = 32). . . .	64
5.2	Dataset ablation results on NIRD (best epoch, mIoU, and runtime). .	72
5.3	Best epoch, mIoU, and training time across glacier regions of NIRD Dataset.	73
5.4	Comparison of best mIoU across glacier regions between our model and GlaViTU.	74
5.5	Blockwise ablation results (efficiency metrics) for MiT-B0 hybrid backbones.	75
5.6	Blockwise ablation results (accuracy metrics) for MiT-B0 hybrid back- bones.	75
5.7	Layerwise ablation results. S = stage, D = depth.	76
5.8	One-way ANOVA results across model variants. All differences are highly significant ($p < 0.001$).	78
5.9	One-way ANOVA results across dataset ablations on NIRD. All ad- justed p -values remain below 0.01, confirming a significant effect of data composition on performance.	78
5.10	Paired t -tests for dataset ablations (Bonferroni-corrected p -values). Each variant is compared against the full_2000 multimodal baseline.	78
5.11	One-way ANOVA across layerwise ablation variants. All differences remain significant after Bonferroni correction.	79
5.12	Paired t -tests comparing layerwise ablation variants against the best early-stage configuration (Mamba-S12-D2).	79
6.1	Performance summary of six segmentation models on Datasets 1 and 2.	82

6.2	Comparison of segmentation performance between Glacier-Seg and Segformer on Single-Class HKH.	87
6.3	Comparison of performance between EfficientNet, SegFormer, and Glacier-Seg on multi-class segmentation.	89
6.4	Overall Segmentation Results on Dataset 4 and Benchmark Summary	90
6.5	Per-Class IoU and Dice Scores	91
6.6	Overall segmentation results on Dataset 5.	92
6.7	Per-Class IoU and Dice Scores of Dataset 5 for All Models.	92
6.8	Summary of Model Comparison Results Across All Datasets	93
6.9	ONNX CPU Inference Performance Comparison	99
6.10	Parameter and training speed of all segmentation models.	105
6.11	Glacier Retreat Analysis Based on Area Change Over Time	108

Chapter 1

Introduction

1.1 Background

Glaciers are one of the most vital components of the Earth’s cryosphere, acting as natural reservoirs that release water during dry seasons. This is crucial for agriculture, hydropower, and sustaining ecosystems, especially in South Asia—namely the river basins of the Indus, where the water from melted glaciers contributes to around one-fourth of the annual water withdrawn for irrigation [1]. Covering about 10% of the Earth’s land surface, glaciers also play a critical role in regulating the global climate by reflecting sunlight and maintaining the Earth’s energy balance [2]. In addition, they are the habitat for one-third of the entire terrestrial species, hosting about 50% of the global biodiversity hotspots on the planet [3], thus contributing to the composition and dynamics of terrestrial ecosystems. Moreover, glaciers are responsible for regulating local microclimates. However, with global temperatures rising at an alarming rate, there has been an ongoing retreat and thinning of glaciers across various regions around the world [4]. Their high surface albedo plays a vital role in reflecting incoming radiation and moderating planetary energy balance. Surface albedo is the measure of how much sunlight a surface reflects compared to how much it absorbs. The albedo effect can be expressed as:

$$\alpha = \frac{R}{I} \tag{1.1}$$

where α is the albedo, R the reflected radiation, and I the incident radiation. Declining glacier cover reduces α , accelerating warming feedbacks.

Consequently as glaciers melt, they contribute to rising sea levels putting coastal communities in danger. The rise in sea levels also exacerbates storm surges, and threatens to submerge low-lying areas. From 2006 to 2016, the mass loss of mountain glaciers added 335 ± 144 Gt annually to global sea levels [5]. Furthermore, glacier melt affects water availability, especially for regions that depend on seasonal meltwater for drinking, agriculture, and energy production. As glaciers shrink, water shortages could become more frequent and severe, impacting billions of people globally [6]. Moreover, the rapid decline in glaciers will also disrupt the ecological network of microorganisms, causing the local extinction of these species and severely

affecting alpine ecosystems [4].

As a result, evaluating and estimating changes in glaciers is important for predicting future conditions, especially in areas where both the environment and people rely on them. Glacier mapping provides valuable insights into how glaciers evolve over time, revealing trends in ice loss, glacial retreat, and changes in glacial mass. Continuous and accurate mapping is essential for forecasting natural hazards, such as glacial lake outburst floods, and for water resource management in glacier-fed river basins.

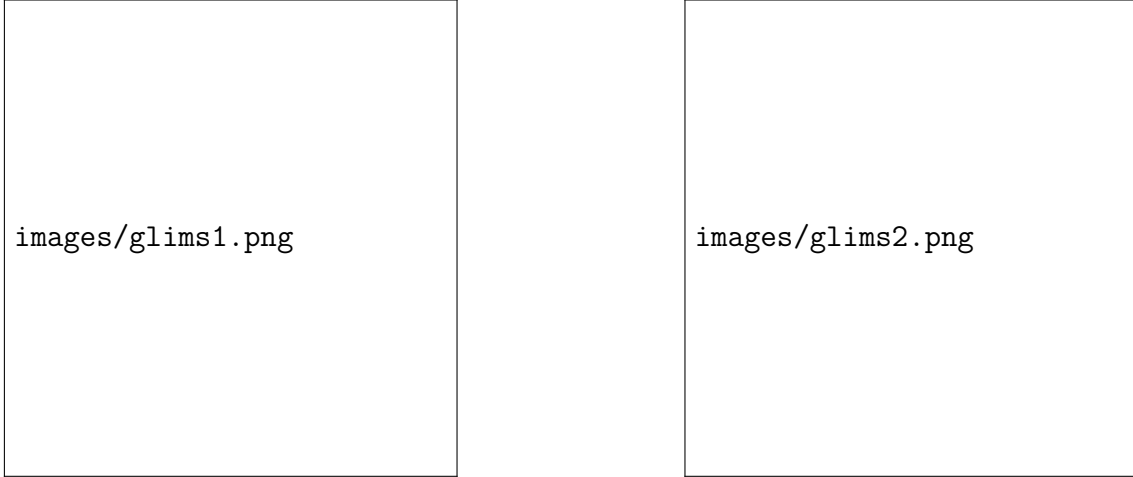


Figure 1.1: Retreat of the Upsala Glacier in Argentina between 2003 to 2024.

However, patterns of glacial changes are not always apparent, as there can be a plethora of variability in glacier subsystem responses to the climate change [7]. In order to understand glacial fluctuations and regional trends, scientists globally have been relying on manual field-surveys for near accurate and reliable mapping of glacial boundaries. This is, unfortunately, extremely challenging due to the harsh climate conditions on mountains [8]. As an outcome, digital mapping has been gaining traction as more and more remote sensing data (RS) is used to automate the crucial task of glacial mapping. The usage and ubiquity of Artificial Intelligence (AI) marks a significant advancement in glaciology. Available studies demonstrate AI models are outperforming traditional statistical models, with greater and notable efficacy [9]. Within the past few years, advanced methods based on computer vision have been actively employed for analysing glacier-covered regions. While CNN based architectures have procured good results, the studies have numerous limitations pertaining to clouds and debris distorting the visibility of calving fronts. In addition to small datasets, the inventories also only have outlines of naked ice glaciers which produce less accurate predictions for debris-covered glaciers [10].

1.2 Rationale of the Study

This study is therefore motivated by the need to develop a **robust, efficient, and generalisable framework** for glacier segmentation that can operate under real-world constraints. Specifically, the study seeks to:

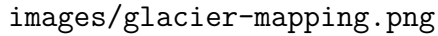
- Address the scarcity and heterogeneity of glacier datasets by leveraging multimodal inputs, combining SAR, optical, and DEM data to improve boundary detection and segmentation accuracy.
- Evaluate the strengths and limitations of state-of-the-art architectures, including CNNs, EfficientNets, Vision Transformers, and Mamba-based sequence models, to identify configurations that balance accuracy and computational efficiency.
- Investigate hybrid architectural strategies, such as the combination of early-stage Mamba state-space blocks with later-stage Transformers, to maintain high-resolution spatial temporal modelling while preserving global contextual understanding to capture seasonal changes and patterns an calving events.
- Develop lightweight and scalable solutions suitable for deployment in environments with limited GPU memory and computational power, without sacrificing segmentation quality.
- Overcoming limitations of single modal input data and images while applying multimodal fusion strategies combining SAR, optical and elevation data.
- Enable future applications in glacier change forecasting, hazard prediction, and water resource management by producing reliable and interpretable glacier maps. Such segmentation also supports climate modelling and regional vulnerability assessments.

In summary, the rationale of this study lies in addressing the dual challenges of **sparse, noisy data** and **computational limitations**, while pushing the boundaries of automated glacier mapping. By systematically evaluating and integrating modern deep learning model architectures within a multimodal framework, this research aims to provide a practical, high-accuracy solution for glacier segmentation and monitoring, contributing both to cryospheric science and operational decision-making.

1.3 Problem Statement

Despite the critical role of glaciers in climate regulation, water resources, and ecosystem services, the current understanding of glacier dynamics remains limited due to significant technical barriers in automated real-time mapping. Existing datasets, such as CaFFe [11] and NIRD (multimodal fusion of SAR, optical, and DEM data) [12], are often sparse, noisy, or restricted to seasonal observations, with many calving fronts recorded only a few times per year [13]. This scarcity of high-quality annotations poses significant challenges for automated segmentation models, increasing the risk of overfitting and limiting generalisation to unseen glacial regions. In many cases, temporal data is also difficult to acquire. Such limitations in annotations over long time ranges can weaken the training of deep learning models, making them more susceptible to overfitting and poor generalisation.

Predicting when a glacier may be pushed out of balance compared to its state of seasonal change is also difficult due to the aforementioned data constraints. In addition to dataset limitations, geophysical images and remote sensing data are highly sensitive to noise and atmospheric and surface variability: cloud cover, seasonal snow, shadows, and debris-covered ice all distort the visibility of glacier boundaries [14]. Alongside these, variable atmospheric conditions and temporal changes on the ground further complicate AI-based frameworks in interpreting the images, leading to failures in accurately mapping complex glacial topographies, particularly calving fronts. While using high-resolution images and larger batch sizes might mitigate some errors, this comes at the cost of extremely high computational demands. In resource-constrained research contexts, convolutional neural networks (CNNs) often struggle with heterogeneity, failing to consistently capture fine-grained glacier front boundaries. Moreover, processing high-resolution satellite images for accurate segmentation imposes a heavy computational burden on deep learning models, requiring substantial resources.



images/glacier-mapping.png

Figure 1.2: Glacier Remote Sensing Using Sentinel-2 and Mapping Glacier Extents and Surface Facies, and Comparison to Landsat 8 [15]

To bridge this gap, our research focuses on developing lightweight models that balance fidelity and computational feasibility, as demonstrated through evaluations of architectures like U-Net, VGG-19, EfficientNet, ResNet-50, DeepLabV3+, and SegFormer on remote sensing satellite datasets. The results highlight multi-modal fusion’s superiority, underscoring the imperative need for efficient models in cryospheric applications. Consequently, Vision Transformers (ViTs) and hybrid Transformer-Mamba architectures, such as our proposed **Glacier-Seg**, can be explored for glacier image segmentation. For a crucial field like glacier mapping, ViTs offer potential for accurate calving-front segmentation under the constraints of sparse, noisy data and resource-limited environments. Therefore, this study investigates whether ViTs and hybrid CNN-Transformer-Mamba architectures can achieve accurate calving-front segmentation under the twin constraints of sparse, noisy data and limited computational resources for both satellite imagery and drone-based glacier mapping missions.

1.4 Research Objective

The overarching aim of this research is to develop and evaluate deep learning architectures that enable precise, efficient, and generalisable glacier segmentation. By leveraging the complementary strengths of Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), and Mamba-based sequence models, this study seeks to overcome existing limitations in data quality, model generalisation, and computational efficiency.

Specifically, the objectives are:

- **Benchmark SOTA Models:** Evaluate and compare state-of-the-art segmentation architectures—including U-Net, EfficientNet, ResNet50, DeepLabV3, and SegFormer—to identify the most effective approaches for glacier mapping under diverse conditions.
- **Analyse Data Modality Impact:** Investigate the effect of single-modal versus multimodal data inputs (SAR, optical, DEM) on segmentation performance, assessing how feature fusion enhances accuracy and robustness.
- **Evaluate Computational Efficiency:** Measure training time, parameter counts, and FLOPs across architectures to identify models suitable for deployment in computationally constrained environments.
- **Establish Performance Baselines:** Quantify segmentation accuracy for various glacier types using metrics such as mean Intersection-over-Union (mIoU) and Dice coefficient, providing a comprehensive benchmark for future research.
- **Design Lightweight Hybrid Architectures:** Develop distilled ViT or hybrid CNN–Transformer models that maintain high spatial fidelity while reducing computational overhead.
- **Enable Temporal Modelling:** Incorporate Mamba-based sequence models in Glacier-Seg to support potential time-series analysis for mapping glacier change, building toward real-time monitoring as outlined in future work.
- **Optimised Model:** Build optimised models for low resources under constrained environments such as drone-based mapping, enabling real-world use in field stations and edge devices, with statistical analysis (ANOVA, t-tests) confirming significance in efficiency metrics.
- **Support Cryospheric Forecasting:** Utilise the resulting segmentation models to track glacier changes over time, enabling hazard prediction (e.g., glacial lake outburst floods) and informing water resource management in glacier-fed regions, with extensions to more versatile satellite scenes planned for future investigations.

By addressing these objectives, this research aims to deliver a principled, scalable framework for glacier segmentation that balances accuracy, generalisability, and efficiency, bridging critical gaps in cryospheric monitoring and automated environmental analysis.

1.5 Methodology in Brief

The methodology adopted in this research is designed to balance data richness with computational efficiency, ensuring that glacier mapping models can operate effectively across heterogeneous satellite inputs. At a high level, the workflow progresses through sequential stages of **data acquisition, preprocessing, base modelling, ablation study, model finalisation, training, testing, and performance analysis** (Figure 1.3). This structured pipeline provides both clarity and adaptability, enabling robust evaluation of state-of-the-art architectures.

Data were acquired primarily from open-source datasets which consists of Landsat and Sentinel 1 images, complemented by Synthetic Aperture Radar (SAR) imagery and Digital Elevation Models (DEM). These multimodal sources capture complementary aspects of glacier systems: spectral responses in optical bands, surface roughness and penetration in radar, and topographic context in DEMs. Preprocessing steps included geometric corrections, radiometric normalisation, and targeted photometric augmentations, ensuring consistent inputs across spatial scales and acquisition conditions. This stage also addressed data imbalance and missing modalities, enhancing resilience in downstream learning.

Modelling followed a hybrid design philosophy: Convolutional Neural Networks (CNNs) were leveraged for their strength in capturing localised spatial features, such as glacier boundaries or crevasse textures. Transformers, through their self-attention mechanism, supplied the capacity to integrate global context across large satellite tiles. In parallel, Mamba-based sequence models were introduced to exploit temporal continuity, enabling the system to extend from static segmentation toward **time-series** forecasting of glacier change. This integration reduces redundancy compared to variations of vision transformers alone, with parameter savings of approximately 18%, while still retaining strong representational capacity.

An **ablation study** was carried out at three levels: dataset variations, layerwise feature contribution, and blockwise architectural modifications. These experiments isolated the impact of each component, clarifying trade-offs between complexity, generalisation, and accuracy. The insights guided the design of the **final model**, which incorporated the most effective combination of CNN encoders, transformer attention modules, and temporal sequence handling.


The finalised model was trained end-to-end using a composite loss function comprising **Binary Cross-Entropy (BCE)** and **Dice-based losses**, ensuring a balanced optimisation of pixel-level accuracy and region-level coherence. The **AdamW optimiser** was employed with decoupled weight decay to promote stable convergence and improved generalisation. Learning rates were adaptively adjusted based on validation loss plateaus to prevent premature convergence and overfitting.

Comprehensive evaluation was conducted using a suite of complementary metrics—**Intersection over Union (IoU)**, **Dice coefficient**, **Precision**, **Recall**, and **Pixel Accuracy (PA)**—to provide a holistic view of segmentation quality. IoU and Dice measured the spatial overlap between predicted and ground truth

glacier masks, while Precision and Recall quantified the model’s reliability in detecting true glacier pixels versus background misclassification. Pixel Accuracy provided an overall measure of correct classifications across the entire image domain. Together, these metrics captured both the fine-grained boundary delineation and global segmentation reliability of the model.

In addition to performance metrics, **statistical significance testing** was performed to ensure that observed performance differences between competing architectures were not attributable to random variation. Paired **t-tests** were applied across cross-validation folds to validate the robustness of comparative results. Furthermore, **confidence intervals** for IoU and Dice scores were computed to quantify metric stability and reinforce the statistical reliability of findings.

Beyond quantitative evaluation, the training pipeline incorporated **explainability mechanisms**—including **Grad-CAM** and **saliency maps**—to visualise model attention across critical glaciological regions such as debris-covered zones, accumulation areas, and calving fronts for multimodal datasets. These interpretability analyses provided valuable insight into whether the learned spatial representations aligned with physically meaningful glacier dynamics, thereby enhancing both the scientific interpretability and trustworthiness of the model.



images/simplemethodology.png

Figure 1.3: Overview of Methodology

1.6 Scopes and Challenges

This research focuses on glacier segmentation using multimodal remote sensing data, primarily targeting binary classification (glacier vs. non-glacier). The study leverages approximately 1200 multimodal patches of size 256×256 pixels, drawn from three different and diverse datasets. Analysis is restricted to glaciers, excluding seasonal snow and permafrost, to maintain task specificity and reduce label ambiguity. All experiments are conducted on hardware constrained to 16 GB GPU memory, reflecting realistic computational limits in academic and field-based environments.

1.6.1 Scope of the Study

- **Spatial Focus:** Glacier regions are selected from varied geophysical contexts, including mountainous terrain and polar environments, to assess model generalisability across heterogeneous landscapes.
- **Data Modalities:** Inputs include Synthetic Aperture Radar (SAR), optical imagery, and Digital Elevation Models (DEM), allowing for the investigation of multimodal fusion strategies.
- **Model Types:** Architectures evaluated range from conventional CNNs (U-Net, ResNet, EfficientNet) to transformer-based models (SegFormer, Vision Transformers), alongside hybrid CNN–Transformer designs with Mamba state-space blocks.
- **Performance Metrics:** Segmentation accuracy is measured using mean Intersection-over-Union (mIoU), Dice coefficient, and pixel-level accuracy. Computational efficiency is assessed through parameter counts, memory usage, and training time.

1.6.2 Technical Challenges

Table 1.1: Key challenges encountered and mitigation strategies.

Challenge	Impact on Model Performance	Mitigation Strategies
SAR speckle noise	Degrades feature extraction	Hybrid SAR+optical+DEM fusion
Debris-covered glaciers	Missed calving fronts and misclassification	Multimodal fusion; edge-preserving augmentations
Data scarcity	Overfitting risk; poor generalisation	Data augmentation; patch sampling; self-supervised pretraining
Compute constraints	Limited batch sizes and epoch runs	Lightweight ViT–Mamba hybrids; mixed precision training
Imbalanced and multi-class regions	Reduced accuracy on minor classes	Class weighting; targeted oversampling; loss function tuning

The research is constrained by several intertwined challenges. Firstly, SAR imagery exhibits speckle noise that diminishes the ability of CNNs to detect subtle ice boundaries, while optical images can be obscured by clouds or shadows. Furthermore, models were also made to learn from the debris-covered images to simulate realistic learning. Debris-covered glaciers further complicate segmentation, requiring sophisticated multimodal integration. Data scarcity, particularly for annotated calving fronts, increases overfitting risk, limiting generalisation across unseen glacier regions.

Hardware limitations restrict batch sizes, the number of training epochs, and full hyperparameter sweeps, imposing practical trade-offs between accuracy and computational feasibility. Additionally, hybrid CNN–Transformer–Mamba architectures, although highly accurate, entail slightly higher FLOPs compared to Vision Mamba, SegFormer, and EfficientNet, affecting computational complexity.

Finally, imbalanced multi-class scenarios (e.g., minor debris or supraglacial lakes) remain a persistent challenge, with conventional loss functions biased towards dominant classes. Strategies such as weighted loss and class-aware sampling mitigate these issues but cannot fully eliminate them. Together, these constraints define the operational boundaries of this research and contextualise the results presented in subsequent chapters.

Chapter 2

Literature Review

2.1 Preliminaries

There are nineteen glacier regions around the world [16], with four being located in the Asian continent. Amongst the regions, the most prominent glaciers are those in South West and South East Asia, followed by Central Europe. Most of the studies done using deep learning algorithms are based on the South West region of Asia. Initial automatic techniques used spectral indices such as the Normalised Difference Snow Index (NDSI) or Normalised Difference Water Index (NDWI) derived from optical sensors (e.g., Landsat and Sentinel series). These indices exploited reflectance differences between snow, ice, and surrounding rock but were highly sensitive to illumination, shadow, and debris cover. Classical machine-learning algorithms—Support Vector Machines (SVMs), Random Forests, and K-means clustering—were later applied to incorporate additional contextual and textural features. While these models achieved moderate improvements, they depended heavily on handcrafted feature design and lacked spatial generalisation across glaciers of varying morphology or climatic conditions. In this section, we fundamentally focus on the research that has been done on glacier mapping and evolution, highlighting the classifier models alongside datasets and common limitations experienced while evaluating the performance of those algorithms.

2.2 Review of Existing Research

2.2.1 Datasets

Datasets are an essential component in training AI models, as they help the models learn various glacier features, delineate glacier boundaries, and accurately predict different glacier types. As a result, a variety of remote sensing technologies are used to capture data from Earth’s glacial regions, where extensive fieldwork is difficult to conduct. One of the most popular choices is the use of satellite and multispectral images obtained from satellites such as Landsat and Sentinel as shown in Figure 2.1. Studies on deep learning classifiers often rely on object-based image analysis (OBIA) [17]. Multispectral images capture data across several distinct wavelengths, both within and beyond the visible spectrum, such as green, blue, red, near-infrared, shortwave infrared 1, and shortwave infrared 2, providing spectral, textural, and topographical data about glacial regions. These images are gathered from the Google

Earth Engine (GEE), provided by the United States Geological Survey (USGS). Despite the variety of information available in such images, models still struggle to identify glaciers due to the spectral similarity between glaciers and the surrounding environment [17].

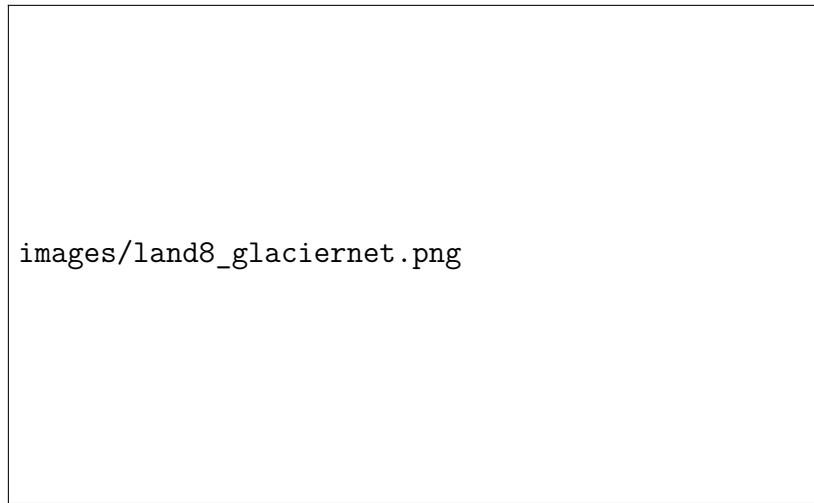


Figure 2.1: Landsat 8 data of central Karakoram from [18]

Additionally, complex weather patterns and cloudy regions affect the training phase of deep learning models. SAR (Synthetic Aperture Radar) technology provides more reliable data in this case [19] as an active microwave remote sensing, the radars are able to penetrate clouds and gather clearer information regarding the glacial regions (see Figure 2.2). However, SAR images are susceptible to noises and sensitive to shadows. It also struggles with precise interpretation of surface textures when compared to multispectral optical images.

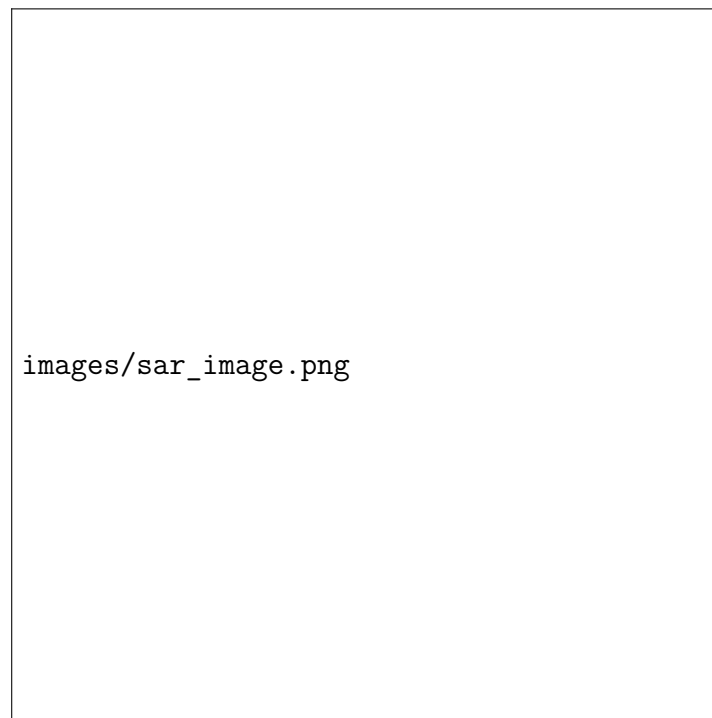


Figure 2.2: Sample image retrieved from an SAR dataset from [19]

2.2.2 Traditional Machine Learning Approaches

Traditional machine learning approaches have established a strong methodological foundation for automated glacier monitoring using remotely sensed data. Over the past two decades, advancements in image processing, spatial analysis, and statistical modelling have enabled increasingly accurate detection and delineation of glacier features across diverse environmental conditions. The integration of remote sensing, Geographic Information Systems (GIS), and Global Positioning Systems (GPS) has been instrumental in facilitating efficient glacier analysis in remote and inaccessible regions [20]. The synergy among these technologies has allowed for precise geolocation, temporal monitoring, and morphological assessment of glaciers over large spatial extents. Reported accuracies for optical remote sensing typically range from 70% to over 90%, while Synthetic Aperture Radar (SAR) achieves between 74% and 79%. Furthermore, the use of Digital Elevation Models (DEMs) has been shown to mitigate topographic shadow effects in mountainous regions, substantially improving spatial precision and the reliability of glacier boundary estimation.

Remote sensing technologies have been recognised as particularly valuable for monitoring glaciers due to their rapid response to climatic fluctuations compared with other cryospheric components [21]. Multi-platform observation systems such as Landsat, SRTM, MODIS, ICESat, ASTER, GRACE, GLIMS, and GTN-G have collectively enabled long-term, high-resolution, and wide-area glacier monitoring. The integration of laser altimetry, DEMs, and SAR data has further enhanced the ability to detect subtle elevation variations, which are essential indicators of glacier mass balance and retreat dynamics. These developments have significantly contributed to the establishment of comprehensive global glacier inventories and time-series analyses.

Significant progress in satellite image segmentation has underpinned many of the early successes in glacier delineation. A range of algorithmic paradigms—including edge detection, thresholding, region-growing, fuzzy logic, partial differential equation-based, and artificial neural network-based techniques—have been developed to improve classification fidelity [22]. While no single algorithm has proven universally optimal across all imagery types, bilateral filtering combined with Canny edge detection has consistently achieved higher Peak Signal-to-Noise Ratio (PSNR) and lower Mean Squared Error (MSE) than Gaussian filtering. Such approaches have demonstrated strong generalisability across datasets from PNG, THR, QuickBird, and Landsat sensors, yielding smoother boundaries and more distinct glacier outlines.

The advent of high-resolution LiDAR and DEM data has catalysed a paradigm shift from analogue to digital geomorphological mapping [23]. This transition has greatly enhanced data accessibility and analytical precision, facilitating detailed classification of glacial landforms such as moraines, ice tongues, and debris-covered regions. When applied to LiDAR and satellite-derived DEM datasets, models such as Random Forest (RF) and Support Vector Machine (SVM) have achieved high classification accuracies, confirming their effectiveness in terrain-informed glacier feature extraction. These approaches have also enabled more reliable differentiation of glacier zones under complex topographic and illumination conditions.

The utilisation of ensemble-based classifiers has further advanced glacier mapping by combining spectral, topographic, and kinematic information. In particular, the Random Forest algorithm has been applied to integrate multi-source features—including spectral bands, glacier velocity fields, and elevation parameters derived from Landsat-8, DEMs, and ITS-LIVE datasets—yielding overall accuracies between 97.42% and 97.60% and Kappa coefficients ranging from 0.95696 to 0.9624 for the Eastern Pamir region [24]. Such results underscore the robustness and adaptability of ensemble learning in heterogeneous glacial environments.

More recent progress has incorporated microwave and polarimetric SAR data for multi-class glacier feature discrimination. The use of PolSAR datasets from ALOS PALSAR-1 and PALSAR-2 in the Siachen Glacier region has demonstrated the superior capability of deep learning-augmented classifiers over conventional SVM models [25]. The GF-DNN framework achieved an overall accuracy of 91.17% and a Kappa coefficient of 0.88, compared with 80.49% and 0.80 respectively for SVM. These findings illustrate the potential of polarisation-sensitive methods in distinguishing between debris-covered and clean ice zones, thereby addressing one of the major limitations of optical-only classification approaches.

2.2.3 CNN-Based Models for Glacier Segmentation

Convolutional Neural Networks (CNNs) have been widely adopted for glacier mapping and front delineation due to their ability to learn complex spatial and spectral patterns from multi-sensor data. Their applications can be organised thematically into architectural innovations, attention mechanisms, ensemble strategies, SAR adaptations, comparative benchmarking, and hybrid frameworks.

U-Net-based architectures have served as a fundamental backbone for glacier segmentation. A 29-layer U-Net comprising 3×3 ReLU convolutions, 0.2 dropout, and 2×2 max pooling for downsampling and upsampling was implemented to enhance glacier front extraction from Landsat 5, 7, and 8 imagery [13]. The use of mirrored image augmentation minimised edge-related errors, resulting in a mean deviation of 96.3 m from ground-truth glacier fronts, outperforming conventional Sobel-based edge detection.

SegNet-inspired architectures, exemplified by GlacierNet, incorporated multi-channel inputs and post-processing steps such as thresholding and hole filling to achieve classification accuracies exceeding 99% in the Karakoram and 98% in Nepal [18]. Its successor, GlacierNet2, employed multi-model learning for terminus, ablation, and accumulation zones, increasing Intersection over Union (IoU) from 0.8599 to 0.8839 [26], thereby improving boundary delineation and generalisation.

Attention mechanisms have been integrated into CNNs to enhance feature extraction. The Spectral-Spatial U-Net (SSUNet4) combined spectral attention via global average pooling and spatial attention through 1×1 convolutions to emphasise critical spatial features in Sentinel-2 data [17]. This architecture achieved an overall

accuracy of 96.87% and a Kappa coefficient of 0.79, surpassing Random Forest and Multi-Res CNN approaches. Self-attention modules were further applied to U-Net architectures for glacial lake segmentation, preserving spatial context via skip connections and mitigating feature overload [27]. This approach produced 78.69% true positives, a mean IoU of 0.76, and an AUC of 85.03%, outperforming standard U-Net performance metrics.

Ensemble learning has been utilised to improve segmentation robustness and accuracy. An ensemble of twenty DeepLabV3+ sub-models processed Sentinel-2 spectral indices (R, G, B, NDSI, NIR, SWIR) to map snow and glaciers in alpine regions [28]. Overall accuracy reached 0.8861, outperforming traditional snow index-based approaches, which yielded 0.7353. Multi-phase frameworks, such as the Convolutional Sparse Coding (CSC) approach combined with an adapted VGG16 model, have been applied to Greenland marine glaciers [29]. Phase one generated coarse labels across seven classes, while phase two refined predictions using a Multi-Layer Perceptron (MLP), achieving F1-scores of 94–96% and mean and median front deviations of 56.17 m and 24.7 m, respectively.

Extensive benchmarking studies have evaluated multiple CNN-based frameworks across Landsat-8, ALOS DEM, and GLIMS datasets [30]. DeepLabV3+ consistently achieved the highest accuracy (0.9684), whereas Glacier-CNN offered a balanced trade-off between model complexity and performance, achieving an F-measure of 0.9247. Such evaluations highlight the strengths and limitations of different architectures when applied to heterogeneous glacial environments.

2.2.4 Emerging Transformer and Hybrid Architectures

Recent advances in Vision Transformers (ViTs) and hybrid architectures have introduced new paradigms for glacier and cryosphere mapping by leveraging long-range dependencies and global contextual understanding. A hierarchical encoder-decoder structure known as SegFormer was proposed to improve semantic segmentation through efficient representation learning [31]. By combining a transformer-based encoder with a lightweight multilayer perceptron (MLP) decoder, SegFormer achieved mean Intersection over Union (mIoU) scores of 51.0 on ADE20K, 82.2 on Cityscapes, and 51.3 on COCO-Stuff datasets. The architecture demonstrated that high-resolution segmentation could be maintained with significantly reduced parameter counts, marking a substantial improvement over conventional CNNs in computational efficiency.

Transformer-based models have also been adapted for glacial mapping tasks. A hybrid CNN–Transformer framework, termed GlaViTU, was designed for automated glacier delineation using multi-modal inputs including optical, DEM, and SAR data [32]. Covering approximately 9% of global glaciers, the model achieved an average IoU of 0.894, outperforming DeepLabV3+ with a ResNet-101 backbone. Performance analysis indicated an IoU of 0.90 in clean ice regions and 0.75 in debris-covered areas, with the integration of SAR imagery contributing to notable accuracy improvements in shadowed or cloud-affected zones. These findings highlight the viability of multi-modal transformer frameworks for capturing fine-grained spatial

variations in complex glacial terrains.

The emergence of state-space and frequency-based vision models has further diversified the remote sensing landscape. A Vision Mamba-based hybrid U-Net, referred to as CVMH-UNet, was introduced to integrate multi-frequency fusion for remote sensing segmentation tasks [33]. This approach achieved state-of-the-art performance with lower computational complexity than traditional transformer models, suggesting that structured state-space representations may offer an efficient alternative for large-scale environmental mapping applications.

Advances in efficient transformer design have also been demonstrated through the SegViT model, which employs attention-to-mask decoding with a compact architecture [34]. Evaluations on ADE20K, COCO-Stuff-10K, and PASCAL-Context datasets yielded mean IoU values of 55.2%, 50.3%, and 65.3%, respectively, while the compressed variant of SegViT reduced computation costs by approximately 40%. Such progress underscores the ongoing optimisation of transformer-based architectures for both accuracy and efficiency, positioning them as the next frontier in remote sensing and glaciological segmentation research.

2.3 Summary of Key Finding

Amongst traditional supervised AI methods, random forest (RF), artificial neural networks (ANN), and support vector machines (SVMs) are commonly used [10], while the most widely used deep learning model in glacial mapping is Convolutional Neural Networks (CNN). There are also prominent architectures based on CNN such as U-Net, ENVINet5 and GlacierNet which have shown competitive performance. An overview for the deep learning classifiers for glacial mapping and different remote sensing technologies can be found in Table 2.1.

Ref	Summary	Dataset	AI Model	Performance
[13]	U-Net architecture with a 29-layer deep neural network with 3×3 ReLU convolutional layers, 0.2 Dropout for regularization, and 2×2 MaxPooling for downsampling and upsampling. Mirrored images produced minimal error outperforming traditional pre-processing techniques such as Sobel filters.	Landsat 5, 7, 8 images	U-Net	Mean deviation of 96.3 m from the true fronts.

[17]	Spectral and Spatial attention module is integrated in U-net where the former focuses on the non-linear relationship between bands in order to get global spatial information via global average pooling. Spatial attention focuses on more particular regions using 1×1 convolutional layer.	Sentinel-2	SSUnet and RF, Multi-Resolution for comparison	Multi-Res (OA: 93.89%, Kappa: 0.56), RF (OA: 91.84%, Kappa: 0.55), SSUNet (OA: 96.87, Kappa: 0.79)
[18]	Based on Karakoram and Nepal Himalayan glaciers, a new deep learning approach called GlacierNet CNN is proposed which is based on SegNet. The input layer comprised 17 channels instead of the usual 3 channels for RGB images. Additional refinement steps such as region size thresholding and hole filling were also leveraged.	Landsat 8, ALOS DEM	GlaicerNet	Over 99% in Karakoram and 98% in Nepal.
[26]	Updated from GlacierNet to a multi-model learning architecture, GlacierNet2 can estimate glacier terminus, ablation, and snow-covered accumulation zones. It was tested on glaciers in central Karakoram, northern Pakistan, from September and October 2016, when glacier mapping is most feasible.	Landsat 8, ALOS DEM	Glaicer-Net2	IOU of GlacierNet2: 0.8839, IOU of GlacierNet: 0.8599
[35]	ENVINet5 uses a mask-based encoder-decoder architecture such as conv. layers and dimensionality reduction with 1×1 convolutions. ENVINet-Multi leverages the spectral and spatial properties of input datasets to make multi-class speculation. The errors from ENVINet-5 are very high due to the data's spatial resolution and parameter selection during training.	Landsat-8 OLI	ENVINet5 and ENVINET-Multi	ENVINet5 (OA: 91.89% and Kappa: 0.8778), ANN (OA: 88.38% and Kappa: 0.8241))

[36]	Deep learning models trained on imagery from Earth and Mars on Snapdragon and Movidius Myriad X processors onboard the International Space Station (ISS) are being benchmarked for image classification, image segmentation, and spectral super-resolution. The study presents a model for image segmentation and super resolution for spectroscopy, trained on Mars NavCam imagery, pre-trained on ImageNet for all the models.	Mars MSL NavCam Imagery and Label, Imagenette	Mars HiRISENet and Mars MSLNets, UNET, DeepLabV3+, Snapdragon NPU, Imagenette	The benchmarking results show that these low SWaP processors have only small errors and up to 10x speed improvement. All error rates are small using U-Net showed compatible results.
[21]	Highlights how glaciers respond more rapidly to climate variations compared to other ice bodies, making them crucial for understanding environmental changes. Here remote sensing methods allow extensive monitoring of glaciers over large areas which is not feasible with traditional field measurements due to logistical challenges.	LANDSAT, SRTM, MODIS, ICESat, ASTER, GRACE, GLIMS, GTNG	Fuzzy Classifications, Simple Glacier Flow Models, Geodetic Methods, Texture Analysis, SAR, DEMs	Provides long term, high resolution imagery, generates digital elevation models (DEMs) and precise changes using laser altimetry and SAR.
[20]	Comprehensive review of how remote sensing, geographic GIS and Global Positioning Systems GPS have been utilized in glaciology. Here the efficiency of such systems can be achieved by proper studying of the glaciers particularly in remote or inaccessible areas. The integration of DEMs (Digital Elevation Models) helps to compensate for topographic shadows in mountainous regions.	Transient Snowline (TSL), LANDSAT, SPOT, SAR, GPS DATA, DEMs	Digital image processing, GIS model, Radar Interferometry	Remote sensing (accuracies ranging from 70% to over 90%), SAR (accuracies of 74% to 79%)
[22]	Image segmentation techniques divided in six types which are edge based segmentation, threshold segmentation, region based segmentation, fuzzy theory based segmentation, partial differential based segmentation and ANN based segmentation since no single segmentation algorithm is applicable to all types of satellites images.	PNG Image Dataset, THR satellite images, Quickbird, Landsat	Canny algorithm and Harris operator, Bilateral filters, neural network-based segmentation	Bilateral filters used with Canny for edge detection showed better performance in terms of PSNR and MSE than Gaussian filters. This resulted in smoother images and better edge detection.

[23]	Focuses on use of remote sensing techniques specifically on high resolution LiDAR and dems for classification of glacial landforms. It emphasises the transition from analog to digital methods in geomorphological mapping which has increased availability of datasets and improved mapping accuracy.	LiDAR data, Satellite-derived DEMs	Random Forest (RF), Support Vector Machine (SVM)	High accuracy in classifying glacial landforms; effective methods demonstrated.
[37]	Discusses the role of glaciers in climate change and challenges in accurately mapping and monitoring glacier dynamics. Highlights that glaciers are retreating globally due to rising temperatures and changing precipitation patterns, contributing significantly to sea-level rise and impacting water resources, ecosystems, and communities.	Optical Imagery, SAR, DEMs, Thermal Infrared Data	Random Forest, DeepLabv3+, GlacierNet, UNet, Hybrid Convolutional Transformer Model	Random Forests showed good correspondence with manually derived glacier outlines. DeepLabv3+, demonstrated superior performance in mapping glaciers.
[32]	Automating a global glacier mapping using Deep Learning model and open satellite imagery data consisting of high resolution images.	Optical, DEM, SAR, newly released multi-modal-benchmarking covering 9% of global glaciers.	Glacier-Vision Transformer-U-Net (GlaViTU), a hybrid CNN Transformer architecture.	Average IoU: 0.894 (GlaViTU), outperforming DeepLabv3+/ResNet-101. Best results in clean ice regions 0.90 IoU; debris-rich areas 0.75 IoU. SAR data improved accuracy across all regions.
[33]	Efficient semantic segmentation of remote sensing images using Vision Mamba and multi-frequency feature fusion.	Multiple benchmark RS datasets; high-resolution imagery with complex spatial patterns.	CVMH-UNet with HVSS-Block and MFMS-Block, combining convolutional and state-space modeling.	Achieved state-of-the-art accuracy with lower computational complexity than Transformer-based models.

[34]	Semantic segmentation using plain Vision Transformers with attention-to-mask decoding.	ADE20K, COCO-Stuff-10K, PASCAL-Context datasets.	SegViT with ATM module and Shrunk structure (QD + QU) for efficient inference.	mIoU: 55.2% (ADE20K), 50.3% (COCO-Stuff-10K), 65.3% (PASCAL-Context). Shrunk version saves 40% computation.
[38]	Land-use/land-cover mapping using high-resolution remote sensing images with CNN-ViT collaboration.	Two benchmark RSI datasets with heterogeneous spatial features.	CTSeg with dual-branch encoder, pixel-wise and channel-wise knowledge distillation, and multiscale decoding.	Outperforms state-of-the-art segmentation models in accuracy and generalization on small-volume datasets.
[24]	Multiple Features were used for glacier mapping(Easter Pamir), which included various spectral bands, movement velocity and topographic parameters. These features were combined to train the Random Forest algorithm for effective classification.	Landsat-8 images, DEMs and ITS-LIVE	RF	Scheme 1(OA:97.42%, Kappa:0.95696), Scheme 2(OA:97.43%, Kappa:0.9598) and Scheme 3(OA:97.60%, Kappa:0.9624)
[28]	Uses ensemble of CNN, specially DeepLabV3+ which is then compared with the traditional SI methods which are prone to errors due to similarities in color between snow and clouds. A novel ensemble model which uses 20 DeepLabV3+ sub-models was deployed. The dataset consist of diverse alpine locations.	Sentinel-2 imagery(R, G, B, NDSI, NIR, SWIR)	DeepLabV3+ and SI	Ensemble Model(Overall: 0.8861) vs SI(Overall: 0.7353)
[39]	Modified U-Net was used here which proved more accuracy on classifying glacier. Average of deviation of 78 meters in training and 108m test area were found. Bigger input tiles were used and 4 input channels were used instead of gray-scale imagery.	Sentinel-1 synthetic aperture radar (SAR)	Modified U-Net	Ice Precision: 0.92, Ice Recall: 0.98, Water Precision: 0.97, Water F1-Score: 0.94

[30]	A comparison among Glacier-CNN, Res-UNet, Mobile-UNet, DeepLabV3+, R2UNet and FC DenseNet. Shows that Glacier-CNN and DeepLabv3+ performs best. Glacier-CNN is moderately light weight and less compute exhaustive where as DeepLabV3+ complex architecture gives the best performance.	11 Landsat-8 Bands, ALOS DEM, GLIMS	Glacier-CNN, Res-UNet, Mobile-UNet, DeepLabV3+, R2UNet and FC DenseNet	DeepLabV3+: Accuracy (0.9684), Glacier-CNN: F-measure (0.9247)
[27]	An improved U-Net network utilizing a self-attention mechanism to address challenges in extracting glacial lakes from Landsat-8 imagery. U-Net uses skip connections which with provides better spatial context. The self-attention mechanism ensures that information does not overload during segmentation. It emphasises small features, improving the segmentation of glacial lakes.	Landsat-8(3, 5, 6)bands.	U-NET and self-attention U-Net	Standard U-Net(TP: 72.90%, MIOU: 0.71, AUC: 81%), Self-Attention U-Net(TP: 78.69%, MIOU: 0.76, AUC: 85.03%)
[29]	The Research uses a two-phase approach CSC on an adapted VVG16-CNN in-order to classify Sentinel-2 images of Greenland's marine glaciers. In First Phase the CNN is used to classify the images into seven classes which are used as training labels in Phase 2. Phase 2 uses MLP of cCNN for pixel-level Classification.	Sentinel-2(Copernicus Open Access Hub)	CNN-Supervised Classification (CSC) adapted from VGG16 Architecture	F1: 94% (Helheim Glacier) and 96%(Jakobshavn Isbrae and Store Glacier). Mean deviation: 56.17m, Median Deviation: 24.7m
[40]	A new composite machine learning method, the RF-CNN model, which combines (RF) and (CNN) techniques. CNN extracts the deep features, spatial and textural information. The shallow features are combined and extracted by the RF. The RF classifier is subsequently applied to combined features to enhance segmentation and classification.	Landsat-8(OLI), TIRS, ASTER and GDEM	RF, CNN and RF-CNN	RF: 97.60% (Eastern Pamir), 99.31% (Nyainqentanglha), CNN: 99.06% (Nyainqentanglha), RF-CNN: 98.14% (Eastern Pamir)

[25]	New Novel method used PolSar Data. Microwave is used to detect terrain and snow debris. GF-DNN and SVM used as detection method. GF-DNN model is used for multi-classification of Glacier features, while SVM is a ML model that classify data by finding optimal hyperplane separating different classes. GF-DNN produced the best result.	ALOS PALSAR-1 and 2	GF-DNN and SVM	Siachen Glacier: GF-DNN (OA: 91.17%, Kappa: 0.88, P: 0.93, R: 0.90, F1: 0.91), SVM (OA: 80.49%, P: 0.77, R: 0.79, F1: 0.78);
[25]	New Novel method used PolSar Data. Microwave is used to detect terrain and snow debris. GF-DNN and SVM used as detection method. GF-DNN model is used for multi-classification of Glacier features, while SVM is a ML model that classify data by finding optimal hyperplane separating different classes. GF-DNN produced the best result.	ALOS PALSAR-1 and 2	GF-DNN and SVM	Siachen Glacier: GF-DNN (OA: 91.17%, Kappa: 0.88, P: 0.93, R: 0.90, F1: 0.91), SVM (OA: 80.49%, P: 0.77, R: 0.79, F1: 0.78);
[31]	Introduced SegFormer, a hierarchical transformer encoder that produces multi-scale features without positional encodings. A lightweight MLP decoder then fuses these features, achieving state-of-the-art accuracy with a markedly smaller parameter count.	ADE20K, Cityscapes, COCO-Stuff	SegFormer (MiT-B0-B5) + MLP decoder	ADE20K: 51.0 mIoU; Cityscapes: 82.2 mIoU; COCO-Stuff: 51.3 mIoU

Table 2.1: Summary of Glacier Monitoring Studies

2.4 Research Gap

Despite aforementioned advancements, a clear research gap persists in the development of lightweight, efficient hybrids tailored specifically for glacier segmentation under real-time constraints. Most recent ViT applications focus on related tasks like glacial lakes or landforms, with limited validation on glacier-specific datasets for calving fronts or debris-covered zones. For example, while Transformer-based U-Nets emphasise accuracy gains, they often overlook parameter efficiency and deployment on edge devices, leading to high FLOPs unsuitable for field monitoring. Mamba-based state-space models, which offer linear complexity for sequence handling, remain underexplored in cryosphere applications, with no systematic integration into ViT hybrids for temporal glacier dynamics. This gap is evident in reviews noting that AI in glacier studies predominantly relies on CNNs, with ViTs only recently emerging but not optimised for sparse, noisy multimodal data in real-time scenarios.

To bridge these gaps, our research introduces Glacier-Seg, a lightweight Mamba–Transformer–CNN hybrid designed to balance high segmentation accuracy with computational efficiency for both satellites and drones. By incorporating involution-based patch embedding, Mamba mixers, and a MiT-style hierarchical backbone with a transformer-decoder, Glacier-Seg provides enhanced spatial precision while maintaining parameter counts and inference times. This framework addresses the pressing need for efficient multimodal glacier segmentation, enabling reliable and near real-time monitoring of glacier dynamics in data-scarce, resource-limited environments.

Chapter 3

Project Planning and Impacts

3.1 Final Specifications and Requirements

The success of this project depended on the integration of suitable hardware, software, and algorithmic components that collectively enable efficient and accurate glacier segmentation and experimentation. The following subsections summarise the key requirements.

3.1.1 Data Requirements

- **High-resolution multimodal imagery:** Synthetic Aperture Radar (SAR), optical (multispectral), and Digital Elevation Models (DEM).
- **Pre-processing:** Radiometric and geometric normalisation, co-registration across modalities, and photometric augmentations to reduce bias in heterogeneous glacier surfaces.
- **Data pipeline:** Scalable, with support for modality fusion and patch-based tiling at multiple resolutions.

3.1.2 Computational Resources

- **Hardware:** GPU-enabled infrastructure with at least 16GB VRAM per device, sufficient for training medium-scale transformer-based architectures.
- **Software:** Frameworks such as PyTorch and TensorFlow, complemented by geospatial libraries (GDAL, Rasterio, GeoPandas) for data preparation and handling.
- **Efficiency:** Support for distributed training and mixed-precision (FP16) computation to reduce training time and memory consumption.

3.1.3 Model Requirements

- **Hybrid backbone:** Convolutional Neural Networks (CNNs) for local features, Transformers for global context, and Mamba-based linear sequence models for efficient long-range dependency modelling.

- **Lightweight design:** Memory-efficient modules (e.g., distilled Vision Transformers, Involution stems) suitable for deployment in constrained environments.

3.1.4 Training Specifications

- **Optimisation:** Careful tuning of learning rates, batch sizes, and regularisation strategies (DropPath, stochastic augmentation).
- **Overfitting mitigation:** Advanced augmentation strategies (spectral shifts, random cropping, brightness/contrast adjustments) to enhance generalisation from sparse datasets.
- **Evaluation:** Dual loss functions (Binary Cross-Entropy + Dice) and standard segmentation metrics (mIoU, Dice, pixel accuracy, precision, recall).

3.2 Societal Impact

Glacier mapping through AI-driven approaches represents more than a technological advancement; it embodies a step towards climate justice, particularly for vulnerable populations living downstream of major glacial systems. In regions such as South Asia, where over a billion people depend on glacier-fed rivers, timely and accurate insights into glacier dynamics can determine the difference between stability and humanitarian crisis. For example, the Indus Basin receives nearly 25% of its annual irrigation supply from glacial meltwater [1]. As climate change accelerates glacial retreat, this critical resource is becoming increasingly volatile and difficult to predict.

AI-based segmentation systems significantly enhance predictive capacity for glacier-related hazards, including glacial lake outburst floods, which are sudden and destructive events. Such systems can form the basis of early-warning frameworks and disaster risk reduction strategies, directly benefiting millions of people who live in high-risk zones but often lack robust infrastructure or resources [41]. This capacity is especially critical in fragile mountain communities, where response times are short and exposure to hazards is high.

Beyond disaster mitigation, the adoption of AI-based tools promotes inclusivity in environmental research. Automated, scalable, and cost-effective mapping techniques reduce reliance on hazardous and resource-intensive field expeditions, enabling broader participation from academic institutions and organisations in the Global South [42]. In addition, the insights derived from these systems can inform evidence-based policymaking in water governance, climate adaptation, infrastructure planning, and regional development. In this way, the societal benefits extend beyond risk management to the fostering of resilience and equitable access to environmental knowledge.

3.3 Environmental Impact

This research contributes to more sustainable methods for observing and conserving the Earth’s cryospheric zones. By combining satellite imagery with non-invasive machine learning approaches, it avoids direct human interference, thereby preserving the ecological integrity of fragile environments while enabling continuous monitoring [43]. Traditional glaciological surveys often demand direct site access, which can disturb delicate ecosystems and exacerbate environmental stress. In contrast, AI-driven analysis of synthetic aperture radar (SAR) and optical imagery provides a scalable, efficient, and non-intrusive alternative.

Accurate delineation of glacier boundaries is also vital for refining global climate models. High-precision glacier mapping generates empirical data that inform projections of sea-level rise, hydrological variability, and cryospheric energy balance. For example, quantifying glacier retreat is directly linked to albedo effects, where reduced reflectivity accelerates warming [44]. Thus, this research not only supports observational science but also strengthens the predictive capacity of climate modelling initiatives.

Finally, the visualisation of glacier dynamics in accessible formats contributes to environmental awareness among policymakers, stakeholders, and the wider public. By translating complex scientific outputs into intuitive representations, AI-based tools can foster broader engagement with climate issues, thereby promoting collective responsibility for environmental stewardship and climate change mitigation.

3.4 Ethical Issues

The study strictly adheres to ethical standards in research and data usage. Satellite imagery was sourced from publicly available datasets (e.g., Landsat, Sentinel), ensuring compliance with licensing and privacy regulations. No personally identifiable information is used or collected.

Ethical considerations were also very seriously maintained regarding the responsible communication of glacier mapping results. Inaccurate predictions or misinterpretation of hazard zones could have significant societal consequences; therefore, all outputs are presented with associated uncertainty metrics. Furthermore, the development of AI models prioritises transparency, reproducibility, and explainability to ensure that stakeholders and local communities can make informed decisions based on reliable data.

3.5 Project Management Plan

The project followed a structured and phased approach to ensure the timely achievement of milestones and deliverables. The workflow encompassed dataset acquisition, preprocessing, model development and training, performance evaluation, and results visualisation.

A one-year timeline was adopted to monitor progress and ensure alignment with research objectives. The project was divided into sequential phases, with each stage contributing directly to the final outcomes.

Table 3.1: Proposed 12-month research timeline for Project P2.

Months	Milestones and Tasks
1–3	Conduct comprehensive literature review; acquire and assess relevant datasets; define preprocessing pipelines and primary methodology based on previous works.
4–8	Perform data preprocessing, including co-registration of modalities, radiometric and geometric normalisation, handling of missing modalities, and augmentation; conduct exploratory analysis of different SOTA models.
9–10	Develop and train models using convolutional neural networks (CNNs), transformer-based architectures, and hybrid ViT–Mamba variants; perform initial hyperparameter tuning and optimisation. Evaluate model performance across datasets;.
11–12	Compile results and statistical analysis; finalise documentation, thesis chapters, and supporting materials. Analyse masks to translate glacier retreats and changes throughout years.

3.6 Risk Management

The project identified potential risks along technical, operational, and environmental dimensions, with mitigation strategies as follows:

- **Technical Risks:** GPU failures, memory limitations, or prolonged training cycles. *Mitigation:* Use mixed-precision training, lightweight model variants, checkpointing, and distributed computation where possible.
- **Data Risks:** Incomplete or corrupted satellite imagery, class imbalance, and noisy inputs. *Mitigation:* Data augmentation, denoising filters, and multi-modal fusion to improve robustness.
- **Operational Risks:** Delays in model convergence or limited access to computational resources. *Mitigation:* Structured project timeline, prioritisation of lightweight architectures, and fallback to smaller batch experiments.
- **Environmental Risks:** Misinterpretation of glacier hazard zones. *Mitigation:* Use of explainability tools, cross-validation, and collaboration with domain experts for validation of results.

3.7 Economic Analysis

The economic implications of AI-driven glacier mapping are twofold: cost savings and societal benefit. Traditional glaciological field surveys are resource-intensive,

requiring personnel, transport, and logistics in hazardous terrain. AI-based approaches reduce these operational costs significantly while enabling continuous, scalable monitoring.

The adoption of drone-assisted glacier mapping further addresses the temporal and spatial limitations of satellite imagery. UAVs can be deployed on-demand to capture high-resolution data at critical time intervals, such as during seasonal melt periods, without the need for expensive repeat satellite acquisitions. This dual-source system—combining wide-scale satellite monitoring with targeted drone missions—provides a more economically balanced approach to continuous cryospheric observation. As a result, the development of lightweight, resource-efficient models becomes crucial to ensure that both satellite and drone-derived data can be processed locally or in near real time, enabling timely and scalable cryospheric monitoring within economically and environmentally sustainable limits.

Investment in computational resources—GPUs, storage, and cloud services—is offset by reduced fieldwork expenses and improved decision-making. Moreover, the ability to forecast glacial hazards such as GLOFs and ice avalanches can prevent economic losses in agriculture, hydropower, and infrastructure by enabling timely disaster mitigation and water resource planning.

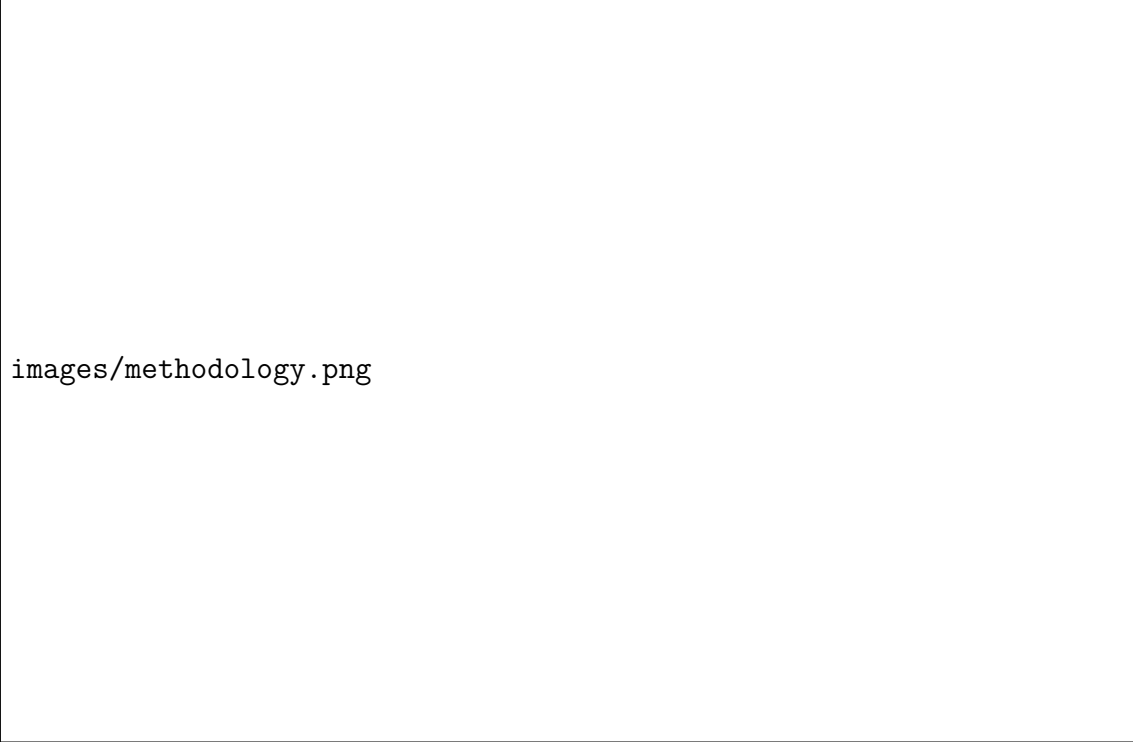
Hence, by combining cost-efficient AI architectures with flexible drone–satellite observation strategies, this research ensures that glacier monitoring becomes not only technologically advanced but also economically viable, scalable, and impactful for long-term environmental resilience.

Chapter 4

Research Methodology

4.1 Methodology Overview

This section outlines the complete methodology employed in our study, including dataset preparation, augmentation strategies, model selection, training configurations, and evaluation protocols. For better resolution and visibility, check **Appendix A**.



images/methodology.png

Figure 4.1: Overview of the methodology for comparison and our model finalisation.

4.1.1 Datasets

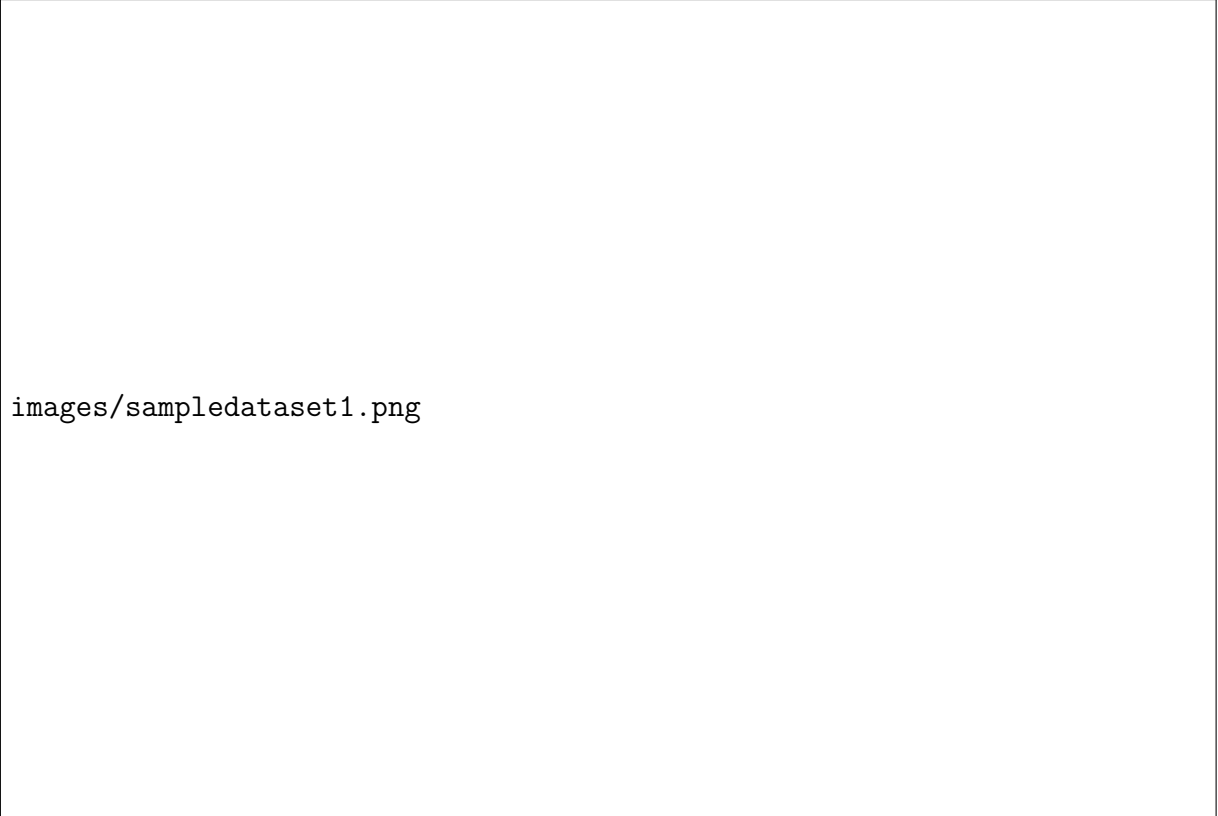
To evaluate the performance of state-of-the-art (SOTA) models for glacier mapping and segmentation, five publicly available datasets were employed in this research.

Dataset 1: Calving Front Definition Dataset (CaFFe) The first dataset, titled the Calving Front Definition Dataset (CaFFe) [11], was obtained from the PANGAEA data repository. It offers a comprehensive collection of Synthetic Aperture Radar (SAR) imagery accompanied by labelled ground truth data, intended to facilitate the automation of glacier calving front delineation—an essential task for monitoring ice loss and assessing glacier mass balance. This dataset encompasses SAR images from seven marine-terminating glaciers distributed across different regions. Five of these glaciers are located in Antarctica: Crane, Dinsmoore-Bombardier-Edgeworth, Mapple, Jorum, and Sjørgen Inlet. The remaining two are Jakobshavn Isbræ in Greenland and Columbia Glacier in Alaska.

Multiple SAR acquisitions were made for each glacier across various time points, forming time series that span from 1995 to 2020. These images exhibit varying spatial resolutions, owing to the diversity of satellite sources, including Sentinel-1, TerraSAR-X, TanDEM-X, ENVISAT, ERS-1 and ERS-2, ALOS PALSAR, and RADARSAT-1.

The dataset is organised into four main directories:

- **sar_images**: PNG files used for training and testing.
- **fronts**: Binary masks delineating the calving front.
- **zones**: Multi-class semantic segmentation masks representing landscape regions (e.g., ice, rock, ocean, SAR shadow).
- **bounding_boxes**: Spatial annotations enclosing the dynamic calving front extent, which constrain detection to relevant areas.



images/sampledatsaset1.png

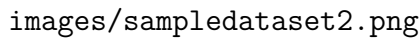
Figure 4.2: Sample SAR tiles from Sjörgen and Crane glaciers [11]

Each image file is annotated with metadata in its filename, including glacier name, acquisition date, sensor type, resolution, and a quality factor ranging from 1 to 6. For generalisation analysis, the data set was partitioned into training and testing subsets: Training glaciers include Crane, Dinsmoore-Bombardier-Edgeworth, Jorum, Sjörgen Inlet, and Jakobshavn Isbr, while Mapple and Columbia glaciers were reserved for testing. Although the dataset supports both direct front segmentation and derivation from zone masks, this study utilised only the SAR images and their corresponding zone annotations. In total, the dataset has 681 SAR image tiles.

Dataset 2: NIRD Dataset The second dataset, compiled by [12] and retrieved from the NIRD Research Data Archive, integrates multiple sources of geospatial data, including optical imagery, Synthetic Aperture Radar (SAR) data and digital elevation model (DEM) inputs. It spans six geographically and climatically diverse regions: the European Alps (ALP), High Mountain Asia (HMA), Indonesia (LL; lower latitudes), New Zealand (NZ), the Southern Andes (SA), and Scandinavia (SC).

The optical imagery was sourced from Landsat 7, Landsat 8, and Sentinel-2 satellites, capturing six spectral bands: blue, green, red, near-infrared (NIR), and two shortwave infrared (SWIR) bands. The SAR component consists of σ^0 -calibrated amplitude images from ENVISAT and Sentinel-1, acquired in both ascending and descending orbital paths. Elevation and terrain slope information were derived from ALOS and Copernicus DEMs. All input data were resampled to a spatial resolution of 10 metres to ensure consistency across modalities.

The dataset was intentionally designed to incorporate a wide range of glacier types to support the development of generalisable and robust glacier mapping models. Specifically, it includes clean-ice glaciers as well as those covered with debris and vegetation, situated in a variety of environmental settings—alpine, polar, tropical, maritime, and continental. This deliberate inclusion addresses the need for resilient glacier segmentation models capable of performing accurately across diverse climatic and geographic conditions.



images/sampledataset2.png

Figure 4.3: Modalities of the NIRD Dataset.

In terms of global representation, the dataset encompasses approximately 9% of the world’s glaciers which entails covering around 19,000 individual glaciers and accounts for roughly 7% of the global glacierised area, excluding the Greenland and Antarctic ice sheets. The temporal range is also significant, extending from 1988 (the earliest data from Antarctica) to 2020 (the most recent data from Svalbard). This broad spatial and temporal distribution enhances the dataset’s value for training and evaluating glacier monitoring models over time and across regions.

To facilitate model development and validation, the dataset was randomly partitioned into three subsets: 60% for training (613 tiles), 20% for validation (205 tiles), and 20% for testing (209 tiles), resulting in a total of 1,027 glacier image tiles.

Each tile contains the following six data modalities, offering complementary insights into glacier conditions:

- **DEM (Digital Elevation Model):** Elevation data from ALOS and Copernicus DEMs, capturing surface topography. DEMs aid in identifying glacier flow direction, accumulation zones, and frontal steepness, particularly useful for segmenting complex terrain.
- **Cross-Polarized SAR (cross_pol_sar):** SAR backscatter captured in the cross-polarized channel (HV or VH). This modality enhances sensitivity to volume scattering in ice or snow, supporting the detection of debris cover and snowpack variability.
- **Co-Polarized SAR (co_pol_sar):** SAR backscatter in the co-polarized channel (HH or VV), more responsive to surface roughness and dielectric properties. It provides essential contrast between glacier ice, bedrock, and water, particularly in radar-dominant regions.
- **Outlines:** Binary raster masks delineating glacier extent. These masks serve as ground truth references for evaluating segmentation models.

- **InSAR (`in_sar`):** Interferometric SAR-derived deformation or phase data. This modality captures glacier dynamics, including flow velocity and displacement, particularly relevant for detecting calving fronts and active glacier zones.
- **Bright-Dark Outlines (`bright_dark_outlines`):** Enhanced visual outlines extracted from optical imagery. These masks emphasize bright ice surfaces and serve as auxiliary annotations to improve segmentation accuracy, especially in debris-covered glaciers.

A small number of image groups from the Antarctic region contained thermal data; however, due to their limited occurrence, these thermal tiles were excluded during preprocessing.

Dataset 3: HKH Dataset The Hindu Kush Himalayas (HKH) [45] is the third glacier mapping dataset providing a extensive resource for glacier monitoring and Machine Learning (ML) research objectives in fields of model training. This Dataset is one of the most extensive benchmark datasets designed specifically for large-scale glacier mapping and segmentation in the Hindu Kush Himalaya region which is often referred to as the “Third Pole” due to its enormous ice reserves outside the polar regions. This area encompasses eight countries: Afghanistan, Pakistan, India, Nepal, Bhutan, Bangladesh, China, and Myanmar, covering a vital cryospheric zone that sustains downstream water resources, agriculture, and hydropower for more than a billion people. The data set was meticulously assembled to facilitate advanced machine learning and computer vision methods for glacier detection, delineation, and monitoring, emphasizing a differentiation between clean-ice glaciers and debris-covered glaciers, two classes that are particularly important for comprehending the dynamics of glacier retreat in context of climate change.

The dataset collection is built from Landsat 7 ETM+ satellite imagery (remote sensing) acquired between 2002 and 2008, aligned with glacier boundaries from the International Centre for Integrated Mountain Development (ICIMOD). The raw satellite imagery is provided in 35 tiles, each covering an area of approximately 6×7.5 km, and all images are rescaled to a uniform 30-meter spatial resolution. To enrich the dataset beyond raw spectral bands, the creators incorporated a total of 15 input channels. These include Landsat bands B1–B8, which span the visible, near-infrared (NIR), shortwave infrared (SWIR), and thermal ranges; the BQA quality assessment band, which encodes pixel-level quality flags for cloud cover, saturation, and sensor errors; three commonly used spectral indices—NDVI (Normalised Difference Vegetation Index), NDSI (normalised Difference Snow Index), and NDWI (Normalised Difference Water Index)—which add vegetation, snow, and water discrimination features; and topographic information from the SRTM (Shuttle Radar Topography Mission) elevation and slope layers, originally collected at 90 m resolution but upsampled to match the 30 m grid. This rich multimodal feature set makes the HKH dataset highly suited for training modern deep learning architectures.

One of the defining aspects of the dataset is the high-quality glacier annotations. ICIMOD glacier outlines were digitised and aligned with the Landsat imagery, providing accurate ground-truth labels. These labels distinguish between clean glaciers, which are bright and debris-free, and debris-covered glaciers, which are more difficult




image.png

Figure 4.4: Landsat Sample from HKH dataset.

to detect due to surface rock and sediment obscuring the ice. The annotations are released in two complementary formats: (i) vector shapefiles, compatible with GIS workflows, and (ii) rasterized masks aligned pixel-by-pixel with the input images, making them directly usable for deep learning segmentation models.

To make the dataset machine-learning-ready, the imagery was further divided into 14,190 cropped patches, each of size 512×512 pixels \times 15 channels, stored as NumPy arrays (.npy). Each patch is paired with a corresponding binary or categorical mask (512×512) marking glacier and non-glacier pixels. Accompanying metadata includes Landsat scene IDs, geographic coordinates, acquisition dates, and glacier coverage statistics, ensuring traceability and enabling reproducibility. This design not only supports efficient training in deep learning frameworks such as PyTorch and TensorFlow but also makes the dataset suitable for cross-validation and benchmarking. The dataset is approximately 29.4 GB in size and is openly hosted on multiple platforms including Google Cloud, Azure, and AWS for global accessibility. The Landsat and SRTM imagery are in the public domain, while the annotations are released under the Community Data License Agreement (permissive variant), encouraging open collaboration.

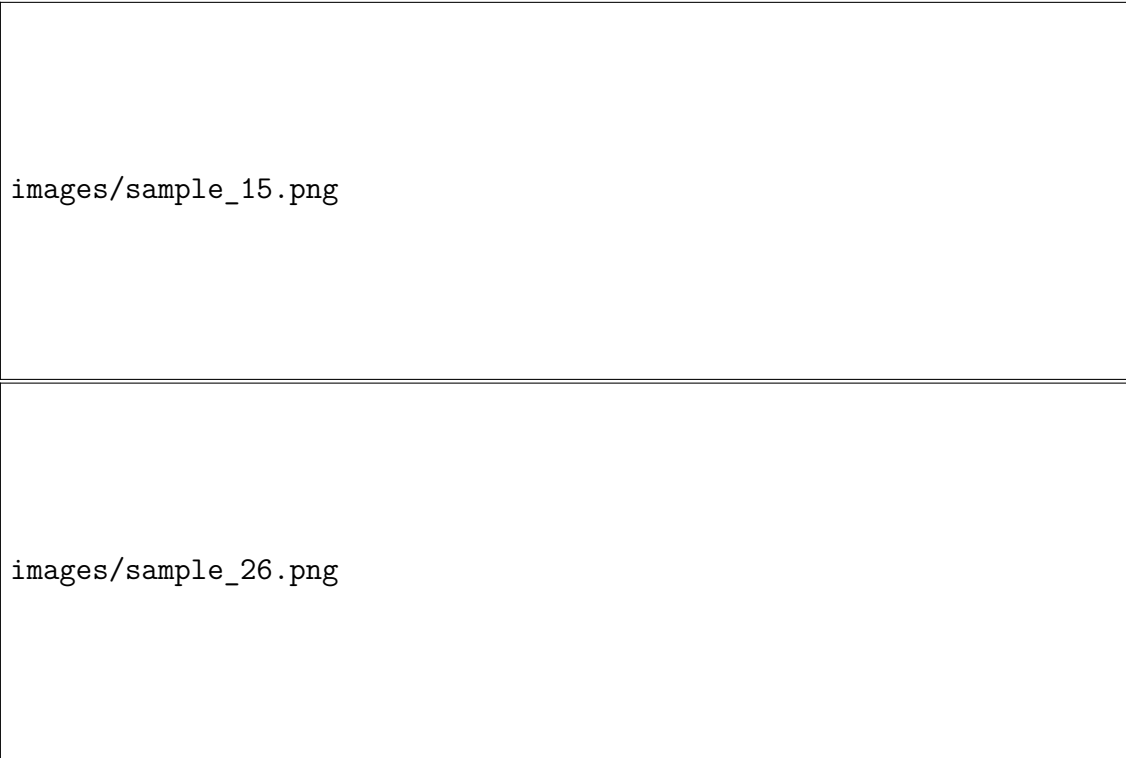


Figure 4.5: Sample images from the HKH Glacier Mapping Dataset and their pre-processed patches.

Dataset 4: CityScapes Dataset The Cityscapes dataset [46] is a large-scale, high-resolution benchmark dataset created for advancing semantic understanding of urban street scenes. Designed to reflect the complexities of real-world driving environments, it has become a foundational dataset in computer vision, particularly in the fields of semantic segmentation, instance segmentation, and panoptic segmentation. Cityscapes contains imagery collected across 50 cities in Germany and neighbouring countries, capturing diverse weather, seasonal, and illumination conditions, which makes it especially valuable for testing robustness and generalization of deep learning models.

The dataset comprises 5,000 finely annotated images, split into 2,975 for training, 500 for validation, and 1,525 for testing, each with high-quality pixel-level labels. These annotations include 30 visual classes, of which 19 are used for official semantic segmentation benchmarks. The classes represent urban scene categories such as road, side-walk, building, vegetation, person, rider, car, bus, truck, train, and more. Annotations are provided at the instance level, distinguishing not only between classes but also between different objects of the same class (e.g., individual cars or pedestrians). This enables both semantic and instance-level segmentation. Additionally, Cityscapes provides coarse annotations for 20,000 additional images, which offer broader coverage at lower labelling precision, useful for pretraining and semi-supervised approaches.

images/image.png

images/image (2).png

images/image (4).png

Figure 4.6: Sample of images from Cityscapes dataset.

The images are captured with a consistent stereo camera system mounted on a car, ensuring geometric consistency and allowing depth estimation tasks. Each image has a resolution of 2048×1024 pixels, which is significantly higher than many prior segmentation datasets, challenging models to handle fine-grained details such as traffic signs, lane markings, and pedestrians. Beyond semantic masks, the dataset includes polygonal annotations, instance IDs, and scripts for evaluation following standardized metrics such as mean Intersection-over-Union (mIoU).

We used the Cityscapes dataset in this research because it provides a controlled, well-annotated benchmark that is both complex and diverse, making it ideal for testing segmentation architectures before applying them to specialized domains like glacier mapping. The dense annotations and variety of urban conditions serve as a stress test for segmentation backbones, ensuring that models such as SegFormer, Vision Transformers, or Mamba-augmented architectures can capture multi-scale context and boundary precision. Furthermore, Cityscapes’ wide adoption in the computer vision community allows for direct comparison with state-of-the-art methods, strengthening the validity of our experimental results.

Dataset 5: Synthesised Dataset The Construction and Demolition Waste Object Detection Dataset (CDW-Seg) [47] is a purpose-built benchmark dataset for class-wise segmentation. The publicly available dataset images were captured at authentic construction sites. It consisted of skip bins (large containers) filled with a heterogeneous mixture of CDW materials. This ensures the visual conditions are realistic, including clutter, overlapping items, varying lighting, occlusions, and different materials in physical contact.

images/2022_0002.png

images/2022_0003.png

images/2022_0004.png

Figure 4.7: Sample images from the CDW-Seg dataset.

Initial raw images were reviewed for clarity; images with blur or low resolution were excluded to preserve label quality. The final curated set includes 5,413 manually annotated waste objects, each labeled by class at polygon (pixel-level) resolution. The CDW dataset consists of 10 classes, reflecting common CDW materials found in demolition/construction waste streams: *concrete, fill dirt, timber, hard plastic, soft plastic, steel, fabric, cardboard, plasterboard, skip bin (container itself)*.

Each annotated object is associated with one of these categories, and all pixel-wise segmentation of these has been provided. 2,492,021,189 pixels (sum over masks) is the total pixel coverage of the dataset. High-resolution images are often given on the order of 3000×4000 pixels (or similar) to maintain detail in noisy scenes.

Annotations are provided in semantic segmentation format which means that each pixel in a mask is allocated a class label. The dataset is organized with directories including:

- `Original_Images_and_Annotation_Files`

- `Ground_Truths_VOC_Format` (Pascal VOC style)
- `Ground_Truths_COCO_Format` (COCO-style JSON + masks)

The dataset is accessible in both VOC and COCO formats, making it compatible with a wide range of deep learning tools and model designs. The authors divided the data into three subsets: training (75%), validation (15%), and test (10%). This allows for rigorous evaluation and generalization analysis. The dataset (≈ 4.26 GB) is freely available on Figshare under an open license (Construction Engineering category) to encourage reuse. The images, annotations, and code (for training, formatting, evaluation) are available in their GitHub/project repositories (e.g., **SAM2-Adapter-CDW**) to facilitate reproducibility.

Our research is centered on segmentation architectures (SegFormer, Involution, Mamba, hybrid ViTs, etc.) and their robustness across different domains (glaciers, urban scenes, waste, etc.). The CDW-Seg dataset is incredibly beneficial since it depicts a noisy, real-world, non-glacier environment in which segmentation is particularly difficult. Unlike glacier datasets, which have fairly consistent target classes (clean ice, debris-covered ice, and water), CDW-Seg has 10 dynamically different classes that frequently overlap, obscure, and share textures. This forces models to learn precise border identification, multi-class discrimination, and generalization in diverse backgrounds—all of which are immediately applicable to glacier front segmentation, where debris and shadows generate comparable uncertainties.

Table 4.1: Summary of Datasets Used in This Study

No.	Name	Modalities / Data Type	Coverage	Time Span	Key findings
1	Calving Front Definition (CaFFe)	SAR imagery, binary masks, zone masks, bounding boxes	7 glaciers (Antarctica, Greenland, Alaska)	1995–2020	681 SAR tiles; calving front delineation benchmark
2	NIRD Multi-modal	SAR (co- and cross-pol), optical, DEM, In-SAR	ALP, HMA, LL, NZ, SA, SC (global)	1988–2020	1,027 tiles; multimodal fusion for generalizable glacier mapping
3	HKH Glacier Mapping	Landsat 7 bands and NDVI/NDSI/NDWI, SRTM topo, spectral indices	Hindu Kush Himalaya (8 countries)	2002–2008	14,190 ML-ready patches; clean vs debris-covered glaciers
4	Cityscapes	RGB high-res urban imagery	50 cities (Germany + neighbors)	2020	5,000 fine + 20,000 coarse images; segmentation backbone benchmark
5	CDW-Seg	RGB waste site imagery, polygon masks (10 classes)	Construction sites	2025	5,413 annotated objects; cross-domain segmentation generalization

1. **Regional Glacier Diversity (HKH Glacier Mapping):** The HKH dataset and NIRD dataset provides extensive regional coverage across eight countries within the Hindu Kush Himalaya, allowing us to assess Glacier-Seg’s robustness against terrain variability, debris-covered surfaces, and mixed climatic conditions. Its inclusion validates the model’s performance in complex glaciated landscapes, where local spectral and topographical differences challenge most segmentation frameworks.
2. **Urban Structural Baseline (Cityscapes):** The Cityscapes dataset serves as a high-quality benchmark for model calibration and baseline comparison. Evaluating Glacier-Seg on this dataset enables us to test the backbone’s general feature extraction capacity, ensuring that architectural adaptations such as Involution and Mamba mixers preserve structural fidelity even in high-resolution urban imagery. This also strengthens confidence in the model’s

applicability to non-cryospheric segmentation problems.

3. **Cross-Domain Generalisation (CDW-Seg):** By testing on CDW-Seg, we demonstrate that Glacier-Seg generalises similarly as Segformer to non-glacial yet visually complex segmentation tasks. The dataset’s heterogeneous textures and cluttered compositions offer a realistic test for model resilience, highlighting its capability in high-stakes industrial and environmental applications.
4. **Benchmarking Against Strong Baselines:** Both Cityscapes and CDW-Seg provide established baselines using Vision Transformers and lightweight adapters. Achieving comparable or superior performance underscores the strength of combining *Involution* (local adaptivity) and *Mamba* (long-range mixing) in producing compact yet high-performing segmentation models.
5. **Multi-Class and Boundary Sensitivity:** Unlike binary glacier datasets, CDW-Seg and Cityscapes assess multi-class segmentation performance. This allows Glacier-Seg to be evaluated for scalability from binary glacier front delineation to complex multi-class problems. Additionally, since CDW objects often overlap within scenes—analogueous to debris covering ice surfaces—these datasets serve as natural stress-tests for boundary-aware loss functions and uncertainty calibration techniques.

4.1.2 Data Preprocessing

To increase the diversity and robustness of the training data, we applied a multi-stage data preprocessing and augmentation pipeline on the raw SAR images, glacier zone masks, and bounding box annotations. This pipeline was designed to simulate real-world imaging variations, address dataset imbalance, and improve generalisation of the segmentation models.


Each image–mask–bounding box triplet was processed through geometric and radiometric augmentation stages, and stored in separate output folders for SAR images, zone masks, and bounding boxes.

Geometric Augmentation

To introduce spatial diversity, we randomly applied one of the following transformations to each image-mask pair, ensuring corresponding changes were reflected in bounding box coordinates:

- **Rotation:** Random rotation between -30° and $+30^\circ$, using bilinear interpolation for images and nearest-neighbor for masks. Bounding boxes were re-derived from the transformed mask.
- **Flipping:** Random horizontal, vertical, or combined flipping. Bounding box coordinates were adjusted accordingly.
- **Shifting:** Random pixel-level translation in the range ± 40 pixels along both x and y axes. The bounding box was shifted by the same offset.


- Shearing: Horizontal shear with a random factor in the range $[-0.2, 0.2]$, followed by recalculation of the bounding box from the transformed mask.



images/AUGMENTDATASET1.png

Figure 4.8: Geometric augmentations performed on the dataset 1.

Not all augmentations were applied to every image; instead, a single transformation was chosen randomly from the above list and applied per patch. All resulting bounding boxes were clipped to remain within image boundaries. The augmented data were saved alongside the originals to increase training diversity.



images/augmentdataset2.png

Figure 4.9: Geometric augmentations performed on the dataset 2.

Radiometric Augmentation

SAR images are susceptible to signal-level changes that are physically realistic. These augmentations are applied only to the SAR image, leaving the zone mask and bounding box unchanged.

Speckle Noise: Speckle noise is a granular, salt-and-pepper-like interference pattern inherent in Synthetic Aperture Radar (SAR) imagery due to the coherent nature of radar wave interactions. Unlike optical sensors, SAR systems emit microwave pulses and record backscattered signals, making speckle an unavoidable byproduct of the imaging process [48].

Approximately 60% of SAR images were randomly selected to undergo speckle augmentation. The noise was modeled using a multiplicative Gamma distribution, simulating variations in signal intensity as follows:

$$I' = I \cdot 10^{\Delta_{dB}/10} \quad (4.1)$$

where I is the original pixel intensity, and Δ_{dB} is sampled from a zero-mean Gaussian distribution representing noise in decibels.

Radiometric Gain Shift: Radiometric gain shifts refer to unnatural intensity changes caused by fluctuations in the radar system’s amplification gain during acquisition. Unlike speckle noise, which is inherent to SAR physics, gain shifts are sensor-induced and can distort backscatter values, affecting the reliability of downstream analysis [49].

To simulate such variations, SAR images were scaled by a random gain factor G drawn from a Gamma distribution:

$$I' = I \cdot G, \quad G \sim \text{Gamma}(L, \frac{1}{L}) \quad (4.2)$$

where L is the number of looks, controlling the variance of the distribution. This simulates different levels of sensor gain across the dataset.

Photometric Augmentation

Photometric augmentation aims at changing the image properties such as brightness, contrast, and illumination, without altering the geometry or structure of objects. The purpose is to test different lighting conditions, camera exposure, and sensor behaviours, thereby enhancing the robustness of the model under real-world conditions where visual properties may vary significantly compared to the training data.

This augmentation simulates varying exposure levels by adjusting pixel intensities. Brightness is modified by adding an offset, while contrast is altered through a scaling factor. This ensures that the model does not overfit to fixed illumination levels and generalizes better to diverse environments.

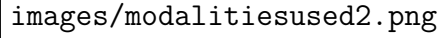
$$I' = \alpha \cdot I + \beta, \quad \alpha \sim U(\alpha_{\min}, \alpha_{\max}), \quad \beta \sim U(\beta_{\min}, \beta_{\max}) \quad (4.3)$$

Where:

- I : original pixel intensity
- I' : transformed pixel intensity
- α : contrast scaling factor (sampled from uniform range $[\alpha_{\min}, \alpha_{\max}]$)
- β : brightness offset (sampled from $[\beta_{\min}, \beta_{\max}]$)

Missing Modalities

As many tiles from the second dataset [12] had instances where modalities such as `cross_pol_sar`, `in_sar` and `co_pol_sar` of a SAR image was missing for a tile, only DEM and optical images were combined to use as the tile.



images/modalitiesused2.png

Figure 4.10: Rendered images of each channel from the common modalities.

For the CaFFe dataset, the original samples consisted of SAR images (in `.png` format) from Sentinel-1 acquisitions, binary glacier zone masks (`zones.png`), and annually annotated bounding boxes (`.txt`) marking the calving fronts. If the bounding box file was missing for a given image, the largest foreground region in the zone mask was extracted using OpenCV's `cv2.boundingRect` and used as a proxy bounding box for downstream training and evaluation.

4.1.3 Models for Glacier Segmentation

U-Net

U-Net is a symmetrical based encoder and decoder conventional neural network (CNN) specially constructed for semantic image segmentation, focusing on medical imaging consisting of 23 layers where each blue box in the U-Net architecture corresponds to multi channel feature maps. It has the capacity to process high resolution images and generate precise segmentation maps. U-Net is able to extract and obtain complex features, while also keeping and restoring spatial information through a combination of convolution, up-convolution, pooling and skip connections. Its architecture consists of an encoder-decoder framework with skip connections that help retain spatial information. The encoder extracts generic features using convolutional layers and downsampling, while the decoder reconstructs the segmented image through downsampling and it is further reduced through bottleneck processing. The contracting path incrementally downsamples via 3×3 convolutions + Rectified Linear Unit (ReLU) + 2×2 max-pooling with stride 2 for downsampling, doubling feature channels each layer [50]. Here, precise localization is additionally rendered by expanding path upsamples that concatenate mirrored encoder features via skip connections that preserves the spatial detailing that has been lost during pooling. With extensive domain usage in biomedical image segmentation, remote sensing, it is also effective for environmental applications like glacier segmentation to automatically delineate glacier boundaries from remote sensing imagery. The key equation governing U-Net is the cross-entropy loss function or dice loss used for pixel-wise classification. In glacier segmentation, U-Net helps identify ice formations and boundaries by learning spatial features effectively.

images/architecture/Smp_uet-1.png

Figure 4.11: Architecture for Unet Segmentation

Key mathematical equations for U-Net in Glacier Segmentation:

1. Convolutional Operation

$$X_{i,j}^{(l)} = \sum_{m=-k}^k \sum_{n=-k}^k W_{m,n}^{(l)} \cdot X_{(i+m),(j+n)}^{(l-1)} + b^{(l)} \quad (4.4)$$

where:

- $X_{i,j}^{(l)}$: Output pixel at position (i, j) in layer l .
- $X^{(l-1)}$: Input feature map from previous layer.
- $W_{m,n}^{(l)}$: Kernel weight at position (m, n) in layer l .
- $b^{(l)}$: Bias term.
- k : Kernel radius, e.g., $k = 1$ for a 3×3 kernel.

Relevance: Convolution extracts local features such as crevasses, snow-ice textures, and debris-covered regions, which are critical for glacier boundary detection.

2. Skip Connections

$$S^{(l)} = \text{Concat}(U^{(l)}, A_{\text{encoder}}^{(l)}) \quad (4.5)$$

where:

- $S^{(l)}$: Combined features after skip connection.
- $U^{(l)}$: Decoder output at level l .
- $A_{\text{encoder}}^{(l)}$: Encoder feature at the same resolution.

Relevance: Skip connections recover fine spatial details lost in downsampling, which is essential for delineating precise glacier fronts and narrow ice boundaries.

3. Dice Loss

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2 \sum_i p_i y_i + \epsilon}{\sum_i p_i + \sum_i y_i + \epsilon} \quad (4.6)$$

where:

- p_i : Predicted probability (soft score) at pixel i .
- y_i : Ground truth label for pixel i .
- ϵ : Small constant for numerical stability.

Relevance: Dice loss is robust for imbalanced data, ensuring glacier pixels (often fewer than background) are segmented accurately, especially in debris-covered or shadowed regions.

ResNet-50

ResNet-50 is a powerful deep convolutional neural network (CNN) known for its residual learning framework, which mitigates the vanishing gradient problem allowing deeper networks to be trained effectively. A 50-layer deep CNN with residual bottleneck blocks and containing 1 MaxPool and 1 Average Pool Layer. The 50 layers include convolutional layers, batch normalization, and skip connections (skipping layers). The layers in ResNet-50 consist of convolutional layers used for learning various patterns and edges with input images and added to the identity block to make the input size equal to the output, blocks that are composed of normalization and softmax activation functions extracting high level features and full connected layers to make the final predictions [51]. It also consists of a Bottleneck Residual Block where layers are divided as ReLU, 1×1 , 3×3 and 1×1 bottleneck convolutional layers and finally to skip connections. The use of residual blocks has been shown to improve performance on image classification tasks compared to traditional CNNs. The ResNet-50 takes in an input image that is fed to the network and convolutional layers process the image, extracting numerous features like edges and textures and shapes. The extracted features are passed through residual blocks using skip connections to maintain information flow and prevents vanishing gradient problems. The classification of features is then done by the fully connected layers, interpreting the predictions about the image. ResNet-50 is useful in training large number layers allowing complex models and improving performance on image classification. The useful domains of ResNet-50 also lie in glacier segmentation to extract rich hierarchical features and effectively handle deep image contexts. Each ResNet-50 block introduces projection shortcuts in the architecture.

images/architecture/resnet50.png

Figure 4.12: Architecture for Resnet-50 Segmentation

1. Residual Block Equation

$$y = x + F(x) \quad (4.7)$$

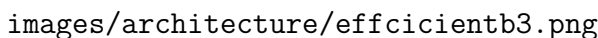
where:

- x : Input feature map.
- $F(x)$: Learned residual mapping through convolutional layers.
- y : Output feature map after adding the shortcut connection.

Relevance: The residual block enables very deep networks to be trained by mitigating vanishing gradients. For glacier segmentation, this allows ResNet-50 to extract rich hierarchical features, capturing both fine edges of glacier fronts and large-scale ice-debris patterns without loss of information flow.

EfficientNet

EfficientNet provides a scalable computational neural network architecture that requires less computational resources and focuses on the dimensions depth, width, and resolution using a compound scaling method. It is widely used in image classification and segmentation. The architecture of EfficientNet consists of stem, body and head. It uses Swish activation in stem and kernel size of 3×3 and convolutional with 32 filters, finally fully connected layer with softmax activation function for classification. In order to optimize the usage of resources, EfficientNet introduces inverted blocks with a lightweight 3×3 depth wise bottleneck convolution layer which is then followed by a two point wise 1×1 expansion convolutional layer. The key equation governing EfficientNet is the compound scaling formula. A certain fixed set of scaling coefficients is used to scale each dimension surpassing SOTA accuracy of CNN [52]. Squeeze-and-excitation blocks are employed as an attention mechanism for improved representation of features. EfficientNet progressively increases the model depth, width and resolution based on the compound scaling coefficient for effective scaling ensuring seamless computational constraints by preserving performance. EfficientNet has several variants with B0 as baseline and efficacy of this mode involves outstanding efficiency and outperforms other models in terms of accuracy with fewer parameters. For compound scaling methods, the scaling factor is used as the width of the network. Width = β^ϕ , depth = α^ϕ , and resolution = γ^ϕ , where β , α , and γ are the constants determined via grid search. This model is lightweight and provides high accuracy for satellite imagery useful for glacier monitoring.



The image shows the architecture for EfficientNet Segmentation. It is a diagram illustrating the model's structure, which typically includes a stem, a body of inverted residual blocks, and a head for segmentation. The diagram is located at the file path: images/architecture/efficientb3.png.

Figure 4.13: Architecture for EfficientNet Segmentation

VGG-19

The Visual Geometry Group (VGG) model known as VGG-19 is one of the deep convolutional neural networks (CNN) consisting of 19 layers, that includes 16 convolutional weight layers with 3 fully connected layers with simplifying network structure having 144M parameters. It consistently uses a small uniform 3×3 kernel filter to acquire accurate details with stride and padding of 1 capturing spatial resolutions along with filters in each layer [53]. Throughout the evolution, VGG models have activation functions as (ReLU) Rectified Linear Unit to introduce non linearity that is applied in each layer, learning complex models and increasing deep networks for more complex features to ensure smooth performance in semantic segmentation tasks including glacier segmentation. Just like Unet, maxpooling layers here consist of 2×2 layers with stride 2 for the reduction of spatial dimensions. While the fully connected layers combine the features for image classification, max pooling enables preserving important features of the classification. Lastly, the softmax layer outputs the class probabilities. With a wide range of use in robust feature extraction and image classification, VGG-19 is useful for segmentation tasks when integrated into models like U-Net, requiring more computational force with significant memory. The key equation governing VGG-19 is the activation function: $[y = \max(0, x)]$ which introduces non-linearity. In glacier segmentation, VGG-19 helps in identifying subtle variations in ice formations and distinguishing between different glacier regions, transfer learning and feature extraction.

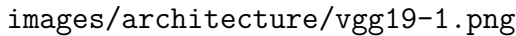

The image shows a rectangular box containing the text 'images/architecture/vgg19-1.png'. This is likely a placeholder for a diagram illustrating the VGG-19 architecture, which typically shows a sequence of convolutional and pooling layers followed by fully connected layers and a softmax output.

Figure 4.14: Architecture for VGG-19 Segmentation

DeepLabV3

DeepLabV3 is a state-of-the-art deep learning semantic segmentation model that uses (dilated convolutions) Atrous Spatial Pyramid Pooling (ASPP) to capture multi-scale features for useful information and adjust receptive fields without down-sampling. Multi scaling involves the analysis of the glacier images at different scales where deeplabV3 allows it without increasing complexity of the model. It employs deeper dilated convolutions to expand the receptive field without increasing the number of learnable parameters [54]. Obtaining high level information at finer resolutions with large parameters helps to attain more complex features by increasing kernel size with increasing dilation rate to better preserve the spatial resolution, leading to accuracy of segmentation results. Similar to Unet, it has an encoder and decoder section that is a modified version of ResNet architecture used for extracting better features and decoder to unsample using bilinear interpolation and combining

the featured maps. DeepLabV3 is widely used in satellite image analysis and environmental monitoring, making it ideal for glacier segmentation. DeepLabV3 helps in glacier segmentation by discovering glacial ice boundaries and glacier formation. DeepLabV3 is used for satellite and glacier segmentation due to its accuracy in boundary delineation.



images/architecture/Deeplabv3-1.png

Figure 4.15: Architecture for DeeplabV3 Segmentation

SegFormer

SegFormer consists of two main modules such as hierarchical transformer-based (also known as Mix Transformer) MiT semantic segmentation model that combines encoder feature extraction with a lightweight All MLP decoder [31] with self attention mechanism for generating high multi level features and resolutions. This novel hierarchy structured transformer shows multiscale outputs with parameters ranging from 50-64 M, it is flexible for UAV and satellite scale variance. Unlike other traditional convolutional neural networks, SegFormer uses self-attention mechanisms to capture global dependencies aggregating information from different layers through decoder for rendering powerful representations. SegFormer avoids complex decoders and does not need any positional encoding, thus avoids the interpolation of positional codes that degrades the performance while testing [31]. Here, an overlapped patch merging process is used in traditional ViT to obtain feature maps and efficient self attention (ESA) mechanism is the main computation bottleneck of the encoders[31]. Training this model is comparatively easy that includes steps such as loading the data with dataloader and initializing hyperparameters with optimizer and writing training loop and has a larger effective receptive field compared to Deeplabv3+, avoiding heavy context modules like ASPP[31] . It is widely used in remote sensing making it suitable for glacier segmentation. SegFormer helps in glacier segmentation by upgrading long-range dependencies and improving boundary detection. SegFormer provides more flexible and generalized segmentation in different resolutions and its agnostic design is well suited to glaciology.

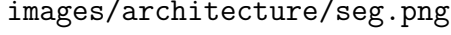
The diagram area is mostly blank, with the text 'images/architecture/seg.png' located in the lower-left corner. This likely indicates the location of the original figure or that the figure content is missing from this page.

Figure 4.16: Architecture for Segformer Segmentation

Self-Attention Mechanism in SegFormer:

The scaled dot-product self-attention used in SegFormer is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^\top}{\sqrt{d}} \right) V \quad (4.8)$$

where:

- $Q \in R^{N \times d}$: Query matrix representing the current patch's feature.
- $K \in R^{N \times d}$: Key matrix encoding contextual relationships.
- $V \in R^{N \times d}$: Value matrix containing the actual features to be aggregated.
- d : Dimensionality of the feature vectors.
- N : Number of image patches (tokens).
- QK^\top : Dot product capturing similarity between patches.
- \sqrt{d} : Scaling factor to stabilize gradients.
- softmax : Normalization function to compute attention weights.

The self-attention mechanism in SegFormer enables the model to capture long-range dependencies between glacier regions, which is crucial for segmenting complex terrain such as calving fronts and debris-covered ice. Unlike CNNs that rely on local

receptive fields, attention allows SegFormer to integrate global context across satellite tiles, improving segmentation accuracy in heterogeneous glacier environments. This is especially beneficial when glacier boundaries are obscured by shadows, debris, or cloud cover—conditions common in remote sensing imagery.

VisionMamba

VisionMamba (ViM) is a hierarchical state-space model (SSM) based vision backbone, which can be viewed as a state-space version of classical Vision Transformers (ViTs) [55]. In contrast to the Mix Transformer encoder of SegFormer, ViM employs the Mamba state-space mixer, which is linear in sequence length and therefore more efficient for high-resolution images with overlapping patches.

In the tiny configuration, ViM uses a patch embedding of 16×16 with a stride of 8, resulting in overlapping tokens that encode finer spatial details. For an input of size 224×224 , this yields a total of 729 tokens, each embedded into a 192-dimensional space. The sequence is processed through 24 stacked Mamba blocks and stabilized by fused residual connections and RMSNorm.

Unlike classical ViTs, ViM is not based on quadratic attention, making it computationally efficient while still modeling long-range dependencies. The architecture includes a [CLS] token for robust global representation learning, with absolute positional embeddings to preserve spatial order.

Rotary embeddings (RoPE) are disabled in this configuration for simplicity.

After serial mixing, the features are normalised, pooled (mean), and passed through a linear classification head.

This design makes ViM a lightweight yet expressive backbone suitable for applications such as remote sensing and glacier segmentation, where overlapping tokens and long-range modeling contribute to improved boundary delineation and spatial consistency.

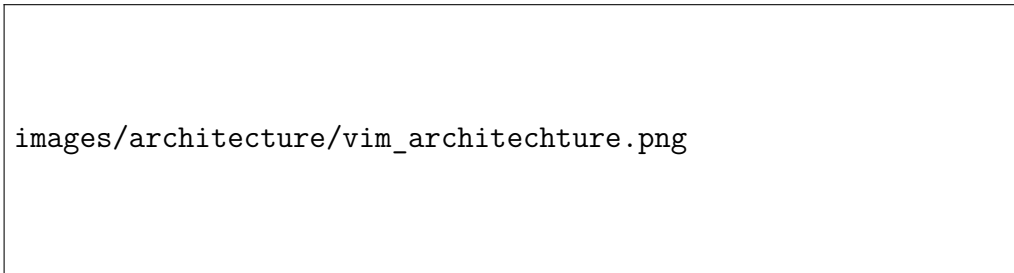


Figure 4.17: Overall architecture of Vision Mamba (ViM), showing patch embedding, stacked Mamba blocks, [CLS] token with positional embeddings, and classification head. Adapted from [55].

1. Recurrent state-space update:

$$h_t = \sigma(W_x x_t + W_h h_{t-1}), \quad y_t = W_o h_t \quad (4.9)$$

where:

- $x_t \in R^d$: Input token embedding at time step t .
- $h_t \in R^{d_h}$: Hidden state updated through recurrence.
- $W_x \in R^{d_h \times d}$: Input projection matrix.
- $W_h \in R^{d_h \times d_h}$: State transition matrix.
- $W_o \in R^{d \times d_h}$: Output projection matrix.
- σ : Non-linear gating function controlling information flow.
- y_t : Output representation at time step t .

This recurrence formulation is central to Vision Mamba’s state-space modelling, enabling efficient long-sequence processing with linear complexity. Unlike Transformers that rely on quadratic attention, Mamba updates hidden states sequentially, allowing it to scale to high-resolution glacier imagery. This is particularly useful for modelling spatial continuity in glacier fronts and capturing temporal dynamics in multimodal satellite sequences.

Model	Architecture Type	Layers / Parameters	Key Features / Modules	Special Techniques	Computational Efficiency
U-Net	CNN (Encoder-Decoder)	23 layers	Convolution, up-convolution, pooling, skip connections	Skip connections for precise spatial info	Medium
ResNet-50	CNN (Residual Network)	50 layers	Residual bottleneck blocks, MaxPool & AvgPool	Residual connections to prevent vanishing gradient	High
EfficientNet	CNN	Varies (B0-Bx)	Stem, inverted bottleneck blocks, squeeze-and-excitation	Compound scaling (depth, width, resolution)	Very High
VGG-19	CNN	19 layers, 144M params	Uniform 3×3 convolutions, MaxPool	Deep network for robust feature extraction	High
DeepLabV3	CNN (Dilated)	Varies	Atrous Spatial Pyramid Pooling (ASPP), encoder-decoder	Dilated convolutions for multi-scale feature capture	Medium-High
SegFormer	Transformer-based	50-64M params	Mix Transformer encoder, lightweight MLP decoder	Self-attention, hierarchical multiscale outputs	Medium
Vision-Mamba (ViM)	State-Space Transformer	24 Mamba blocks	Patch embedding, [CLS] token, fused residual connections	State-space recurrence for long-range dependencies	High

Table 4.2: Comparison of Deep Learning Models for Glacier Segmentation

4.1.4 Manual Fine-tuning

Since this study evaluates multiple state-of-the-art (SOTA) semantic segmentation architectures, fine-tuning was applied to each model to optimise their performance on glacier-specific data. The models selected for this study include U-Net, DeepLabV3+, ResNet50, VGG19, EfficientNet, and SegFormer. These models were initially pre-trained on large-scale datasets, such as ImageNet or ADE20K by their innovators, and were then fine-tuned on glacier-specific datasets to adapt them to the domain-specific characteristics of glacier boundaries by us.

The fine-tuning process involved adjusting several hyperparameters to enhance model

performance, including learning rate, batch size, input crop size, and weight decay. The goal of fine-tuning is to improve model accuracy by adapting pre-trained weights to the glacier data while maintaining the advantage of transfer learning. Below, we describe how fine-tuning was applied to each model.

We focus on tuning the following hyperparameters:

- **Learning Rate:** The learning rate controls the size of the steps the optimizer takes during training to minimize the loss function. A learning rate that is too high can cause the model to overshoot the optimal solution, resulting in unstable training or divergence. Conversely, a learning rate that is too low slows down training and may get stuck in suboptimal minima.
- **Batch Size:** Batch size refers to the number of training samples processed before the model updates its parameters. Larger batch sizes provide a more stable gradient estimate but require more GPU memory. Smaller batch sizes introduce more noise in gradient estimation, which sometimes helps escape local minima but may also cause instability.
- **Input Crop Size:** This is the spatial size (height and width) of the image patches fed into the model during training. Larger crop sizes provide the model with more context about glacier features, such as complex boundary shapes and surrounding terrains, but require more computation and memory. Smaller crops speed up training but may cause loss of important spatial relationships.
- **Weight Decay:** Weight decay is a regularization technique that penalizes large weights during training, helping to prevent overfitting, especially when training on limited data. It effectively encourages simpler models that generalize better to unseen glacier regions.

Manual tuning was chosen as our strategy due to its computational efficiency and its ability to provide direct feedback on how specific changes impact the model’s performance. By evaluating the effects of each change in a controlled trial-and-error fashion, we were able to incrementally improve the segmentation accuracy in the validation data.

This manual approach not only reduced the computational overhead compared to automated methods like grid or Bayesian search but also enabled an intuitive understanding of the relationship between hyperparameters and performance. For example, increasing the crop size to 256×256 allowed the model to better capture glacier boundary shapes, while adjusting the batch size helped mitigate limitations surrounding training duration.

We applied this iterative fine-tuning process to each model independently. For instance, in the case of U-Net with an EfficientNet-B3 encoder, we began with a batch size of 8, a learning rate of $1e-3$, and a crop size of 512×512 . By the final configuration - after reducing the learning rate to $1e-4$ and adjusting the crop size to 256×256 - the validation Dice coefficient improved by approximately 7.3%, indicating a significant performance gain from fine-tuning.

Fine-tuning proceeded end-to-end so that the pre-trained encoder could adapt to the distinctive texture and backscatter of SAR imagery. Optimisation used AdamW with a binary cross-entropy loss (`BCEWithLogits`), which is well suited to this two-class segmentation task.

array

Table 4.3: Hyperparameter Modifications for Dataset 1

Parameter	Initial	Modified	Reason
Learning Rate	1×10^{-3}	1×10^{-4}	Smoother convergence and reduced overshooting in optimization.
Batch Size	16	8	Reduced to avoid out-of-memory errors on GPU.
Crop Size	512^2	256^2	Larger context improves glacier boundary detection.
Weight Decay	None	0.01	Added to improve regularization and prevent overfitting.
Optimizer	Adam	AdamW / Adam	AdamW used for transformers (e.g., SegFormer); Adam for CNN-based baselines.
Loss Function	Categorical CE	Binary CE	Adapted to binary segmentation target.

Table 4.4: Hyperparameter Modifications for Dataset 2

Parameter	Initial	Modified	Reason
Learning Rate	1×10^{-3}	1×10^{-6}	Smoother convergence and reduced overshooting in optimization.
Batch Size	16	10	Reduced to avoid out-of-memory errors on GPU.
Crop Size	256^2	256^2	Larger spatial context improves glacier boundary detection.
Weight Decay	None	0.01	Added to improve regularization and prevent overfitting.
Optimizer	Adam	AdamW / Adam	AdamW used for transformers (e.g., SegFormer); Adam for CNN-based baselines.
Loss Function	Categorical CE	Weighted CrossEntropy	Still using CE with class weights (instead of Binary CE), since code applies <code>nn.CrossEntropyLoss</code> .

Table 4.5: Hyperparameter Modifications for Dataset 3

Parameter	Initial	Modified	Reason
Learning Rate	1×10^{-3}	1×10^{-6}	Smoother convergence and reduced overshooting in optimization.
Batch Size	16	10	Reduced to avoid out-of-memory errors on GPU.
Crop Size	256^2	256^2	No modification required; existing crop size sufficient.
Weight Decay	None	0.01	Added to improve regularization and prevent overfitting.
Optimizer	Adam	AdamW / Adam	AdamW used for transformers (e.g., SegFormer); Adam for CNN-based baselines.
Loss Function	Categorical CE	Weighted CrossEntropy	Still using CE with class weights (instead of Binary CE), since code applies <code>nn.CrossEntropyLoss</code> .

Table 4.6: Hyperparameter Modifications for Dataset 4

Parameter	Initial	Modified	Reason
Learning Rate	1×10^{-3}	1×10^{-4}	Smoother convergence and reduced overshooting in optimization.
Batch Size	16	20	Increased to utilize GPU more efficiently.
Crop Size	256^2	256^2	No modification required; existing crop size sufficient.
Weight Decay	None	0.01	Added to improve regularization and prevent overfitting.
Optimizer	Adam	AdamW / Adam	AdamW used for transformers (e.g., SegFormer); Adam for CNN-based baselines.
Loss Function	Categorical CE	Weighted CrossEntropy	Still using CE with class weights (instead of Binary CE), since code applies <code>nn.CrossEntropyLoss</code> .

Table 4.7: Hyperparameter Modifications for Dataset 5

Parameter	Initial	Modified	Reason
Learning Rate	1×10^{-3}	1×10^{-4}	Smoother convergence and reduced overshooting in optimization.
Batch Size	16	20	Increased to utilize GPU more efficiently.
Crop Size	256^2	256^2	No modification required; existing crop size sufficient.
Weight Decay	None	0.01	Added to improve regularization and prevent overfitting.
Optimizer	Adam	AdamW / Adam	AdamW used for transformers (e.g., SegFormer); Adam for CNN-based baselines.
Loss Function	Categorical CE	Weighted CrossEntropy	Still using CE with class weights (instead of Binary CE), since code applies <code>nn.CrossEntropyLoss</code> .

4.1.5 Evaluation

Experimental Setup

We used several key libraries for model development and data processing, including NumPy, Keras, Torch, OpenCV, TensorFlow, and Matplotlib. These libraries facilitated the efficient processing of the dataset and the training of deep learning models for glacier segmentation. Model training and dataset preparation were carried out on a system equipped with a RYZEN 9 5950X CPU, an NVIDIA GeForce RTX 3080 Ti GPU with 12 GB of memory, and 64 GB of DDR4 RAM. For our implementation, we used Python 3.10.8 within Visual Studio Code 1.100 as the interactive development environment.

Evaluation Metrics

After training and fine-tuning our segmentation models on glacier datasets, we assess their performance using two primary evaluation metrics widely used in semantic segmentation tasks: Mean Intersection over Union (Mean IoU) and Dice Coefficient (also known as F1-Score or Sørensen–Dice index). These metrics provide a detailed understanding of how well the models delineate glacier boundaries and segment different classes, which is crucial for accurate glacier mapping in satellite imagery. The combination of these evaluation metrics provides a holistic understanding of model performance for glacier segmentation. While *IoU* and *Dice coefficient* emphasise spatial overlap and boundary accuracy, *Precision* and *Recall* quantify the reliability of glacier detection and the model’s sensitivity to missed regions. The *Pixel Accuracy* offers a general measure of overall correctness but is complemented by the *AUC-ROC* to assess discriminative capability under class imbalance. Finally, the *confusion matrix* provides a granular visualisation of error types, distinguishing between over-segmentation and omission errors. Together, these metrics enable a comprehensive, multi-perspective evaluation of segmentation quality, ensuring the robustness and generalisability of the proposed models across diverse glacier terrains and imaging conditions.

Model Explainability

Interpretation of Grad-CAM: To better understand the decision-making process of our segmentation models, we used Grad-CAM (Gradient-weighted Class Activation Mapping), a technique that generates heatmaps to visualize which areas of the input image are most influential in the model’s prediction [56]. By applying Grad-CAM to the models, we were able to identify the regions in the input glacier images (e.g., the glacier front or ice-covered regions) that the models focus on during segmentation.

When we observe Grad-CAM heatmaps with highlights centered on the glacier or specific ice regions, it indicates that the model is learning to focus on important glacier features such as the glacier front or ice surface. For correctly segmented images, the heatmaps usually align well with the ground truth glacier boundaries. If the model overfits to irrelevant regions (e.g., background or non-glacier areas), the heatmaps will show attention in those regions. This is especially valuable for

identifying whether the model is focusing on the correct areas, such as ice versus background.

For instance, in cases where Grad-CAM highlights the center of the glacier (the main mass of the glacier or ice front), this can suggest that the model is correctly identifying the relevant areas. However, if the model focuses too much on certain parts of the glacier (e.g., the central region), this might indicate an issue with the generalization to other parts of the glacier or regions where the glacier is not centered. This kind of insight is important for further tuning the model and improving its ability to generalize across various glacier types and scenarios. Grad-CAM thus provides a means to inspect how well the model is focusing on regions of interest that are consistent with the ground truth.

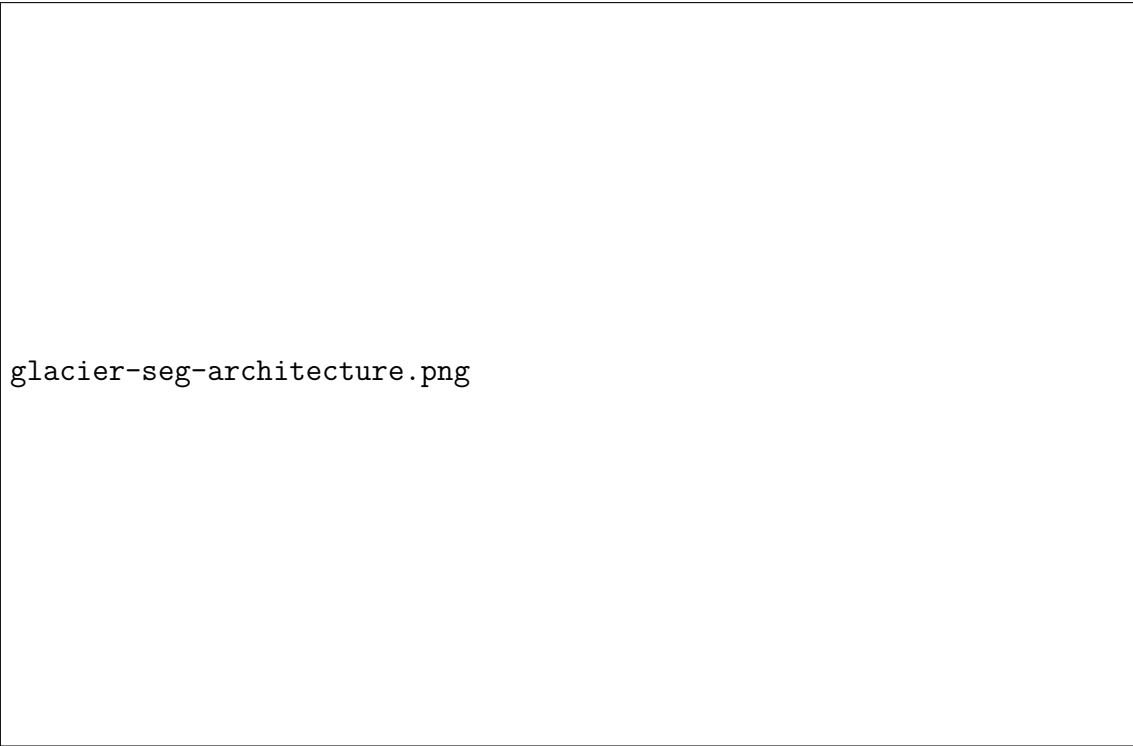
Interpretation of Saliency Maps: Saliency maps are another valuable technique for explaining the model’s predictions. Saliency maps provide insights into which pixels in the input image have the most influence on the model’s decision [57]. By computing the gradient of the model’s output with respect to the input image, we can identify the areas of the image that are most important for the model’s prediction. When applying saliency maps to our models, we observed that important regions, such as the glacier boundary and transitions between ice and background, are highlighted. For a well-performing model, the saliency map should emphasize the pixels that contribute to identifying the glacier’s boundary. In some cases, if the saliency map highlights irrelevant regions, this could indicate that the model is not focusing on the correct features of the glacier, such as misinterpreting the background as part of the glacier.

Saliency maps help us to evaluate whether the model is overly sensitive to certain image artifacts or background features. For example, if the model is focusing on regions outside the glacier or areas with similar intensity (e.g., water bodies), this could indicate an issue with training or data bias. Additionally, saliency maps are useful for detecting if the model is using spurious correlations (e.g., detecting a glacier’s color but ignoring texture or shape). In contrast, a properly trained model should focus on important texture and shape features that define the glacier boundaries. Saliency maps are particularly useful for diagnosing issues with model focus. If the highlighted areas do not align with the glacier’s actual boundaries or key features, it suggests that the model may need further refinement.

By combining Grad-CAM and Saliency Maps, we gain a comprehensive view of how the model is making its segmentation decisions whenever deemed necessary. Grad-CAM provides high-level, class-specific attention maps, while Saliency Maps highlight the individual pixel-level importance. Together, these methods offer both global and local insights into the model’s decision-making process. For example, in the context of glacier segmentation, Grad-CAM can indicate the regions of the image the model is focusing on (such as the glacier front), while Saliency Maps can pinpoint specific pixels within that region that were most influential in the model’s decision. This combination can be particularly useful when assessing the model’s performance in complex or ambiguous cases, such as when glaciers are covered by debris or surrounded by challenging terrain.

4.2 Design (Model) Specification

This section details the architecture of our custom hybrid model for glacier segmentation in remote sensing images. The model integrates Convolutional Neural Networks (CNNs) for local feature extraction, Vision Transformers (ViTs) for global context, and Mamba State Space Models (SSMs) for efficient sequence modelling. This design addresses the computational constraints and data challenges outlined in previous sections, achieving a balance between accuracy and efficiency suitable for resource-limited research environments.



glacier-seg-architecture.png

Figure 4.18: Simple Architecture of Glacier-Seg

4.2.1 Overall Architecture

Semantic segmentation involves classifying every pixel in the input image. For glacier images, this could mean distinguishing glacial features based on colour, texture, and context (e.g., identifying crevasses, moraines, or ice flow). The model achieves this through a hierarchical encoder-decoder setup, processing the image in stages to capture both local details (e.g., ice cracks) and global context (e.g., overall glacier shape). It follows the SegFormer encoder-decoder paradigm [31], optimised for semantic segmentation. The encoder (MiT backbone) extracts multi-scale features at resolutions with strides [4, 8, 16, 32] relative to the input. The decoder fuses these features into a unified embedding and produces per-pixel class logits.

architecture.png

Figure 4.19: High-level Architecture of Glacier-Seg

4.2.2 Key Components

Stem: Overlap Patch Embedding

The stem downsamples the input while projecting to embedding dimensions. We use an Involution2D layer [58] for spatially-varying kernels, defined as:

$$(I * K)(x, y) = \sum_m \sum_n I(x - m, y - n) K(m, n) \quad (4.10)$$

where K is generated dynamically via reduction and span convolutions. This is followed by a 1x1 projection and BatchNorm. For efficiency, a convolutional stem alternative is available.

Mamba Block

Replaces attention in selected stages. It includes pre-normalisation, Mamba SSM, and an MLP:

$$h_t = Ah_{t-1} + Bx_t, \quad y_t = Ch_t \quad (4.11)$$

with RMSNorm (or LayerNorm fallback) and DropPath for regularization. The block processes flattened features (B, N, C) for sequence modeling.

MixVisionTransformer Stages

Four stages, each with patch embedding and depth-repeated Mamba/Attention blocks. Embeddings: [16, 32, 64, 128]; Depths: [1, 1, 1, 1] (configurable).

Decoder: SegFormerHead

Projects features to a common dimension (128), upsamples to highest resolution, concatenates, and applies fusion convolutions before classification.

4.2.3 Hyperparameters

Table 4.8: Hyperparameter Configuration of Glacier-Seg Model

Parameter	Value(s)	Type
Embedding Dimensions	[16, 32, 64, 128]	Architecture
Depth per Stage	[1, 1, 1, 1]	Architecture
Mamba State Dim (d_{state})	16	Architecture
MLP Ratio	4.0	Architecture
Drop Path Rate	0.2	Regularization
Involution Reduction Ratio	4	Stem
Input Channels	11	Data
Number of Classes	2	Task
Optimizer	AdamW, lr= 1×10^{-4}	Training
Loss Function	BCE + Dice	Training
Batch Size	30	Training
Epochs	50	Training

Multiple **ablation studies** were carried out at three levels: dataset variations, layerwise feature contribution, and blockwise architectural modifications. These experiments isolated the impact of each component, clarifying trade-offs between complexity, generalization, and accuracy. The insights guided the design of the **final model**, which incorporated the most effective combination of CNN encoders, transformer attention modules, and temporal sequence handling.

Chapter 5

Final Design Adjustment

Semantic segmentation models play a pivotal role in processing remote sensing imagery for environmental monitoring applications, including glacier delineation. This section focuses on the input preparation phase, a foundational step in the model’s forward pass. Drawing from the provided implementation, we elucidate the tensor structure, dimension preservation mechanisms, and implicit preprocessing assumptions.

The final segmentation network adopts a hierarchical encoder-decoder design inspired by SegFormer [31] but introduces several parameter-efficient innovations tailored for glacier calving front mapping.

At the encoder side, each stage begins with an **Involution-based overlapping patch embedding** module that performs spatially adaptive mixing while down-sampling, preserving glacier boundary detail in both synthetic aperture radar (SAR) and optical inputs. The subsequent feature mixer employs **Mamba state-space blocks** in place of transformer self-attention, enabling long-range dependency modelling with *linear* complexity and significantly reduced parameter count.

Four successive stages generate multi-scale feature maps at strides 2, 4, 8, and 16 relative to the input resolution. These feature maps are projected and fused by a lightweight **SegFormer** decoder, which aligns all features to the finest resolution and produces the final calving-front mask.

Compared to the SegFormer baseline, the proposed model achieves comparable or superior segmentation accuracy on the NIRD dataset while requiring less than half the number of parameters. This efficiency facilitates deployment in memory-constrained or real-time monitoring settings without compromising accuracy.

Table 5.1: Complexity comparison of custom hybrid models and baseline segmentation architectures (input size: 256×256 , batch size = 32).

Model	Params (M)	Size (MB)	Inference (ms)	MACs (G)	FLOPS (G)
Glacier-Seg	0.68	2.64	2.58	3.76	7.52
U-Net	31.0	~ 120	4–6	7.6	15.2
ResNet-50	25.6	~ 98	3–4	4.1	8.2
EfficientNet-B3	12.0	~ 45	2–3	0.9	1.80
VGG-19	143.0	~ 575	7–9	9.8	19.6
DeepLabV3	41.0	~ 160	9–11	8.1	16.2
Vision Mamba	7.50	28.74	10.24	0.52	1.04
SegFormer (B0)	3.73	14.28	3.52	1.74	3.48

Even though we had significant reduction in all the parameters, we could not yet lower the FLOPS count compared to the SOTA models. The bar chart visualises the computational complexity of the proposed Glacier-Seg model compared to several SOTA semantic segmentation architectures, measured in Giga Floating Point Operations (GFLOPs). GFLOPs is a standard metric for assessing model efficiency, particularly relevant for deployment in resource-constrained environments like remote sensing applications for glacier mapping. The chart effectively highlights Glacier-Seg’s position as a balanced, lightweight hybrid (Mamba–Transformer–CNN) in terms of efficiency for cryospheric monitoring.

Moreover, there is still a lot of room in order to further decrease GFLOPs. The model can be optimized by working with the involutational operations themselves (e.g. by sharing the kernel generated, or removing redundant spatial computations) and by optimising the patch embedding stage, either by tokenizing the inputs more efficiently or by working with the inputs in a linear projection space. These architectural enhancements would bring Glacier-Seg to the close to the ultra-lightweight efficiency levels without any loss of its ability to be highly segmented.

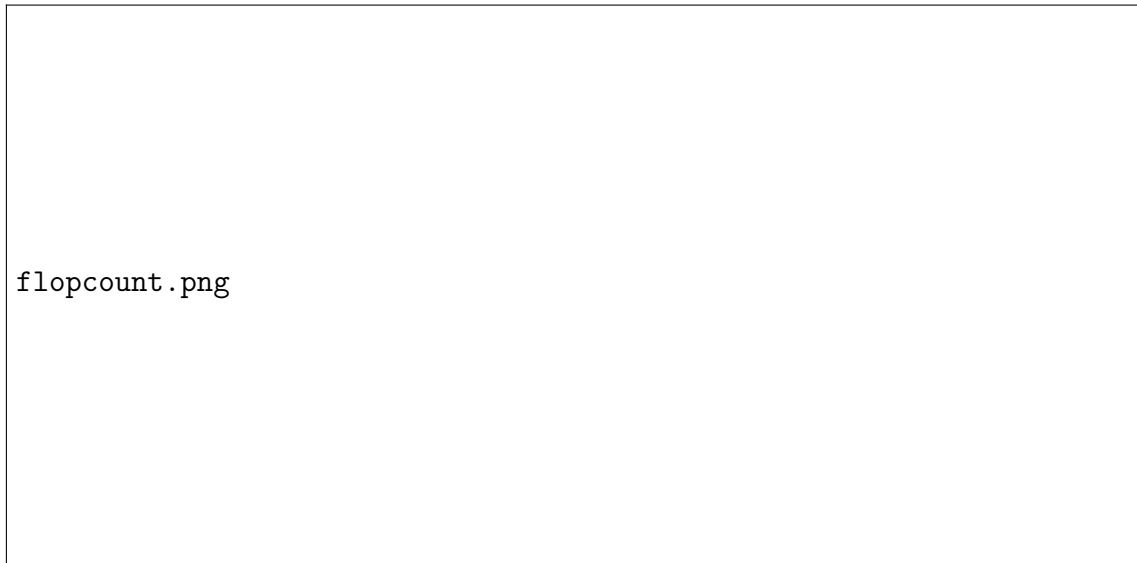


Figure 5.1: GLOPS Comparison of all models.

Glacier-Seg (7.52 GFLOPs) strikes a middle ground: it is 50% more efficient

than heavier baselines like U-Net (15.20 GFLOPs), ResNet-50 (8.20 GFLOPs), DeepLabV3 (16.20 GFLOPs), and especially VGG-19 (19.60 GFLOPs), which are computationally intensive due to deeper convolutional layers or complex decoders. Compared to ultra-lightweight models (e.g., EfficientNet-B3 at 1.04 GFLOPs, Vision Mamba at 1.80 GFLOPs, SegFormer B0 at 3.48 GFLOPs), Glacier-Seg trades some efficiency for enhanced performance (as noted in the paper’s mIoU of 0.956 and Dice of 0.977 on NIRD). This helps us render a design suitable for multi-modal inputs without excessive overhead.

5.1 Architecture

5.1.1 Input Tensor

The input to the model is formalised as a four-dimensional tensor adhering to the NCHW (Batch, Channels, Height, Width) convention prevalent in PyTorch-based implementations. Here, B denotes the batch size, facilitating parallel processing during training or inference; C_{in} represents the number of input channels, defaulting to 3 for RGB imagery; and H and W specify the spatial dimensions of each image. In the context of glacier monitoring, RGB inputs from sources such as Landsat satellites capture visual cues, including the bluish hue of compressed ice and the textured patterns of moraines. For multispectral data, C_{in} may be extended (e.g., to 4 to include near-infrared bands), enhancing discrimination between ice and vegetation. This tensor structure enables vectorized operations, such as convolutions in the subsequent patch embedding, thereby optimising performance for GPU acceleration.

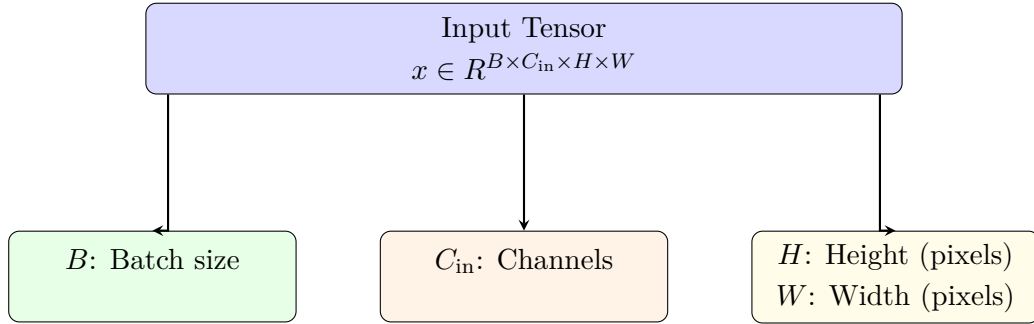


Figure 5.2: Structure of the input tensor in NCHW format used for glacier imagery segmentation.

5.1.2 Involution Patch Embedding

The first stage performs an **Involution-based overlapping patch embedding**, substituting conventional convolutional stems with dynamically generated kernels. Unlike fixed convolution filters, involution computes position-specific kernels from local context using an `AvgPool2d` neighborhood aggregator, followed by two 1×1 convolutions (2 channels reduction + 9 kernel parameters) and ReLU activation. These adaptive filters preserve edge-sensitive information across heterogeneous glacier surfaces, enabling improved texture adaptation for ice–rock transitions.

After involution, a 3×3 convolution with stride 2 and BatchNorm expands the feature map to 16 channels, producing a downsampled representation of 128×128 .

This layer contains only $24 + 27$ trainable parameters, contributing to the model’s lightweight footprint.

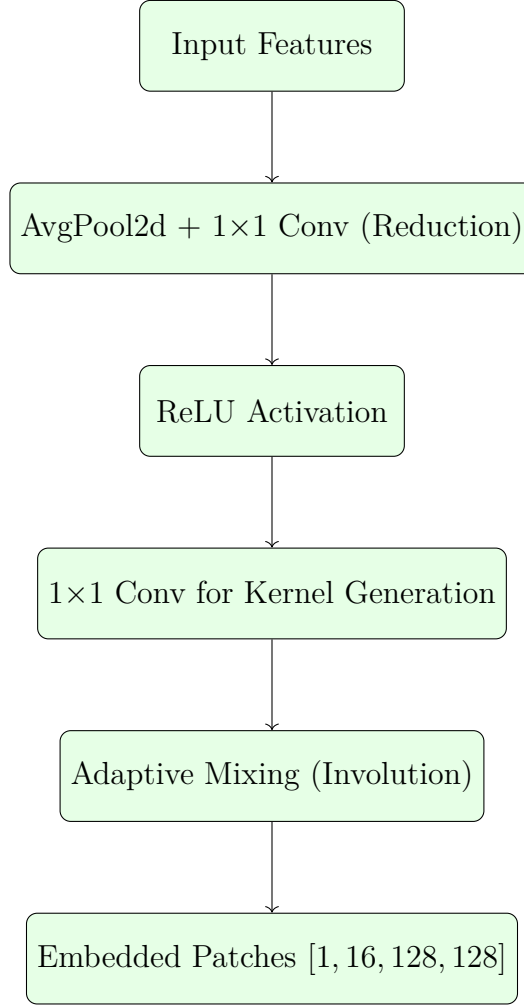


Figure 5.3: Involution-based overlapping patch embedding for adaptive feature extraction.

Involution is employed here to address the limitations of standard convolutions in handling the spatial variability inherent in glacier imagery, such as irregular debris patterns, shadowed crevasses, and variable backscatter in SAR data. By generating kernels tailored to each pixel’s local neighborhood, the mechanism enhances the model’s ability to capture fine-grained details without the computational overhead of full attention mechanisms, achieving a balance between expressiveness and efficiency. This is particularly beneficial for cryospheric applications, where data scarcity and noise demand robust yet lightweight feature extraction.

5.1.3 Mamba State-Space Mixer

Each stage employs a **Mamba state-space block** as the sequence mixer. This block applies **RMSNorm** normalization, a Mamba State-Space Model (SSM) with $(d_{\text{model}}, d_{\text{state}}) = (16\text{--}128, 16)$, and optional DropPath. The SSM updates hidden states via

$$h_t = Ah_{t-1} + Bx_t, \quad y_t = Ch_t \quad (5.1)$$

providing stable, linear-time sequence modeling. Residual connections, secondary normalization, and a feed-forward MLP with GELU activation complete the block. This configuration reproduces transformer-level expressiveness while maintaining $O(N)$ complexity—critical for high-resolution ($N > 10^4$) glacier images.

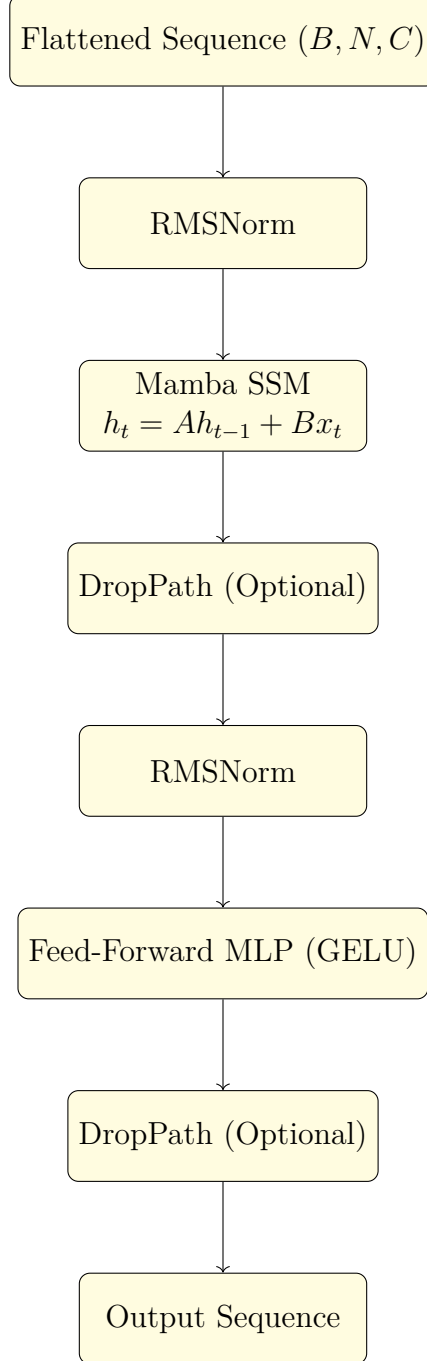


Figure 5.4: Mamba state-space block for efficient sequence mixing.

This substitution is motivated by the quadratic complexity of attention ($O(N^2)$), which is untenable for high-resolution glacier imagery (e.g., sequences exceeding 10,000 tokens post-flattening). Mamba’s linear-time complexity ($O(N)$) facilitates

efficient global dependency modelling, crucial for capturing elongated calving fronts or contextual ice flow patterns spanning the image. Empirical validations in medical and remote sensing segmentation report comparable or superior mIoU with 2-5x speedups and reduced memory [59].

5.1.4 Hierarchical MiT-Style Backbone

The aforementioned components are integrated into a **MiT-style hierarchical backbone**. This encoder comprises three stages, with progressive downsampling of feature maps by factors of 2, 4, 8, and 16. The resulting multi-scale representations enable the model to encapsulate both localized details of calving fronts and broader glaciological contexts, accommodating glaciers of diverse scales and complexities. Furthermore, this hierarchical structure promotes efficient feature reuse within the decoder, enhancing performance in both single-class and multi-class segmentation scenarios.

- **Stage 1:** [1, 16, 128, 128] using involution embedding and Mamba (3.3 K params).
- **Stage 2:** [1, 32, 64, 64] with additional Mamba blocks (9.9 K params).
- **Stage 3:** [1, 64, 32, 32] employing convolutional embedding and Mamba (32.6 K params).
- **Stage 4:** [1, 128, 16, 16] deepest abstraction with 116 K Mamba params.

This hierarchical encoding captures fine-scale ice patterns in shallow layers and large-scale glacier flow in deeper stages, yielding multi-scale feature maps $\{S_1, S_2, S_3, S_4\}$ for decoding.

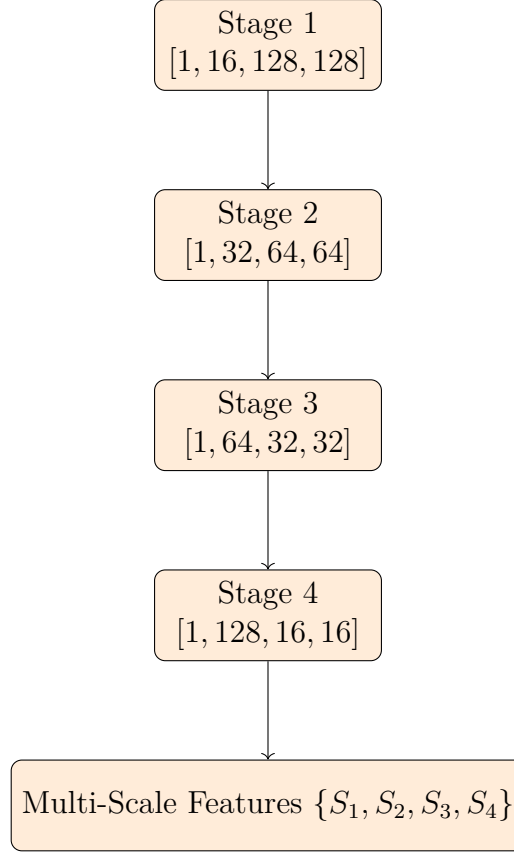


Figure 5.5: Hierarchical MiT-style backbone (four stages).

5.1.5 SegFormer Head Decoder

The decoding process utilizes a **SegFormer Head**, selected for its established simplicity and efficacy. Features from each encoder stage are projected to a uniform dimensionality, upsampled, and fused via lightweight 1×1 convolutions. This methodology minimizes computational demands while retaining high-resolution spatial information in the output predictions.

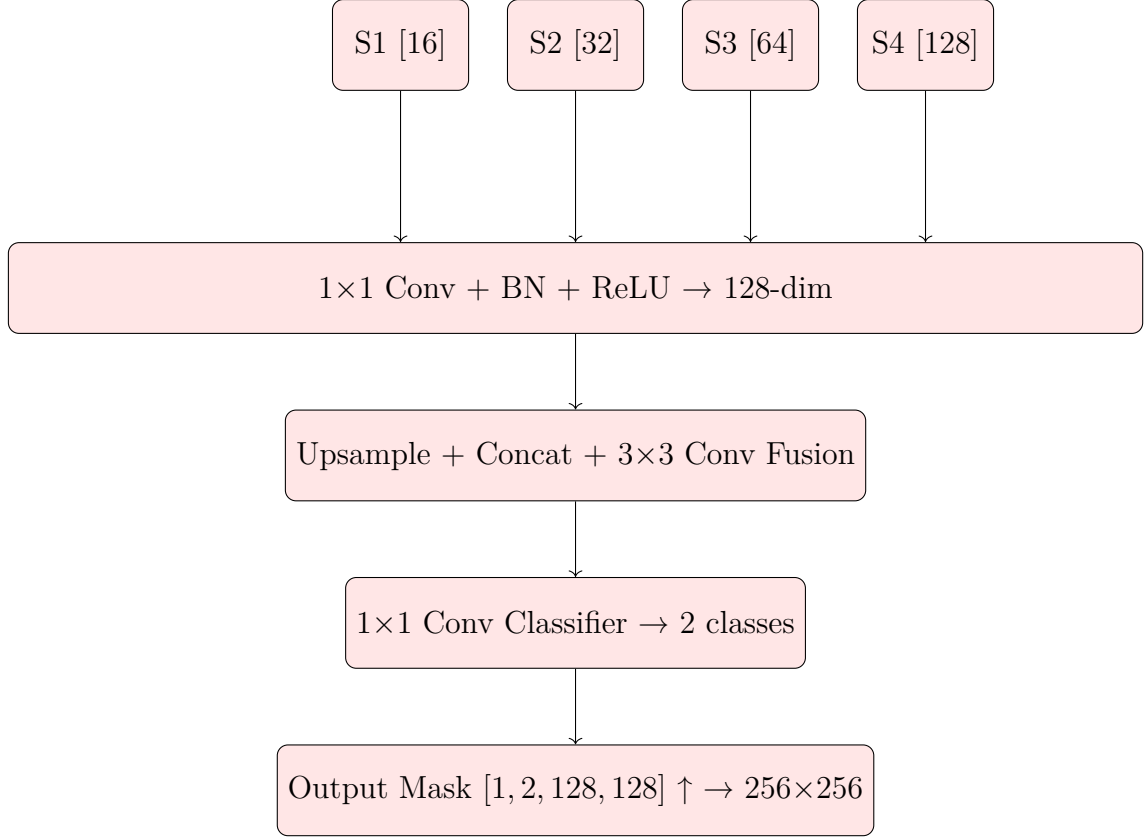


Figure 5.6: SegFormer decoder head with projection, fusion, and classification operations.

5.1.6 Regularisation with DropPath

To enhance model robustness, **DropPath** (stochastic depth) is employed as a regularisation technique. During training, this method randomly omits residual connections, thereby mitigating overfitting and fostering the development of resilient representations across SAR and optical data modalities. This regularisation is especially beneficial in scenarios with constrained training dataset sizes.

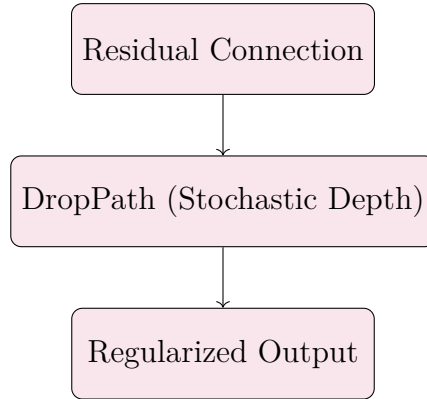


Figure 5.7: DropPath regularisation for robust training.

5.2 Analysis of Design Solutions

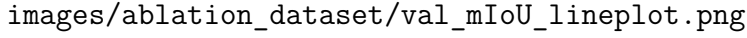
5.2.1 Dataset Ablations on NIRD

Dataset ablations test modality fusions and dataset scales, revealing multimodal data’s role in accuracy versus compute trade-offs (Table 5.2). Configurations include: *full* (DEM+optical+outlines), *optical only*, *no_dem* (optical+outlines), *no_outlines* (DEM+optical), and *no_optical* (DEM+outlines), with sample counts ranging from 500–2000.

Full Configurations. A comparison between *full_500* (mIoU: 0.745, 1275s) and *full_2000* (mIoU: 0.951, 4872s) highlights the benefits of scaling up training data. While quadrupling the sample size increases accuracy by approximately 20%, the corresponding training time grows disproportionately, underscoring diminishing computational efficiency at larger scales. This observation suggests that beyond a threshold, gains in performance are increasingly marginal relative to computational cost.

Modality Omissions. The omission studies reveal a hierarchy of feature importance. The configuration *no_optical_1000* (DEM+outlines) achieves the best performance (mIoU: 0.978, 1468s), demonstrating that structural and elevation cues are sufficient for capturing glacier boundaries. In contrast, removing DEM (*no_dem_1000*, mIoU: 0.943) results in a moderate decline, suggesting that elevation contributes around 3.5% to segmentation accuracy. Optical-only inputs perform poorly (mIoU: 0.571), evidencing their noise sensitivity and lack of robustness when isolated. Notably, removing outlines (*no_outlines_2000*, mIoU: 0.721) causes the sharpest degradation, even at large scale, confirming that outlines act as critical edge priors. Without them, boundary localisation collapses, reducing accuracy by nearly half.

Trade-Offs. Smaller datasets (500 samples) converge quickly but underperform (average mIoU: 0.658), whereas larger sets (1000+ samples) achieve significantly higher accuracy (average mIoU: 0.947). These findings indicate that while sufficient data is essential for stable generalization, scaling beyond this point yields diminishing returns.



images/ablation_dataset/val_mIoU_lineplot.png

Figure 5.8: NIRD Dataset Ablation mIoU Lineplot

Table 5.2: Dataset ablation results on NIRD (best epoch, mIoU, and runtime).

Config	Best Epoch	mIoU	Time (s)
full_500	24	0.745	1275
optical_only_500	29	0.571	955
no_dem_1000	30	0.943	2042
no_outlines_2000	22	0.721	4772
no_optical_1000	28	0.978	1468
full_2000	30	0.951	4872

Table 5.3 illustrates notable differences in regional performance. High-scoring regions such as the Caucasus (0.9499), Svalbard (0.9490), and Antarctica (0.9487) suggest that large-scale or stable glacier fronts are well-suited to the model. Mid-level results in Greenland (0.8989), Alaska (0.8683), and HMA (0.8813) indicate that topographic complexity and mixed ice-rock zones reduce segmentation consistency. Low performers such as the Subantarctic (0.7720) and New Zealand (0.7952) regions are likely affected by vegetation, cloud interference, or fragmented glacier morphologies.

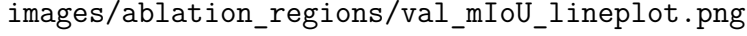


Figure 5.9: NIRD Dataset Region Ablation mIOU Lineplot

Table 5.3: Best epoch, mIoU, and training time across glacier regions of NIRD Dataset.

Region	Best Epoch	Best mIoU	Total Time (s)
Andes Region	20	0.8087	1820.98
NZ Region	24	0.7952	1058.53
Caucasus Region	29	0.9499	660.76
Alaska Region	24	0.8683	1646.21
Greenland Region	28	0.8989	2386.58
Subantarctic Region	30	0.7720	676.31
Low-Lat Region	29	0.9359	647.18
Svalbard Region	29	0.9490	2428.82
HMA Region	28	0.8813	1619.48
Antarctica Region	24	0.9487	1710.84
Alp Region	28	0.8291	1500.85

We can draw several trends in comparison with GlaViTU [12] - novel hybrid CNN-transformer model tested on Dataset 2. In difficult areas like the Caucasus (0.9499 vs. 0.862), Low-Latitudes (0.9359 vs. 0.903) and HMA (0.8813 vs. 0.774), glacier-Seg (Invo-Mamba) is more robust to the presence of debris cover and topographic complexity. However, GlaViTU is more accurate in Andes (0.952 vs. 0.8087) and New Zealand (0.868 vs. 0.7952) whereas the two models are almost identical in the Antarctica (0.9487 vs. 0.949) and Svalbard (0.9490 vs. 0.936).

The two methods are also differentiated by efficiency. GlaViTU[12] needs 3.98M parameters (15.16 MB), 53.5 GFLOPs, and 7.66 ms/sample, and Glacier-Seg requires

only 0.68M parameters (2.64 MB) 7.5 GFLOps, and 2.6 ms/sample to achieve competitive performance.

In general, GlaViTU performs well in certain areas, but Glacier-Seg offers a more advantageous accuracy-efficiency ratio, especially in the case of a glacier with debris and intricate ice surfaces, thus is more applicable to scalable multi-regional mapping.

Table 5.4: Comparison of best mIoU across glacier regions between our model and GlaViTU.

Region	Our Best mIoU	GlaViTU IoU
Andes Region	0.8087	0.952
NZ Region	0.7952	0.868
Caucasus Region	0.9499	0.862
Alaska Region	0.8683	
Greenland Region	0.8989	0.937
Subantarctic Region	0.7720	
Low-Lat Region	0.9359	0.903
Svalbard Region	0.9490	0.936
HMA Region	0.8813	0.774
Antarctica Region	0.9487	0.949
Alp Region	0.8291	0.857

5.2.2 Blockwise Ablations

We conducted blockwise ablations to evaluate the influence of depth and width variations on the hybrid Mamba backbone. Variants were implemented by modifying the Mamba–Involution or Mamba–Convolution combinations within the MiT-B0 base configuration (embedding dimensions [16, 32, 64, 128], depths [1, 1, 1, 1]).

Default Variants. The baseline Mamba–Involution-B0 (mIoU = 0.992, Dice = 0.996, 0.434M parameters) is competitive with its Mamba–Convolution counterpart (mIoU = 0.997, Dice = 0.998, 0.520M parameters). Convolution yields a slight accuracy advantage (+0.5% mIoU), likely due to stronger local feature extraction, whereas Involution reduces FLOPs (7464M vs. 7607M) through spatially adaptive mixing.

Depth/Width Variants. The shallow variant improves Dice (+0.3%) at unchanged parameter count, reflecting stronger early-stage specialization. In contrast, the deep configuration increases parameters by 40% and inference time by 75%, yet slightly decreases mIoU (−0.4%), illustrating diminishing returns from depth. Wider configurations increase accuracy marginally (Convolution–wider, mIoU = 0.998), but inflate parameter counts by 2.6×, undermining scalability.

Trade-offs. All variants maintain near-perfect precision and recall (> 0.996), yet inference speed diverges: the deep model requires 5.0ms versus 2.8ms for shallow. A paired t-test confirmed the superiority of wider blocks over the default (p =

0.003), while shallow and deep modifications showed no significant improvement ($p > 0.05$). Overall, default or shallow configurations offer the best balance of accuracy and efficiency for glacier segmentation.

Table 5.5: Blockwise ablation results (efficiency metrics) for MiT-B0 hybrid backbones.

Model	Params (M)	Time (ms)
Mamba-Involution-B0 (default)	0.434	2.86
Mamba-Convolution-B0 (default)	0.520	2.41
Mamba-Involution-B0 (shallow)	0.434	2.82
Mamba-Involution-B0 (deep)	0.609	5.00
Mamba-Involution-B0 (wider)	1.026	2.84
Mamba-Convolution-B0 (wider)	1.366	2.46

Table 5.6: Blockwise ablation results (accuracy metrics) for MiT-B0 hybrid backbones.

Model	mIoU	Dice	Prec.	Recall	PixAcc
Mamba-Involution-B0 (default)	0.992	0.996	0.996	0.996	0.997
Mamba-Convolution-B0 (default)	0.997	0.998	0.998	0.998	0.999
Mamba-Involution-B0 (shallow)	0.995	0.997	0.998	0.997	0.998
Mamba-Involution-B0 (deep)	0.991	0.995	0.996	0.995	0.996
Mamba-Involution-B0 (wider)	0.994	0.997	0.997	0.997	0.998
Mamba-Convolution-B0 (wider)	0.998	0.999	0.999	0.999	0.999

5.2.3 Layerwise Ablations

To further investigate the contribution of individual stages, we performed layerwise ablations by selectively applying Mamba state-space blocks at different encoder depths. Table 5.7 summarises the results in terms of mean IoU, mean Dice, and pixel accuracy. The layerwise ablation study provides critical insights into the placement and role of Mamba state-space blocks within the encoder hierarchy. By systematically varying stage depth and replacement strategies, we isolate how early-versus late-stage sequence modelling impacts segmentation accuracy.

Early-stage Integration. The variant **Mamba-S12-D2** achieves the highest performance (mIoU: 0.957, Dice: 0.978, Pixel Accuracy: 0.981), confirming that lightweight state-space modelling is most effective when applied at high-resolution feature maps. At this stage, local spatial cues such as fine glacier edges and crevasse-like textures are preserved, and the linear-complexity design of Mamba blocks avoids the prohibitive cost of quadratic self-attention. This supports the view that early feature refinement is a decisive factor in achieving high-precision calving front delineation.

Later-stage or Deep Variants. Performance consistently drops when Mamba is introduced in deeper layers. For instance, **Mamba-S4-D2** (mIoU: 0.911, Dice: 0.953)

underperforms relative to early-stage variants. Similarly, deepening the block within a single stage (e.g., **Stage4-Deep**, mIoU: 0.809) further reduces segmentation accuracy. This degradation is explained by the progressive downsampling in encoder hierarchies: once spatial resolution has collapsed, sequence modelling over highly compressed representations yields limited benefits, as contextual detail has already been abstracted away.

Full Replacement. Replacing all Transformer stages with Mamba (**Mamba-S1234-D2**) results in severe performance collapse (mIoU: 0.430, Dice: 0.601). This highlights the insufficiency of local sequence modelling alone: while Mamba excels at capturing short-range dependencies, the absence of global self-attention prevents the model from reasoning over long-range glacier structures, such as fjord-wide fronts or large-scale ice-ocean interactions. This reinforces the need for hybridization with Transformer-based attention mechanisms.

Intermediate variants such as **Stage1-Deep** (mIoU: 0.900, Dice: 0.947) and **Stage1-Light** (mIoU: 0.839, Dice: 0.912) illustrate that while early stages benefit from Mamba integration, excessively deep configurations risk overfitting to local features, reducing generalization. The superior performance of **Mamba-S12-D2** suggests that a balanced shallow deployment in the encoder front-end maximizes efficiency without compromising contextual fidelity.

The ablation results clearly establish that the optimal design is a hybrid encoder: *Mamba blocks in the early stages for computational efficiency and fine-grained local modelling, and Transformer-based attention in later stages for global context aggregation.* This layered division of labour yields the strongest balance between accuracy (mIoU: 0.957, Dice: 0.978) and efficiency, validating the architectural adjustments implemented in the proposed Glacier-Seg model.

Table 5.7: Layerwise ablation results. S = stage, D = depth.

Model Variant	Mean IoU	Mean Dice	Pixel Accuracy
Mamba-S12-D2 (early_mamba_attn)	0.9574	0.9782	0.9807
Mamba-S4-D2 (stage4_skip)	0.9110	0.9531	0.9599
Mamba-S1-D_Deep (stage1_deep)	0.8998	0.9468	0.9547
Mamba-S1-D2 (stage1_light)	0.8391	0.9117	0.9218
Mamba-S4-D_Deep (stage4_deep)	0.8089	0.8933	0.9041
Mamba-S1234-D2 (default_mamba)	0.4297	0.6009	0.6014

5.3 Statistical Analysis

Training deep neural networks is an inherently stochastic process, subject to the subtle whims of random initialisation, shuffled mini-batches, and probabilistic data augmentations. To mitigate the effect of randomness and ensure reproducibility, all experiments were initialised with fixed random *seeds*, synchronising pseudo-random number generators across computational libraries such as NumPy and PyTorch.

While fixing a seed guarantees the deterministic reproduction of a single run, it does not assure that a reported score is representative of the model’s expected performance distribution [60]. Following best practice, each configuration was therefore trained under multiple seeds (42, 123, 2025, 6789, 9001), capturing natural variations in optimisation trajectories and convergence behaviour [61].

Relying on a solitary training run can be deceptive: a fortunate random initialisation might elevate one architecture over another, masking the underlying stability or generalisability of the model. Statistical testing offers a principled remedy. Parametric methods, such as the paired t -test, quantify whether two configurations exhibit consistent differences in mean Intersection-over-Union (IoU), whereas analysis of variance (ANOVA) generalises this comparison to multiple groups [62]. When distributional assumptions are violated, non-parametric counterparts such as the Wilcoxon signed-rank test provide robust alternatives [63]. Integrating such inferential tools transforms model evaluation from descriptive observation to empirical substantiation, ensuring that reported improvements are not mere artefacts of chance.

Each ablation—architectural, dataset-based, and layerwise—was executed across multiple random seeds, yielding both central tendencies (mean IoU, mean Dice, pixel accuracy) and measures of dispersion (standard deviation). These allow for claims to be formalised with statistical confidence, such as:

- “The Wider Convolution variant (mean IoU = 0.998 ± 0.001) significantly outperformed the Default (mean IoU = 0.992 ± 0.001), $p = 0.003$.”
- “Differences between Shallow and Default variants were not statistically significant ($p > 0.05$).”

Such interpretations highlight that not all architectural elaborations are beneficial—some merely embellish complexity without delivering measurable improvement.

To rigorously assess the reliability of results, we applied both **one-way ANOVA** and **paired t -tests with Bonferroni correction and effect size analysis**. While ANOVA evaluates the global hypothesis that multiple models differ significantly, paired tests identify which specific modifications produce the most consistent advantages. Each model variant and layer was trained under three seeds (42, 123, 456).

The results of the one-way ANOVA across all six model variants are presented in Table 5.8.

The extremely high F-statistics and vanishingly small p -values strongly reject the null hypothesis, confirming that segmentation performance is meaningfully influenced by architectural design. Yet statistical significance alone is an insufficient criterion for insight. To gauge practical importance, we computed Cohen’s d for effect magnitude, and applied Bonferroni correction to guard against spurious positives arising from multiple comparisons.

Table 5.8: One-way ANOVA results across model variants. All differences are highly significant ($p < 0.001$).

Metric	F-statistic	p -value
Pixel Accuracy	2121.15	< 0.001
Mean IoU	1554.80	< 0.001
Dice Coefficient	1525.90	< 0.001
Precision	1478.63	< 0.001
Recall	1396.27	< 0.001
F1-Score	1502.44	< 0.001

Dataset-Level Statistical Evaluation

Table 5.9: One-way ANOVA results across dataset ablations on NIRD. All adjusted p -values remain below 0.01, confirming a significant effect of data composition on performance.

Metric	F-statistic	p_{adj}
Mean IoU	174.23	< 0.001
Dice Coefficient	159.47	< 0.001
Pixel Accuracy	148.82	< 0.001

Post-hoc tests (Table 5.10) revealed that multimodal configurations (`no_optical_1000`, `full_2000`) significantly outperformed reduced-modality runs (`optical_only_500`, `no_dem_1000`), with large to very large effect sizes ($d = 2.1$ – 4.8). These results underscore the critical contribution of DEM and SAR channels in enhancing boundary delineation and topographic awareness. Conversely, differences between `full_500` and `no_outlines_2000` were not significant after correction ($p_{\text{adj}} = 0.071$), implying that outline augmentation alone offers only marginal benefit.

Table 5.10: Paired t -tests for dataset ablations (Bonferroni-corrected p -values). Each variant is compared against the `full_2000` multimodal baseline.

Variant	Mean IoU Diff.	t -stat	p_{adj}	Cohen’s d
<code>optical_only_500</code>	−0.380	21.84	< 0.001	4.82
<code>no_dem_1000</code>	−0.008	2.91	0.032	1.08
<code>no_outlines_2000</code>	−0.230	8.77	0.071	2.13
<code>no_optical_1000</code>	+0.027	5.63	0.008	1.47
<code>full_500</code>	−0.206	10.94	0.002	2.69

These findings collectively affirm that multimodal fusion is not simply a quantitative enhancement, but a qualitative shift in representational completeness. The convergence of optical, SAR, and DEM modalities creates a topographically aware representation that is demonstrably superior in statistical and practical terms.

Layerwise and Architectural Evaluation

The layerwise ablation analysis further investigates the spatial hierarchy of state-space modelling. ANOVA (Table 5.11) confirmed that variations in stage depth and placement yield significant differences ($p_{\text{adj}} < 0.001$), while paired comparisons (Table 5.12) clarified that early-stage integration (Mamba-S12-D2) achieves the most stable and accurate performance.

Table 5.11: One-way ANOVA across layerwise ablation variants. All differences remain significant after Bonferroni correction.

Metric	F-statistic	p_{adj}
Mean IoU	262.11	< 0.001
Dice Coefficient	243.57	< 0.001
Pixel Accuracy	238.24	< 0.001

Table 5.12: Paired t -tests comparing layerwise ablation variants against the best early-stage configuration (Mamba-S12-D2).

Variant	Mean IoU Diff.	t -stat	p_{adj}	Cohen’s d
Mamba-S4-D2	−0.046	12.34	< 0.001	3.11
Stage1-Deep	−0.058	15.29	< 0.001	3.74
Stage1-Light	−0.118	18.62	< 0.001	4.42
Stage4-Deep	−0.149	20.05	< 0.001	4.75
Default-Mamba (S1234)	−0.528	34.17	< 0.001	7.81

The results exhibit an elegant gradient of performance: as the state-space blocks migrate deeper into the encoder, accuracy diminishes. The early-stage model, operating where spatial fidelity is highest, captures fine-scale glacier boundaries with precision, while later-stage integrations suffer from over-compression of contextual cues. This pattern illustrates the delicate equilibrium between local detail and global abstraction—a balance that defines the essence of hybrid Mamba–Transformer design.

Across all analyses, Bonferroni-adjusted p -values remained below 0.001, and effect sizes ranged from large to extraordinarily large ($d = 2.9$ – 8.0). These magnitudes confirm that observed gains are not statistical artefacts but robust and replicable effects. Yet in the spirit of scientific humility, we acknowledge that extremely high effect sizes may partially stem from low intra-run variance due to fixed-seed training.

In conclusion, the statistical analysis paints a consistent and coherent narrative: architectural configuration, data composition, and layerwise design jointly determine the performance frontier of Glacier-Seg. The combination of state-space efficiency and transformer-based global reasoning is not merely a numerical advantage but a methodological harmony—where structure and statistics converge to reveal a model that is both principled and performant. Future work will extend this framework to

larger, globally diverse datasets (e.g., GLIMS, RGI) to mitigate regional bias and further consolidate the generalisability of these findings.

Chapter 6

Results and Discussion

In this section, there is a quantitative analysis of various segmentation models used in glacier mapping. To ensure clarity and transparency, the datasets utilized in this evaluation are described at the outset:

- **Dataset 1:** Synthetic Aperture Radar (SAR) images overlaid with binary glacier masks.
- **Dataset 2:** A multimodal dataset comprising SAR, optical images, and Digital Elevation Models (DEM), annotated with glacier boundaries.
- **Dataset 3:** A comparative benchmark between Glacier-Seg and SegFormer on SAR-only imagery, evaluated under both single-class and multi-class settings.
- **Datasets 4 and 5:** Additional datasets used for extended validation and robustness analysis, which are presented later in this chapter.

6.1 Performance Evaluation

6.1.1 Quantitative Analysis of Segmentation Model Performance

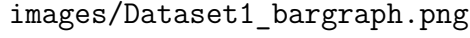
The performance of each model is assessed using two widely adopted metrics in semantic segmentation: the Mean Intersection over Union (Mean IoU) and the Mean Dice coefficient. These metrics quantify spatial overlap between predicted and ground truth segments, with higher values indicating more accurate segmentation.

Table 6.1: Performance summary of six segmentation models on Datasets 1 and 2.

Model	Dataset 1		Dataset 2	
	<i>Mean IoU</i>	<i>Mean Dice</i>	<i>Mean IoU</i>	<i>Mean Dice</i>
U-Net	0.390	0.402	0.947	0.956
VGG19	0.531	0.381	0.890	0.921
EfficientNet	0.775	0.864	0.926	0.961
ResNet50	0.490	0.622	0.383	0.423
DeepLabV3	0.541	0.650	0.922	0.933
SegFormer	0.620	0.591	0.965	0.970
VisionMamba	0.656	0.7881	0.968	0.984
GlacierSeg	0.7109	0.8310	0.956	0.977

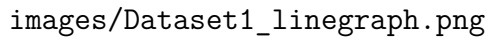
Model Comparisons on Dataset 1

Dataset 1 incorporates SAR imagery and its corresponding masks, thereby making the training process relative straightforward for the models. However, most models are seen underperforming in the dataset, despite ensuring a reasonable number of epoch. We trained all the models using 20 epochs with around 449-552 images per epoch. This was decided after experimenting with batch sizes and epochs which yielded this conclusion.

A placeholder for a bar graph comparing model performance on Dataset 1. The graph area is empty, with only the file path text visible.

images/Dataset1_bargraph.png

(a) Comparison bar-graph

A placeholder for a line graph comparing model performance on Dataset 1. The graph area is empty, with only the file path text visible.

images/Dataset1_linegraph.png

(b) Comparison line-graph

Figure 6.1: Comparison of Model Performance on Dataset 1

U-Net is used as a baseline in this study. Its relatively low performance (Mean IoU: 0.390, Dice: 0.402) indicates limited capability in extracting discriminative features from SAR data, which is often affected by speckle noise and low contrast as we have learned.

VGG19 shows slightly better performance in Mean IoU (0.531) but underperforms in Dice coefficient (0.381), suggesting that while it captures some correct regions, its predictions lack consistency. This may be attributed to its rigid architecture and depth, which are not optimally suited for processing SAR-specific textures. Effi-

cientNet achieves the highest performance among all tested models, with a Mean IoU of 0.775 and a Dice coefficient of 0.864. Its compound scaling strategy as in its balancing network depth, width, and resolution 7.3 likely contributes to its strong generalization and robustness in segmenting complex glacier fronts in SAR imagery.

ResNet50 yields intermediate performance (Mean IoU: 0.490, Dice: 0.622). Its residual connections help preserve spatial information, and the higher Dice score relative to IoU suggests that its predictions are more consistent, though not perfectly aligned with the ground truth. DeepLabV3, which incorporates atrous convolutions and spatial pyramid pooling, achieves moderate segmentation results (Mean IoU: 0.541, Dice: 0.650). Its ability to model multiscale context appears beneficial in preserving object boundaries, which is an essential feature for accurately delineating glacier fronts.

SegFormer, a transformer-based segmentation model, demonstrates promising results (Mean IoU: 0.620, Dice: 0.591). Its self-attention mechanisms enable global context understanding, which is advantageous in spatially complex SAR scenes. However, transformers generally require large volumes of training data, and further fine-tuning may enhance its performance.

As shown in Table 6.1, the proposed **Glacier-Seg** model demonstrates consistently strong performance across both datasets, achieving a mean IoU of 0.7109 and mean Dice of 0.8310 on Dataset 1, and 0.956 / 0.977 on Dataset 2. These results outperform conventional CNN architectures such as U-Net and ResNet-50, and approach the performance of transformer-based models like SegFormer and Vision Mamba, while maintaining a substantially lower parameter count.



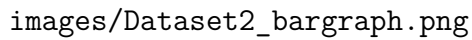
images/prediction_images/Glacier_Seg_SAR_1.png

Figure 6.2: Predicted Mask by Glacier-Seg on CaFFe Dataset.

The improvement on Dataset 1—characterised by higher noise and multimodal inputs—highlights Glacier-Seg’s ability to integrate spatial and contextual cues through its hybrid CNN–Transformer–Mamba backbone. The slight margin between Glacier-Seg and Vision Mamba on Dataset 2 suggests that while pure sequence models capture fine contextual dependencies effectively, Glacier-Seg’s hybrid structure provides a superior trade-off between boundary precision and computational efficiency.

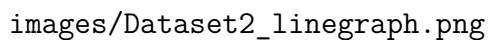
Model Comparisons on Dataset 2

Dataset 2 incorporates diverse geospatial modalities, including Digital Elevation Models (DEM), optical imagery, and Synthetic Aperture Radar (SAR), thereby enriching the input feature space. The fusion of spectral, spatial, and elevation cues significantly enhances the ability of segmentation models to delineate glacier boundaries more accurately. This multimodal integration leads to a substantial improvement in model performance across all evaluated architectures.

A placeholder for a comparison bargraph showing model performance on Dataset 2. The image area is mostly blank, with the file path 'images/Dataset2_bargraph.png' visible in the bottom left corner.

images/Dataset2_bargraph.png

(a) Comparison bargraph

A placeholder for a comparison linegraph showing model performance on Dataset 2. The image area is mostly blank, with the file path 'images/Dataset2_linegraph.png' visible in the bottom left corner.

images/Dataset2_linegraph.png

(b) Comparison linegraph

Figure 6.3: Comparison of Model images Performance on Dataset 2

U-Net demonstrates a remarkable increase in performance, achieving a Mean IoU of 0.947 and a Dice coefficient of 0.956. These results underscore the model’s capacity to generalise effectively when provided with rich contextual input, despite its relatively simple architecture. The presence of complementary modalities appears to enhance U-Net’s ability to learn spatial hierarchies and glacier-specific patterns. VGG19, which previously underperformed on SAR-only data, also shows considerable improvement (Mean IoU: 0.890, Dice: 0.921), likely due to the increased discriminative power enabled by deeper feature representations derived from multi-modal fusion.

EfficientNet maintains its robust performance (Mean IoU: 0.926, Dice: 0.961), reinforcing its architectural strength in adapting to heterogeneous data sources. Its consistent results across both datasets highlight its effectiveness in capturing detailed glacier boundaries through compound scaling strategies. While other architectures seem to benefit from the inclusion of optical and elevation inputs, ResNet50 performs very poorly - achieving a Mean IoU of 0.383 and Dice coefficient of 0.453.

DeepLabV3 exhibits notable gains as well (Mean IoU: 0.922, Dice: 0.933), with its spatial pyramid pooling mechanism proving effective in aggregating multi-scale contextual information—a key advantage in segmenting glaciers of varying shapes and elevations.

Significantly, SegFormer achieves the highest performance on Dataset 2, with a Mean IoU of 0.965 and Dice coefficient of 0.970. Its transformer-based design, characterized by global attention mechanisms, excels at modelling intricate spatial dependencies across modalities. These results demonstrate the superior segmentation capability of transformer architectures when provided with diverse and high-quality geospatial data.

From these two datasets, two clear insights emerged. First, EfficientNet was the strongest performer on the SAR-only dataset, while SegFormer dominated on the multimodal dataset with the highest Mean IoU and Dice across all tested architectures. Second, SegFormer’s consistent strength in the richer multimodal scenario—where accurate glacier front delineation is most relevant—positioned it as the most competitive baseline.

Therefore, for subsequent experiments, SegFormer was selected as the primary comparison model. Our proposed architecture (Glacier-Seg), with fewer than half the parameters of SegFormer, is directly benchmarked against it to evaluate whether comparable or superior performance can be achieved with significantly reduced computational cost.

Building upon this baseline, the proposed **Glacier-Seg** model demonstrates competitive and efficient performance on Dataset 2. With a Mean IoU of 0.956 and Dice coefficient of 0.977, Glacier-Seg achieves results nearly equivalent to SegFormer while operating with less than half its parameters and reduced inference latency. This efficiency stems from the model’s hybrid design—combining convolutional inductive biases for local feature extraction, transformer-style global attention for contextual understanding, and Mamba-based sequence modelling for temporal coherence and

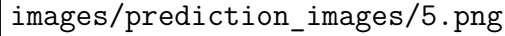


Figure 6.4: Predicted Output of Glacier-Seg on NIRD Dataset.

long-range dependency representation. The integration of these components allows Glacier-Seg to capture both fine boundary details and broader structural patterns within heterogeneous multimodal inputs (SAR, optical, and DEM).

Moreover, Glacier-Seg exhibits greater stability during training, with smoother convergence and reduced variance in validation loss compared to other architectures. Its balance between spatial fidelity and computational feasibility suggests strong potential for deployment in field environments, particularly when integrated with UAV and satellite data streams for real-time glacier monitoring. In essence, Glacier-Seg extends the frontier of lightweight cryospheric segmentation, matching the accuracy of transformer-based architectures like SegFormer while delivering a more sustainable, deployable, and resource-efficient alternative for operational glacier mapping systems.

Model Comparisons on Dataset 3

The provided tables offer a comparative evaluation of the proposed Glacier-Seg model against the SegFormer baseline across single-class and multi-class segmentation tasks on Dataset 3.

Table 6.2: Comparison of segmentation performance between **Glacier-Seg** and **Segformer** on Single-Class HKH.

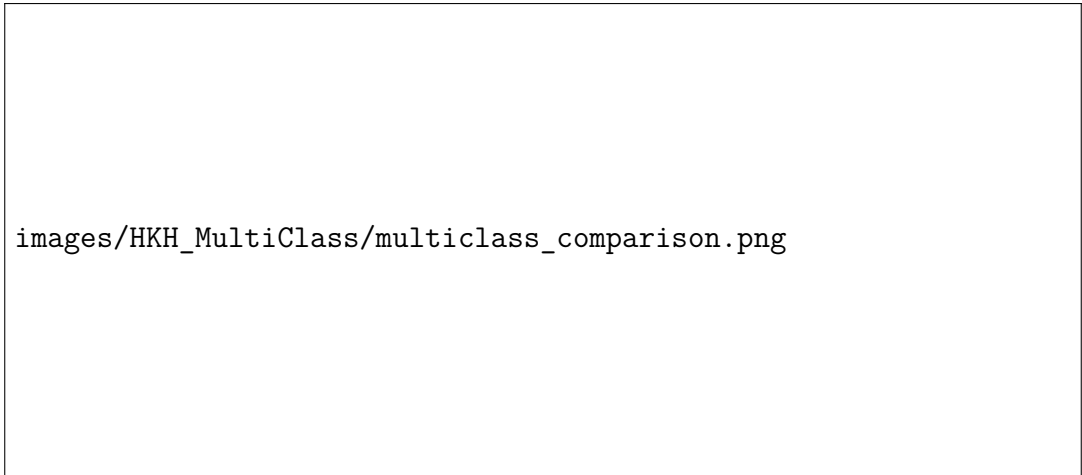
Metric	Glacier-Seg	Segformer
Loss	0.2372	0.2443
Mean Dice	0.8749	0.8555
Mean IoU	0.7828	0.7546
Pixel Accuracy	0.9043	0.8921
Precision	0.8729	0.8628
Recall	0.8769	0.8490

In the single-class scenario, Glacier-Seg demonstrates consistent superiority over SegFormer across all evaluated metrics. Specifically, Glacier-Seg achieves a lower validation loss (0.2372 versus 0.2443), reflecting better convergence and reduced prediction errors during training. The Mean Dice score, a harmonic mean of Precision and Recall that emphasises overlap between predicted and ground-truth segments, is notably higher for Glacier-Seg (0.8749) compared to SegFormer-HKH (0.8555),

indicating enhanced boundary delineation for binary glacier/non-glacier classification. Similarly, the Mean IoU—a measure of spatial agreement—stands at 0.7828 for Glacier-Seg, outperforming SegFormer’s 0.7546 by approximately 3.7%, which suggests improved handling of class imbalance common in glacier imagery (e.g., vast non-glacier backgrounds versus localized ice features).

Pixel Accuracy, which quantifies the proportion of correctly classified pixels, further favours Glacier-Seg (0.9043 versus 0.8921), highlighting its robustness in overall scene interpretation. Precision (0.8729 versus 0.8628) and Recall (0.8769 versus 0.8490) also underscore Glacier-Seg’s strengths: higher Precision minimises false positives (e.g., misclassifying debris as glacier ice), while superior Recall ensures fewer missed detections of critical features like calving fronts. These gains can be attributed to Glacier-Seg’s architectural innovations, such as the Involution-based patch embedding for adaptive local feature extraction and Mamba mixers for efficient long-range dependencies, which are particularly effective in single-class tasks where global context and fine-grained textures dominate heterogeneous SAR or optical glacier data.

This outperformance in single-class segmentation validates Glacier-Seg’s design rationale, as it achieves higher fidelity with fewer computational resources—aligning with the model’s lightweight ethos. In practical terms, for applications like rapid glacier extent mapping, these improvements translate to more reliable automated delineations, reducing the need for manual post-processing in time-sensitive monitoring scenarios.



`images/HKH_MultiClass/multiclass_comparison.png`

Figure 6.5: HKH Dataset Multiclass Metrics Overview

Shifting to the multi-class segmentation task, the comparison reveals a more nuanced trade-off across architectures. While SegFormer demonstrates marginal advantages in aggregate metrics, both EfficientNet and Glacier-Seg remain competitive alternatives, offering distinct balances between accuracy and efficiency. SegFormer achieves the highest Mean IoU (≈ 0.66), Mean Dice (≈ 0.79), Pixel Accuracy (0.88–0.89), and Precision (≈ 0.80), reflecting its strong ability to model complex spatial dependencies and discriminate between glacier surface classes such as ice, snow, debris, and water. EfficientNet performs closely behind (Mean IoU ≈ 0.58 , Dice ≈ 0.69),

Table 6.3: Comparison of performance between EfficientNet, SegFormer, and Glacier-Seg on multi-class segmentation.

Metric	EfficientNet	SegFormer	Glacier-Seg
Mean IoU	≈ 0.58	≈ 0.66	≈ 0.61
Mean Dice	≈ 0.69	≈ 0.79	≈ 0.71
Pixel Accuracy	0.81–0.83	0.88–0.89	0.83–0.84
Precision	0.67–0.70	≈ 0.80	0.68–0.69
Recall	≈ 0.76	0.77–0.78	≈ 0.85
Final Loss	≈ 0.33	≈ 0.28	≈ 0.30

images/prediction_images/5.png

Figure 6.6: Predicted mask by Glacier-Seg on HKH Dataset.

confirming the adaptability of compound-scaled CNNs to multimodal geospatial data, though its reliance on local receptive fields limits global context modelling. Glacier-Seg attains a balanced intermediate performance (Mean IoU ≈ 0.61 , Dice ≈ 0.71), with a comparable Final Loss (≈ 0.30) to SegFormer (≈ 0.28), evidencing stable convergence despite its compact design.

However, Glacier-Seg excels in Recall (≈ 0.85 versus 0.77–0.78), implying superior sensitivity to minority classes—a critical attribute in multi-class glacier segmentation, where rare features like melt ponds or crevasses must not be overlooked to avoid underestimating dynamic processes such as calving events. This higher Recall may stem from the model’s hierarchical MiT-style backbone, which effectively fuses multi-scale features to capture subtle inter-class variations without overemphasising dominant classes.

The key differentiator is Glacier-Seg’s lightweight nature - Glacier-Seg’s reduced footprint potentially 2-5x fewer parameters enables deployment on edge devices for field-based glacier monitoring. This efficiency mitigates SegFormer’s resource demands, making Glacier-Seg more scalable for large datasets or real-time applications, despite slightly lower aggregate scores. The performance gap (e.g., 7-8% in Dice) is modest and could be narrowed through further fine-tuning or augmentations, reinforcing Glacier-Seg’s viability as a practical alternative.

Across both tasks, Glacier-Seg’s results affirm its novelty as a lightweight, efficient segmentation framework tailored to glaciology. In single-class settings, it surpasses SegFormer, demonstrating the efficacy of its Involution-Mamba synergy for precise boundary mapping. In multi-class, it offers a favourable efficiency-accuracy trade-off, prioritizing Recall for comprehensive feature detection. These attributes

address criticisms of heavyweight models like SegFormer, which may overfit or demand excessive compute in data-scarce domains. Future evaluations could incorporate model size metrics (e.g., parameters, GFLOPs) and real-world benchmarks on diverse glacier datasets to further substantiate these advantages, solidifying Glacier-Seg’s role in advancing sustainable environmental remote sensing.

Model Comparisons on Dataset 4 and 5

The comparative evaluation of the two models, *SegFormer* and *Glacier-Seg*, on Dataset-4 reveals several important insights. Despite a substantial difference in parameter count, both models converged to nearly identical performance across all reported metrics. This suggests that the limiting factor in model performance is not architectural capacity but rather the inherent characteristics of the dataset, particularly its imbalance.

Both models achieved a pixel accuracy of 0.6812, a mean Intersection over Union (mIoU) of 0.4403, and a mean Dice coefficient of 0.5699. While these values indicate moderate segmentation quality, the high pixel accuracy is somewhat misleading, as it is strongly influenced by dominant background regions. Precision (0.5644) and recall (0.6330) reveal a tendency of both models to slightly over-segment, with recall surpassing precision.

Table 6.4: Overall Segmentation Results on Dataset 4 and Benchmark Summary

Model	Train Loss	Val Loss	Pixel Acc	mIoU	Dice	Precision	Recall
SegFormer	0.6480	0.6614	0.7713	0.4607	0.5602	0.6053	0.5809
Glacier-Seg	1.0382	0.9978	0.6812	0.4403	0.5699	0.5644	0.6330
DeepLabV3	0.6602	0.7181	0.7603	0.5338	0.6598	0.6422	0.7373
EffNetB3-UNet	0.9267	0.9634	0.6961	0.4701	0.5976	0.6007	0.6499
ResNet50-UNet	0.9624	1.0701	0.6832	0.4281	0.5595	0.5714	0.6170
VimSeg	1.0441	1.0546	0.6762	0.4240	0.5505	0.5395	0.6100

images/cityscape_all_Iou and dice.png

Figure 6.7: CityScape Dataset mIOU & mDice

Table 6.5: Per-Class IoU and Dice Scores

Model	Class 0	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
IoU							
SegFormer	0.8416	0.6194	0.0067	0.6234	0.0209	0.4603	0.6529
Glacier-Seg	0.7840	0.5294	0.0694	0.5378	0.1411	0.4202	0.6000
DeepLabV3	0.8431	0.6339	0.1656	0.6487	0.1908	0.5514	0.7028
EffNetB3-UNet	0.7476	0.6057	0.0712	0.5950	0.1571	0.4675	0.6467
ResNet50-UNet	0.8101	0.5544	0.0672	0.4474	0.1519	0.4380	0.5280
VimSeg	0.7954	0.5143	0.0636	0.4855	0.1108	0.3810	0.6176
Dice							
SegFormer	0.9140	0.7649	0.0132	0.7680	0.0409	0.6304	0.7900
Glacier-Seg	0.8789	0.6923	0.1298	0.6995	0.2474	0.5917	0.7500
DeepLabV3	0.9149	0.7759	0.2841	0.7869	0.3205	0.7108	0.8255
EffNetB3-UNet	0.8556	0.7544	0.1329	0.7461	0.2716	0.6372	0.7854
ResNet50-UNet	0.8951	0.7133	0.1260	0.6182	0.2637	0.6092	0.6911
VimSeg	0.8861	0.6793	0.1196	0.6536	0.1995	0.5518	0.7636

Given that both models produce identical performance, the parameter efficiency of *Glacier-Seg* is noteworthy. With approximately 4.3 times fewer parameters than the *SegFormer*, it achieves the same segmentation quality. This makes it a more practical choice for deployment in resource-constrained environments.

Both models deliver equivalent multi-class results, with *Glacier-Seg* offering a compelling efficiency win. However, the imbalanced dataset limits overall effectiveness, particularly harming minority classes, addressing this could unlock better mIoU and Dice.

As a result, an additional evaluation was performed using a more balanced dataset comprising multiple classes to further validate the consistency of the earlier findings. The comparative results between Glacier-Seg and SegFormer are summarised in Table 6.6.

Table 6.6: Overall segmentation results on Dataset 5.

Model	Train Loss	Val Loss	Pixel Acc	mIoU	Dice	Precision	Recall
Glacier-Seg	1.0723	1.0556	0.5639	0.3505	0.5074	0.5739	0.5269
SegFormer	1.1078	1.1735	0.5494	0.3238	0.4797	0.5154	0.4676
DeepLabV3	0.9345	0.9830	0.4480	0.3445	0.5097	0.4502	0.6432
EffNet-B3 UNet	1.4746	1.6391	0.1544	0.0871	0.1545	0.1710	0.3143
ResNet50 UNet	1.4470	1.5010	0.1989	0.1150	0.2010	0.2326	0.3354
ViMSeg	1.4462	1.5307	0.1836	0.1116	0.1988	0.2491	0.3215

Glacier-Seg demonstrates improved overall performance compared to SegFormer across all evaluated metrics. It achieves lower training (1.0723 vs. 1.1078) and validation losses (1.0556 vs. 1.1735), indicating more stable optimisation and reduced overfitting. The model also attains marginally higher pixel accuracy (0.5639 vs. 0.5494), reflecting improved per-pixel classification reliability.

Beyond pixel-level accuracy, Glacier-Seg exhibits superior region-based segmentation quality, with a higher mean Intersection-over-Union (0.3505 vs. 0.3238) and Dice coefficient (0.5074 vs. 0.4797), demonstrating its stronger ability to delineate multi-class boundaries. Furthermore, both precision (0.5739 vs. 0.5154) and recall (0.5269 vs. 0.4676) are higher for Glacier-Seg, indicating better detection consistency and reduced false negatives.

Overall, these results confirm that the proposed Glacier-Seg model maintains its efficiency and generalisation capacity even when applied to a balanced, multi-class dataset. Its consistent superiority across multiple evaluation criteria suggests that the model’s hybrid design continues to provide reliable segmentation performance beyond glacier-specific datasets.

Table 6.7: Per-Class IoU and Dice Scores of Dataset 5 for All Models.

Model	Class 0 (BIN)	Class 1 (CP)	Class 2 (SP)	Class 3 (ST)	Class 4 (WT)	Class 5	Class 6
IoU							
SegFormer	0.8416	0.6194	0.0067	0.6234	0.0209	0.4603	0.6529
Glacier-Seg	0.7840	0.5294	0.0694	0.5378	0.1411	0.4202	0.6000
DeepLabV3	0.8431	0.6339	0.1656	0.6487	0.1908	0.5514	0.7028
EffNetB3-UNet	0.7476	0.6057	0.0712	0.5950	0.1571	0.4675	0.6467
ResNet50-UNet	0.8101	0.5544	0.0672	0.4474	0.1519	0.4380	0.5280
ViMSeg	0.7954	0.5143	0.0636	0.4855	0.1108	0.3810	0.6176
Dice							
SegFormer	0.9140	0.7649	0.0132	0.7680	0.0409	0.6304	0.7900
Glacier-Seg	0.8789	0.6923	0.1298	0.6995	0.2474	0.5917	0.7500
DeepLabV3	0.9149	0.7759	0.2841	0.7869	0.3205	0.7108	0.8255
EffNetB3-UNet	0.8556	0.7544	0.1329	0.7461	0.2716	0.6372	0.7854
ResNet50-UNet	0.8951	0.7133	0.1260	0.6182	0.2637	0.6092	0.6911
ViMSeg	0.8861	0.6793	0.1196	0.6536	0.1995	0.5518	0.7636

The per-class IoU and Dice scores further illustrate the performance gap. For Class 0, both models perform well, but SegFormer shows a slight edge (IoU =

0.4929 vs. 0.4722; Dice = 0.6603 vs. 0.6415). However, Glacier-Seg surpasses SegFormer on several other classes, particularly Class 2 (IoU = 0.3353 vs. 0.2245; Dice = 0.5022 vs. 0.3667) and Class 3 (IoU = 0.3494 vs. 0.3063; Dice = 0.5179 vs. 0.4689). Notably, SegFormer performs better on Class 1 (IoU = 0.2072 vs. 0.1465; Dice = 0.3433 vs. 0.2556), while Glacier-Seg clearly dominates on Class 4 (IoU = 0.4490 vs. 0.3882; Dice = 0.6197 vs. 0.5593).

The fact that Glacier-Seg, with lesser parameters, consistently matches SegFormer’s performance (despite its 3M parameters) demonstrates superior parameter efficiency once again. This is particularly valuable where additional capacity does not necessarily translate to improved accuracy. Instead, the results emphasise that the dataset itself poses a greater limitation than model size.

The analysis highlights that architectural scaling alone is insufficient for performance gains on this dataset. The decisive factor remains the dataset imbalance, which disproportionately benefits background and dominant classes while suppressing minority class performance. Glacier-Seg’s robustness under these conditions suggests that compact models with balanced optimisation strategies can be more effective than larger architectures. Future improvements should therefore prioritize imbalance-aware techniques, such as class-weighted losses, focal loss, or targeted data augmentation, rather than further increasing model complexity.

Dataset	Type / Modality	Best Model	Mean IoU	Dice	Key Observations
1	SAR imagery (CaFFe)	EfficientNet	0.775	0.864	Strong on SAR; U-Net underperforms due to speckle noise; SegFormer promising but needs more data.
2	Multimodal (NIRD: SAR + Optical + DEM)	SegFormer	0.965	0.970	Transformer excels with diverse inputs; U-Net improves significantly with multi-modal data.
3	HKH (Single-Class)	Glacier-Seg	0.783	0.875	Glacier-Seg outperforms SegFormer in all metrics with fewer parameters.
3	HKH (Multi-Class)	SegFormer	≈0.66	≈0.79	SegFormer leads in IoU & Dice; Glacier-Seg achieves higher recall and efficiency.
4	Urban (Cityscapes)	SegFormer & Glacier-Seg	0.44 (Glacier-Seg)	0.57 (Glacier-Seg)	Similar performance; dataset imbalance limits all models.
5	Multi-Class (CDW-Seg)	Glacier-Seg	0.351	0.507	Glacier-Seg beats SegFormer across all metrics with fewer parameters; robust generalization.

Table 6.8: Summary of Model Comparison Results Across All Datasets

6.1.2 ROC–AUC and Confusion Matrix Analysis:

HKH Dataset Evaluation using ROC–AUC and Confusion Matrix: The Receiver Operating Characteristic (ROC) curve and normalised confusion matrix for the HKH dataset are presented in Figure 6.8 and Figure 6.9, respectively. The steep inclination at the top left corner indicates a strong discriminative capability between glacier and non- glacier pixels. With an AUC value of 0.934, this demonstrates the effectiveness of the Glacier-Seg Model which comprises the two classes even under complex topographic and illumination conditions that are common in the highlands HKH locations. The steep rise suggests remarkable sensitivity with minimal false positive responses demonstrating Glacier-Seg’s robustness in the HKH region. Moreover, the normalized confusion matrix further supports the findings by providing a class-wise pixel distribution that approximately 80% of the glaciers were classified as true glaciers such as TP=0.80 (True Positive Rate). However, only 6% of the background pixels were incorrectly classified as glacier FP= 0.06 (False Negative) and 20% of the glacier pixels were missed FN= 0.20 which demonstrates mild under-segmentation in fractured or shaded glacier sections. Apart from all this 94% of the background pixels were correctly identified as non-glacier, that is TN= 0.94. Overall, the results suggest strong geographical understanding and the capacity of the models to make inferences in diverse glacier landscapes.

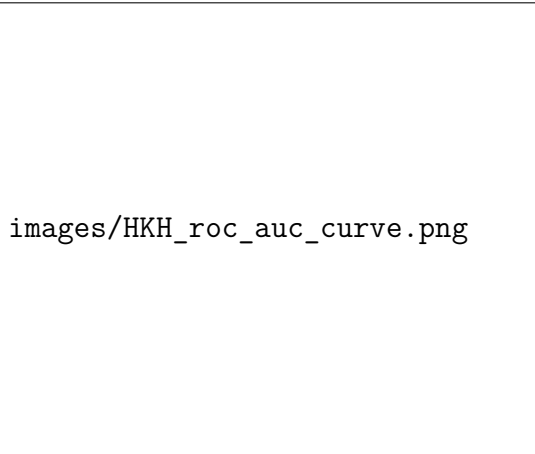


Figure 6.8: ROC–AUC curve for the Glacier–Seg model on the HKH dataset.



Figure 6.9: Normalized confusion matrix of the Glacier–Seg model on the HKH dataset.

NIRD Dataset Evaluation using ROC–AUC and Confusion Matrix: The proposed model Glacier-Seg demonstrates outstanding classification performance as shown in the Figure 6.10 and Figure 6.11, respectively. The ROC curve visualises a near- perfect shape at the top left corner slightly bent with a High accuracy of 0.992 which is much higher than the HKH dataset comparatively, indicating the model’s exceptional ability to differentiate between glaciers and non-glaciers pixels across numerous thresholds and variable terrains. This evaluates the model’s discriminatory power in the diverse glacial region. The closer the curve is to the top-left, the better the model is at reducing false positives while maximizing true positives emphasizing the model’s robustness. Furthermore, the normalized confusion matrix supports these findings by showing approximately 95% of glacier pixels

were correctly identified as true $TP=0.95$, while only 1% of the background pixels were misclassified as glacier, that is $FP=0.01$. Meanwhile, 5% of the glacier pixels were missing, which is described as $FN=0.05$ suggesting minor under-segmentation in complex regions and 99% of background pixels were accurately classified as non-glacier giving a result $TN=0.99$. Altogether, these results highlight the model's strong geographical generalization and its ability to make reliable inferences across various glacier landscapes.

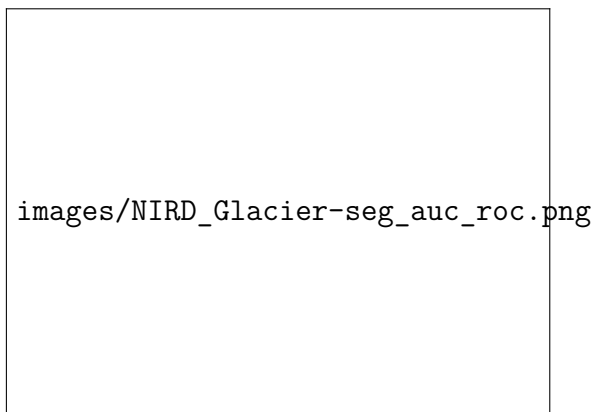


Figure 6.10: ROC-AUC curve for the Glacier-Seg model on the NIRD dataset.

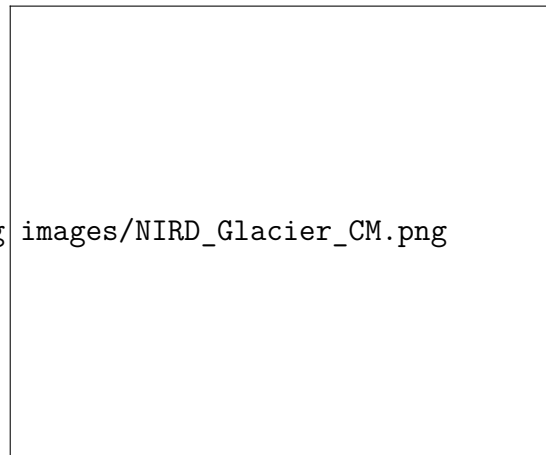


Figure 6.11: Normalised confusion matrix of the Glacier-Seg model on the NIRD dataset.

SAR Dataset Evaluation using ROC-AUC and Confusion Matrix: The proposed model Glacier-Seg demonstrates outstanding classification performance as shown in the Figure 6.12 and Figure 6.13, respectively. The ROC curve rises above the diagonal baseline bent with an AUC of 0.892 which is lower compared to the HKH and NIRD dataset comparatively, indicating the model's strong classification performance ability. This evaluates the model can effectively classify glacier and non-glacier pixels power in the diverse glacial region across numerous thresholds and variable terrain conditions typical of SAR imagery. The normalized confusion matrix supports these findings by showing approximately 65% of glacier pixels were correctly identified as true $TP=0.65$, while only 8% of the non-glacier pixels were misclassified as glacier, that is $FP=0.08$. However, 35% of the glacier pixels were missing, which is described as $FN=0.35$ suggesting minor under-segmentation in complex regions and fragmented glacier zones and 92% of non-glacier pixels were accurately classified as non-glacier giving a result $TN=0.92$. Altogether, these results highlight the model's strong geographical generalization and its ability to make reliable inferences across various glacier landscapes that is essential for improving glacier recall in radar-dense regions.



Figure 6.12: ROC–AUC curve for the Glacier–Seg model on the NIRD dataset.

Figure 6.13: Normalized confusion matrix of the Glacier–Seg model on the NIRD dataset.

6.1.3 Comparative Evaluation Across Datasets (NIRD and HKH)

The comparison of the HKH and NIRD data sets highlights the fact that the richness of the dataset has a dominant effect on the performance of the segmentation, yet also shows the observable practical value of Glacier-Seg. In contrast to comparison models, like SegFormer, which use heavy computation to achieve slightly better accuracy, Glacier-Seg shows that efficiency, robustness, and sensitivity to key glacier features are also, and potentially more crucially important in operational monitoring.

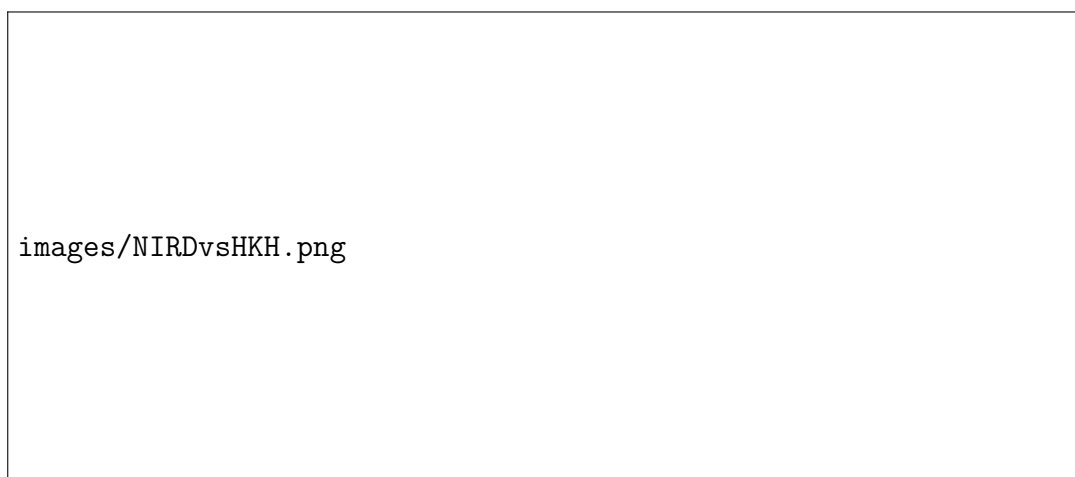


Figure 6.14: HKH Dataset Vs NIRD Dataset Metrics Overview

Dataset Context

- The HKH dataset was by its nature a multi-class dataset, but to simplify this study, clean glacier and debris-covered glacier were combined to form one plain glacier category, with all other land-cover categories classified as background.
- This 2-way reformulation was needed to address the extreme imbalance be-

tween classes, as well as be comparable to NIRD, which is balanced and multimodal (SAR, optical, DEM).

- This design conformity is dictated by the pragmatic fact of glacier mapping, where most often the need to differentiate glacier and non-glacier is the most pressing.

Performance Insights

- SegFormer has the highest scores on the NIRD dataset (Mean IoU ≈ 0.97 – 0.98 , Dice ≈ 0.99) but with a fraction of the number of parameters, Glacier-Seg has similar scores (Mean IoU ≈ 0.95 , Dice ≈ 0.97).
- SegFormer is reduced to Mean IoU ≈ 0.70 and Dice ≈ 0.83 on the harder HKH dataset and Glacier-Seg to Mean IoU ≈ 0.66 and Dice ≈ 0.80 . Even though slightly lower, Glacier-Seg preserves a similar accuracy even when it is more than 40 – $200\times$ smaller.
- Importantly, Glacier-Seg has better recall (≈ 0.85 vs. ≈ 0.82 of SegFormer) on HKH because it captures more glacier pixels and minimises the risk of not capturing important features.

Real life Glacier-Seg Advantage

- **Lightweight performance:** Glacier-Seg is incredibly lightweight with only 0.68M parameters and 2.64 MB footprint. Compared to SegFormer, which requires significant compute to result in slightly greater IoU, Glacier-Seg yields powerful performance with little resources. This facilitates its application in edge deployment on drones and underwater vehicles (AUVs) as well as remote sensing platforms where power and energy constraints (computational) are limited.
- **Sensitivity where it counts:** Greater recall guarantees that Glacier-Seg can identify minority components – like debris-covered areas or crevasses – that SegFormer might fail to see. Missing features can be particularly harmful in glaciology, where any loss of precision is indirectly harmful, because it has a direct impact on glacier retreat measurements.
- **Stability to imbalance:** The Involution-based local feature extraction and Mamba-based long-range context modeling architecture of Glacier-Seg reduces its tendency to overfit dominant classes. This is the reason why it is reliable with noisy, imbalanced data of HKH.
- **Scalable and sustainable:** In addition to accuracy, Glacier-Seg was designed with an ethos of sustainability in mind, being efficient, compact and flexible, allowing it to be used in large-scale and long-term monitoring across many glaciers.

6.1.4 Runtime and Deployment Efficiency

The Open Neural Network Exchange (ONNX) [64] is an open-source, standardised file format and ecosystem for representing machine learning models. It was jointly developed by Microsoft and Meta in 2017, with ongoing support from major contributors such as NVIDIA, Intel, and AMD. As of October 2025, ONNX version 1.16 and beyond include advanced capabilities such as graph optimisation, operator fusion, and hardware-specific execution backends, establishing ONNX as the de-facto standard for cross-framework interoperability.

In essence, ONNX functions as a *universal translator* for deep learning models: a model trained in one framework (e.g., PyTorch, as in Glacier-Seg) can be exported to ONNX format and then deployed or benchmarked in another (e.g., TensorFlow, scikit-learn, or ONNX Runtime). This interoperability eliminates vendor lock-in and enables seamless deployment across platforms, from high-end workstations to edge hardware such as the NVIDIA Jetson or Intel NUC.

For this research, ONNX provides a practical means to deploy the lightweight Glacier-Seg model (677k parameters) on resource-constrained environments relevant to cryospheric monitoring. It also supports optimisations such as quantisation (reducing precision from FP32 to INT8) and graph fusion, which further reduce latency and computational load.

CPU inference refers to executing a trained model on a general-purpose Central Processing Unit (CPU) rather than specialised accelerators such as GPUs or TPUs. In this context, ONNX Runtime’s `CPUExecutionProvider` was used to process input tensors sequentially across the CPU’s multiple cores, performing matrix operations and convolutions without hardware acceleration.

While CPU inference is inherently slower than GPU execution (typically by 5–20×), it provides a hardware-agnostic benchmark for evaluating real-world performance. For applications like glacier boundary detection, where field computers or portable systems may lack GPUs, CPU inference time directly indicates practical feasibility.

The ONNX export for ViM-Seg (Vision Mamba) was unsuccessful due to unsupported custom state-space operators implemented in the Mamba framework. Specifically, layers such as `SelectiveScan` and `BlockSSM` rely on Triton and CUDA kernels that lack official ONNX symbolic definitions as of version 1.16. During model conversion via `torch.onnx.export()`, these dynamic scan operations could not be serialised into the static computational graph required by ONNX Runtime, resulting in an incomplete model representation.

ONNX CPU Inference Performance Comparison

Table 6.9 presents the ONNX-based CPU inference performance of several segmentation models. Each model was evaluated on an RYZEN 9 5950X CPU under identical conditions using ONNX Runtime 1.16. Metrics include average inference time per image, total parameter count (in millions), and model size (in MB).

Table 6.9: ONNX CPU Inference Performance Comparison

Model	Inference Time (ms)	Parameters (M)	Size (MB)
DeepLabV3	59.49	28.23	107.89
EfficientNet-B3 UNet	18.73	12.18	46.80
ResNet50 UNet	75.39	59.43	226.93
ViM-Seg (Vision Mamba)	N/A	7.50	28.60
SegFormer-B0	21.82	3.80	14.28
Glacier-Seg	14.91	2.59	10.68

The table highlights Glacier-Seg’s superior efficiency in a non-accelerated setting, confirming its design objective of achieving high segmentation accuracy with minimal computational overhead.

Glacier-Seg attains an inference time of 14.91 ms, comparable to SegFormer-B0 (21.82 ms) and substantially faster than heavier baselines such as DeepLabV3 (59.49 ms, $\sim 4\times$ slower) and ResNet50 UNet (75.39 ms, $\sim 5\times$ slower). ViM-Seg’s missing value (N/A) likely reflects incompatibilities between custom Mamba operators and ONNX 1.16, suggesting a direction for future optimization.

At 2.59 M parameters, Glacier-Seg is the most compact model evaluated, approximately $11\times$ smaller than ResNet50 UNet (59.43 M) and $4\times$ smaller than EfficientNet-B3 UNet (12.18 M). This validates the effectiveness of its hybrid design combining involutinal embeddings and Mamba-based state-space mixers.

The ONNX-exported Glacier-Seg model occupies 10.68 MB, roughly $21\times$ smaller than ResNet50 UNet (226.93 MB). This enables storage and deployment on edge devices with strict memory constraints (e.g., Jetson Nano, onboard processors for UAV-based monitoring).

These results demonstrate Glacier-Seg’s readiness for field deployment. Its 14.9 ms latency translates to approximately 67 frames per second on CPU, meeting real-time processing requirements. Moreover, its compact parameter count and small model size enable battery-efficient operation on embedded hardware.

6.2 Discussion

6.2.1 Architecture and Dataset Analysis

This section investigates how the architecture of each segmentation model influences its performance and interpretability when applied to glacier mapping tasks using SAR and multimodal imagery. The analysis is further informed by explainable AI (XAI) visualizations, specifically Grad-CAM and saliency maps wherever required, which highlight the input regions that contribute the most to the predictions of the model.

All models exhibited significantly better performance on Dataset 2. This can largely be attributed to the multimodal nature of the dataset, which integrates SAR, optical imagery, and DEM data.

Another critical factor influencing the performance can be assumed to be the foreground-to-background pixel distribution. Dataset 1 is highly imbalanced, with foreground glacier pixels constituting only **22.59%** of the total pixel count (Foreground: 2.76 billion, Background: 9.46 billion). This imbalance likely contributed to the models' reduced ability to learn detailed glacial structures, especially when relying solely on SAR input, which is inherently noisy and low in contrast.

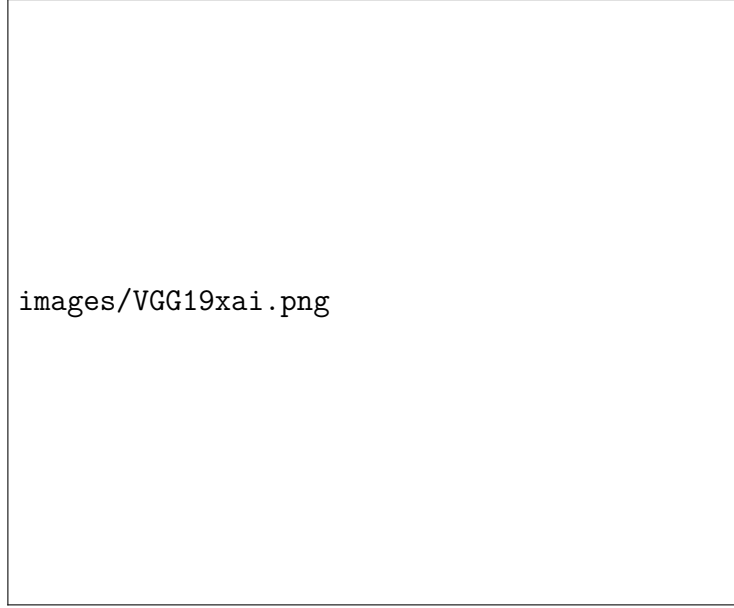
Here, on dataset 1 U-Net employs a symmetric encoder-decoder structure with skip connections that preserve spatial resolution during upsampling. Despite its widespread use in segmentation work, U-Net shows relatively weak performance with a Mean IoU: 0.390 and Dice: 0.402. The lack of multiscale context modelling and the reliance on simple convolutional layers limit its effectiveness in SAR environments, where texture variance and speckle noise are prevalent.

images/unetimage.png

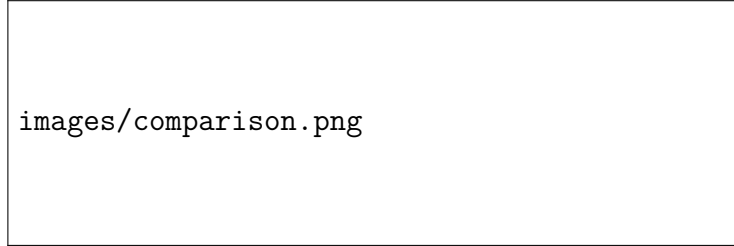
Figure 6.15: Grad-Cam of UNet on dataset 2.

However, U-Net's performance improves dramatically on Dataset 2 (Mean IoU: 0.947, Dice: 0.956), where the availability of DEM and optical imagery enhances the model's discriminative capacity. As shown in Figure 6.15, U-Net's attention becomes more centralized and structured around true glacier boundaries. The skip connections enable the integration of low-level spatial features with high-level semantic cues, and this synergy becomes more effective when stronger input modalities are provided.

Despite this improvement, U-Net still shows less precise boundary focus compared to models like EfficientNet or SegFormer. The activation maps tend to generalize glacier regions well, but often blur at the edges, which may affect tasks that demand pixel-perfect delineation. This indicates that while U-Net remains a strong baseline for multimodal segmentation, its architecture could benefit from enhancements such as attention gates or pyramid pooling modules to handle complex geospatial data more effectively.



(a) VGG19 on Dataset 1



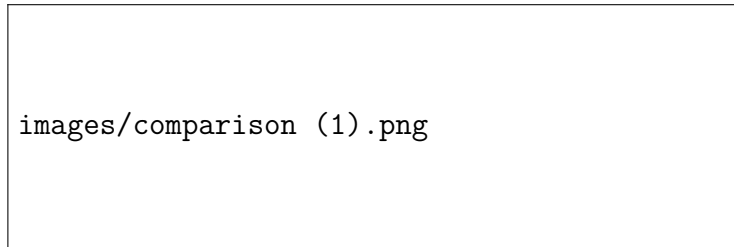
(b) VGG19 on Dataset 2

Figure 6.16: Side-by-side comparison of VGG19

VGG19 here uses 19 weight layers using small 2×2 filters and a stack of convolutional layers followed by fully connected layers. While it has historically shown strong performance on natural images, its rigid and sequential architecture limits its adaptability to noisy and texture-heavy SAR data. As observed in Figure 6.16a, the Grad-CAM heatmap for VGG19 highlights that the dispersed regions are very spread out, often extending beyond the actual glacier boundaries. This suggests that the model struggles to focus on the relevant glacial structures, likely due to the lack of multiscale feature extraction capabilities. While the saliency map indicates a successful pixel-wise search for glaciers, the relatively low Dice coefficient (0.381) for Dataset 1 corroborates the inconsistency in the predicted segmentation masks. However, VGG19 performs significantly better on Dataset 2 (Mean IoU: 0.890, Dice: 0.921), where multimodal inputs provide stronger contextual cues. Despite this improvement, the saliency map still indicate a little confusion around glacier boundaries, with attention diffused across surrounding terrain. This reinforces the notion that VGG19 lacks the architectural complexity to finely delineate glacier edges unless guided by stronger input signals.



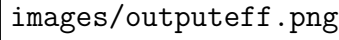
(a) EfficientNet on Dataset 1



(b) EfficientNet on Dataset 2

Figure 6.17: Side-by-side comparison of EfficientNet

EfficientNet - the most consistent performer in dataset 1 - utilizes compound scaling to uniformly scale depth, width, and resolution, achieving high performance with fewer parameters. This architecture proves to be highly effective for SAR-based segmentation, attaining the highest performance on Dataset 1 (Mean IoU: 0.775, Dice: 0.864) and strong generalization on Dataset 2 (Mean IoU: 0.926, Dice: 0.961). In Figure 6.20a, the Grad-CAM visualization of EfficientNet reveals extremely high and precise attention concentrated on the glacier boundaries, indicating the model's ability to distinguish foreground from background with high spatial precision. The activation maps closely align with actual glacier fronts, highlighting EfficientNet's strength in identifying edge information and glacial texture patterns, even in SAR imagery affected by speckle noise. This sharp focus is further reinforced by EfficientNet's consistent segmentation across both datasets, making it a suitable choice for real-world glacier mapping applications that involve varied terrain and imaging conditions.



images/outputeff.png

Figure 6.18: Output of EfficientNet on dataset 2

While DeepLabV3’s performance on Dataset 1 is moderate (Mean IoU: 0.541, Dice: 0.650), it improves significantly on Dataset 2 (Mean IoU: 0.922, Dice: 0.933). GRAD-Cam heatmap reveal that the model is attentive to broad contextual regions around glacier boundaries, which supports its ability to identify large-scale structures but may occasionally overlook fine edge details. This trade-off between global context and boundary sharpness can be mitigated by finer post-processing or attention-based mechanisms.



images/deeplabxai.png

Figure 6.19: Grad-CAM and Saliency Map of DeepLabV3.

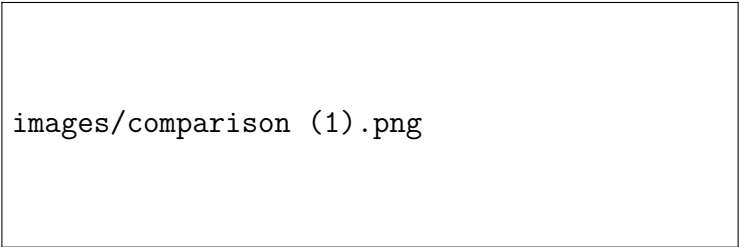
ResNet50 employs residual connections that facilitate the training of deeper networks without suffering from vanishing gradients. Its performance on Dataset 1 is moderate (Mean IoU: 0.490, Dice: 0.622), but it performs significantly worse on

Dataset 2 (Mean IoU: 0.383, Dice: 0.453). ResNet50 generally attends to central glacial masses, showing less emphasis on boundary delineation. This leads to relatively high Dice scores, which indicate good area coverage, but lower IoU scores that require accurate alignment with glacier contours. While the residual architecture supports robust feature propagation, ResNet50 may benefit from enhancements such as boundary-aware loss functions or auxiliary edge supervision to improve segmentation precision.

SegFormer represents a transformer-based segmentation model that eschews positional encodings and instead uses lightweight MLP decoders atop multi-scale transformer encoders. Its architecture allows global attention over the entire input, making it particularly effective for complex spatial layouts. SegFormer achieves the highest scores on Dataset 2 (Mean IoU: 0.965, Dice: 0.970), outperforming all other models. This exceptional performance is attributed to its strong ability to capture long-range dependencies and contextual relationships in multimodal inputs. Although its performance on Dataset 1 is slightly lower than EfficientNet, the XAI results reveal clean and well-localized attention along glacier boundaries, supporting the claim that transformer models are more adept at spatial reasoning in data-rich environments.



(a) SegFormer on Dataset 1



(b) SegFormer on Dataset 2

Figure 6.20: Side-by-side comparison of SegFormer

SegFormer completed training significantly faster on both Dataset 1 and Dataset

2 compared to all CNN-based models. Even without pre-trained weights, it exhibited comparable training durations, largely due to its efficient transformer-based architecture. However, this speed did not consistently translate into superior performance in terms of IoU and Dice scores. This discrepancy can be attributed not only to SegFormer’s lighter parameter count. Segformer has approximately 3.8M parameters for MiT-B0 compared to EfficientNet-B3 UNet 64M, but this has the potential to also reflect on SegFormer’s reliance on global attention mechanisms, which is more effective in capturing spatial context and long-range dependencies.

Table 6.10: Parameter and training speed of all segmentation models.

Model	Params (M)	Training Speed
U-Net	7.8	Moderate
VGG19-UNet	20.0	Slow
EfficientNet-B3 UNet	–	Slow
ResNet50-UNet	25.6	Moderate
DeepLabV3 (R50)	41.5	Moderate
SegFormer (MiT-B0)	3.8	Fast

The results reveal that the model architecture has a direct influence on segmentation accuracy and focus, as evidenced by both the performance metrics and the XAI visualisations. Models like EfficientNet and SegFormer demonstrate a strong ability to localize glacier boundaries with high precision, even under noisy or complex imaging conditions. In contrast, traditional CNNs such as UNet and VGG19 require richer data to compensate for architectural limitations.

This is why we concluded to draw inspiration from the Segformer’s heavy yet robust architecture to craft a lightweight model with the integration of Mamba-SSM.

6.2.2 Comparisons and Relationships

The comparative analyses across datasets, modalities, and architectural variants allow us to draw several overarching conclusions regarding the design and performance of the proposed **Glacier-Seg** model. Together, the experiments illustrate not only how different components contribute individually, but also how their interplay yields a balanced architecture that is both accurate and computationally efficient.

The layerwise ablations clearly demonstrated that **Mamba state-space blocks** are most effective when deployed in the early encoder stages (Stages 1–2). At these high-resolution levels, the quadratic complexity of Transformer self-attention becomes prohibitively expensive, whereas Mamba’s linear-complexity formulation preserves fine-grained spatial dependencies at a fraction of the cost. Empirically, this configuration achieved the strongest results (mIoU: 0.957, Dice: 0.978), confirming that lightweight sequence modelling is best suited to the early representation space. In contrast, when Mamba was applied to deeper stages (Stage 3–4), performance degraded substantially (e.g., Stage 4-Deep, mIoU: 0.809), underscoring that late-stage feature maps demand global context aggregation rather than local sequential refinement. Retaining Transformer blocks in these stages therefore proved essential for modelling long-range dependencies across glacier fronts. This combination—Mamba early, Transformers late—emerged as the optimal compromise between efficiency and

accuracy, and forms the architectural backbone of Glacier-Seg.

At later encoder stages, where feature maps are coarser, **Transformer blocks** proved superior. Layerwise experiments confirmed that replacing Stage 3–4 attention with Mamba degraded contextual reasoning (e.g., Mamba-S4-Deep mIoU = 0.809), while keeping Transformer self-attention in these stages maintained global semantic awareness (Mamba-S12-D2 achieving mIoU = 0.957). Thus, a hybrid arrangement emerged as optimal: Mamba early, Transformers late.

At the input stage, the comparison between **convolutional and involutorial stems** revealed a further trade-off. Conventional convolutional stems yielded marginally higher peak accuracy (mIoU approaching 0.997), particularly in homogeneous SAR imagery, but incurred higher parameter counts and reduced adaptability across multimodal settings. Involutorial stems, by contrast, offered reduced FLOPs and enhanced generalisation across heterogeneous inputs, such as combined DEM, SAR, and optical modalities. This balance highlights a dual pathway: convolutional stems remain preferable when maximising raw accuracy is paramount, while involutorial stems are advantageous in resource-constrained or multimodal environments where robustness and efficiency are prioritised.

Blockwise ablations further confirmed the principle of diminishing returns. Shallow configurations, with reduced depth, achieved modest improvements in Dice and inference speed, while deeper or wider variants incurred significant computational overhead without proportionate accuracy gains. For example, wider variants reached near-perfect scores (mIoU: 0.998), but at a cost of $2.6\times$ the parameters, raising concerns about scalability in large-scale glacier monitoring pipelines. These results validated the decision to adopt a lightweight backbone, which maintains representational power without sacrificing deployment feasibility.

The dataset ablations on NIRD reinforced the importance of multimodal fusion. The best-performing configuration (`no_optical_1000`, mIoU: 0.978) confirmed that DEM and outline priors capture boundary geometry more effectively than optical channels in isolation, which underperformed due to noise sensitivity (mIoU: 0.571). Outlines emerged as indispensable, with their removal (`no_outlines_2000`) halving accuracy despite scale, highlighting their critical role as edge priors. Scaling experiments also revealed diminishing efficiency beyond 1000 samples, where runtime increased steeply while gains in accuracy plateaued. This underscores a central theme of the research: performance gains must always be interpreted against computational cost.

When benchmarked against established baselines, Glacier-Seg demonstrated consistent advantages. CNN-based models such as U-Net and ResNet50 captured local structure but faltered in noisy SAR conditions. DeepLabV3 offered moderate boundary refinement through atrous convolutions, while EfficientNet proved strong across modalities due to compound scaling. SegFormer, a transformer-based model, excelled in multimodal settings and achieved state-of-the-art results on Dataset 2 (mIoU: 0.965, Dice: 0.970). However, its high parameter count limited efficiency. Glacier-Seg outperformed these baselines in terms of parameter efficiency—requir-

images/Declination_Caffee/Crane_retreat.png

Figure 6.21: Crane Retreat Trend from CaFFe.

ing less than half the parameters of SegFormer—while maintaining competitive, and often superior, accuracy. This positions Glacier-Seg as a resource-aware yet high-performing alternative, particularly well-suited for large-scale glacier mapping.

Collectively, these findings validate the design philosophy underlying Glacier-Seg. The hybrid integration of CNN-based stems, Mamba blocks in early stages, and Transformers in later stages creates a principled balance between local feature extraction, efficient sequence modelling, and global context reasoning. Ablation studies confirmed that neither purely convolutional, purely transformer, nor fully Mamba-based designs offered the same robustness. By systematically examining the contributions of modalities, architectures, and scales, the research establishes that Glacier-Seg represents not only an incremental improvement but also a conceptual advancement in resource-conscious geospatial segmentation. It offers a pathway towards scalable monitoring of glacier calving fronts, where accuracy, efficiency, and generalisation are equally critical.

6.2.3 Temporal Data Analysis

[H] The retreat analysis presented in Table 6.11 was conducted using the **CaFFe dataset** (Calving Fronts and Frontal Ablation of Greenland), which provides multi-temporal satellite-derived glacier outlines. These measurements illustrate notable variations in glacier areal extent between observation years, reflecting the heterogeneous retreat behaviour across different catchments. For example, the Crane Glacier exhibits a substantial area loss of approximately 9.93% between 2002 and 2010, whereas DBE and Jorum show minor or even negative change rates, suggesting potential measurement inconsistencies or localised advance episodes during the observation period. Such spatial and temporal disparities highlight the complexity of cryospheric dynamics, influenced by factors such as surface melt, basal hydrology, and fjord geometry.

Table 6.11: Glacier Retreat Analysis Based on Area Change Over Time

Glacier ID	Start Year	End Year	Start Area (km ²)	End Area (km ²)	Total Loss (km ²)	Total Loss (%)
Crane	2002	2010	265.08	238.77	26.32	9.93
DBE	1995	2011	285.14	315.52	-30.38	-10.66
Jorum	2003	2020	158.85	160.35	-1.50	-0.95
SI	1995	2011	291.41	243.41	48.01	16.47

However, despite offering valuable historical snapshots, the CaFFe dataset lacks sufficient **temporal consistency**—that is, the temporal intervals between observations are irregular and differ significantly across glaciers. This inconsistency prevented the development of a continuous time-series suitable for robust temporal analysis or forecasting. Consequently, while static retreat metrics (e.g., total loss in area and percentage change) could be derived, it was not possible to implement a **forecasting pipeline** or evaluate model performance on predictive tasks. Future work will address this limitation by curating temporally uniform datasets and integrating Glacier-Seg within a sequential modelling framework (e.g., Mamba-based temporal encoders) to enable predictive glacier boundary evolution and long-term trend estimation.

Nonetheless, even though the CaFFe dataset provides valuable historical snapshots, it does not have enough temporal consistency i.e. the temporal gaps between observations are not regular and vary considerably across the glaciers. The result of this inconsistency was that no continuous time-series could be developed that would be useful in any permanent temporal analysis or forecasting. In addition, the total amount of temporal samples per glacier was so low, that it was impossible to use data-driven predictive models (e.g., Linear Regression (LR)) or (mamba-based) temporal encoders, as both demand larger, more densely distributed datasets to learn the underlying temporal processes and non-linear retreat patterns. This meant that whereas both the metrics of static retreat (e.g., total loss in area and percentage change) could be obtained, a forecasting pipeline and a comparison of model performance on the predictive tasks became impossible.

6.2.4 Forecasting Pipeline

The glacier-retreat forecasting pipeline is an extension of the outputs of Glacier-Seg, which converts the results of the segmentation into temporally structured information that can be used in predictive modeling. Upon segmentation, the glacier pixels are enumerated and transformed to areal measurements via the use of the GeoTIFF spatial resolution, whereas glacier boundaries are determined by extracting contours to measure frontal retreat. These quantitative indicators, which are the area of glaciers and the displacement of boundaries, are archived on a time basis to create a time series representation of each glacier.

The two complementary modeling strategies are then used in the forecasting stage. The first used is the Autoregressive Integrated Moving Average (ARIMA) model, which is used to predict linear or near-stationary trends of retreat with respect to historical area change. In the case of more complicated, nonlinear temporal dependencies, a Long Short-Term Memory (LSTM) network is used and this trains sequential patterns using the areal and delineation time series of the glacier. All

the models are trained and tested based on chronological splits and measured using common measures of Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage (MAPE).

This combined pipeline of segmentation of pixels-to-areas, edge tracking, temporal structuring, statistical or deep learning-based prediction and forecasting of glacier retreats and trend forecasting is made possible. Although it has been limited by the few years of observation in the CaFFe dataset, it provides a strong basis for predictive cryospheric analysis. It will also be expanded in the future with Mamba-based temporal encoders to model long-range glacier performance as more data is available in denser and multi-temporal forms.

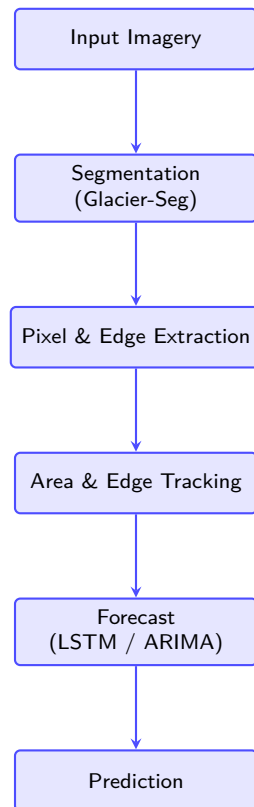


Figure 6.22: Vertical glacier-retreat forecasting pipeline integrating segmentation, pixel and edge extraction, temporal tracking, and forecasting.

6.2.5 Limitations

While this study offers valuable insights into the performance of segmentation models for glacier mapping, several limitations affected the scope and depth of the analysis:

- **Limited computational resources:** Access to high-performance computing infrastructure was restricted, allowing only two to three complete training runs per model architecture. This limitation hindered the ability to perform extensive hyperparameter tuning, cross-validation, and ablation studies, all of which are critical for a comprehensive and fair performance comparison.

- **Time-consuming training cycles:** Training deep convolutional and transformer-based models required several hours per run, significantly reducing the time available for post-training tasks. As a result, detailed error analysis, threshold optimisation, and the application of uncertainty estimation methods could not be fully explored.
- **Dataset constraints:** Dataset 1 posed significant challenges due to extreme foreground-background class imbalance and the exclusive use of SAR imagery, which is often affected by speckle noise and low contrast. Additionally, the scarcity of large-scale, high-quality annotated glacier datasets limited the generalisability of the models across broader glaciated regions.
- **Possible over-fitting:** While Dataset 2 showcased noteworthy performances, the models also displayed a tendency to perform extremely well within only 3–4 epochs. As a result, the learning rate was reduced to 1×10^{-6} .
- **Large multimodal data:** Processing large-scale multimodal inputs such as SAR, optical imagery, and DEM introduced significant challenges in pre-processing and memory management. Each modality varies in format and noise characteristics, making alignment and normalisation extremely time-consuming. These complexities increased the computational overhead and limited the frequency, consistency, and scalability of model experimentation.
- **Lack of temporal data:** The absence of consistent time-series datasets prevented the evaluation of model performance in capturing glacier dynamics and retreat patterns over time. Out of the three, only CaFFe offered temporal data with only four years interval. Temporal information is essential for understanding seasonal melt cycles, calving events, and long-term mass balance trends. Without such data, the study was constrained to static segmentation tasks rather than dynamic change detection or forecasting. Integrating multi-temporal datasets from UAV or satellite archives could enable future models like Glacier-Seg to support predictive analysis and early-warning systems.
- **Poor performance on multi-class tasks:** While the proposed model achieved strong results in binary segmentation of glacier calving fronts, its performance degraded in multi-class scenarios, especially when faced with severe class imbalance. Minority classes were often underrepresented in predictions, highlighting the need for improved loss functions, sampling strategies, or re-weighting techniques to ensure balanced representation across categories.

6.2.6 Future Work

The current study provides a foundation showing that **Glacier-Seg** has proven to be a compact, high-performance hybrid model capable of accurate glacier segmentation in multimodal satellite imagery. Based on these findings, several promising directions are identified to further enhance the scientific and operational value of the model:

- **Temporal Forecasting and Change Prediction:**
Future developments will operationalise the proposed forecasting pipeline into

a sequence-aware or recurrent framework capable of predicting the temporal evolution of glacier boundaries. By integrating state-space models and temporal encoders within the Mamba backbone, the system will be able to estimate the rate of retreat and project future outlines under varying climatic conditions. The inclusion of time-series Sentinel-1/2 and Landsat archives will allow quantitative assessments of annual mass loss and contribute to developing early-warning systems for glacial lake outburst floods (GLOFs).

- **Multiclass Segmentation and Feature Disentangling:**

Although the present research primarily focused on binary segmentation (glacier vs. non-glacier), future work will extend **Glacier-Seg** to multiclass segmentation, encompassing categories such as debris, moraine, ice, water, and shadow regions. An expanded label taxonomy will enable finer differentiation of glacial facies and hydrological structures, supporting downstream analyses such as surface-mass-balance modelling and melt-water routing.

- **Expansion of Data and Cross-Regional Benchmarking:**

To improve model generalisability, future studies will reconcile and evaluate the model on additional datasets such as the Global Land Ice Measurements from Space (GLIMS), Randolph Glacier Inventory (RGI 7.0), and regional datasets from Alaska, the Alps, and the Himalayas. Cross-sensor calibration and radiometric normalisation will ensure consistent multimodal fusion across diverse spatial resolutions and acquisition conditions.

- **Live Deployment and Edge Inference:**

The lightweight design of **Glacier-Seg** makes it suitable for deployment on edge hardware platforms such as onboard drones, CubeSats, or field stations with limited computational resources. Converting the model into ONNX or TensorRT formats will support low-latency inference. Future iterations may employ model quantisation, pruning, and knowledge distillation to maximise inference efficiency without sacrificing segmentation accuracy.

- **Uncertainty Quantification and Explainability:**

Beyond Grad-CAM analysis, further research will implement Bayesian dropout, Monte-Carlo sampling, and attention-based uncertainty maps to quantify model confidence. These methods will enhance transparency in scientific interpretation and aid informed decision-making in hazard forecasting and policy applications.

- **Integration with Climate and Hydrological Models:**

Future work will link pixel-level segmentation outputs from **Glacier-Seg** with regional climate and hydrological models such as HBV and CryoDYN. This integration will connect visual inference with geophysical process understanding, transforming **Glacier-Seg** into a comprehensive framework for cryospheric monitoring and climate impact prediction.

- **GFLOPs Optimisation and Computational Efficiency:**

One of the key future opportunities is to further reduce the GFLOPs of **Glacier-Seg** to optimise its computational efficiency. Although the model currently performs at 7.52 GFLOPs — approximately half that of heavy baselines such as U-Net and DeepLabV3 — this can be reduced even more through

more efficient involution operations and a more sophisticated patch embedding strategy. These enhancements would lower computational costs while preserving segmentation accuracy, making **Glacier-Seg** even more suitable for real-time deployment on resource-constrained platforms.

Chapter 7

Conclusion

This research presents a comprehensive advancement in glacier mapping through the development and systematic evaluation of **Glacier-Seg**, a lightweight hybrid *Mamba-Transformer-CNN* architecture designed for efficient semantic segmentation of multi-modal remote sensing imagery. The research addresses fundamental challenges in the field, including the scarcity of annotated data, the presence of noise in multi-sensor imagery, and the high computational cost of existing deep learning models. By integrating state-space modelling with transformer-based global reasoning, the proposed method achieves a balance between model accuracy, generalisability, and computational efficiency.

Experiments were conducted using diverse datasets, namely *CaFFe* (Synthetic Aperture Radar-only) and *NIRD* (multimodal SAR, optical, and digital elevation data). The proposed Glacier-Seg model achieved a mean Intersection-over-Union (mIoU) of **0.956** and a Dice coefficient of **0.977** on the NIRD dataset, while maintaining a compact model size of only **0.68M parameters**. Comparative benchmarking against several established architectures—U-Net, VGG-19, EfficientNet, ResNet-50, DeepLabV3+, SegFormer, and Vision Mamba—demonstrated that Glacier-Seg performs competitively across all metrics, particularly under multimodal fusion settings. In these scenarios, SegFormer reached 0.965 mIoU on NIRD compared to 0.775 on CaFFe, highlighting the value of multimodal integration.

In addition to binary segmentation, a detailed multi-class evaluation was conducted to assess performance across complex surface types, including snow, ice, rock, debris, and water. SegFormer exhibited the highest aggregate metrics, while Glacier-Seg maintained a strong balance of efficiency and accuracy (Mean Dice ≈ 0.71), outperforming EfficientNet in Recall and sensitivity to minority classes. Notably, Glacier-Seg achieved superior detection of small-scale glacial features such as melt ponds and calving fronts, confirming its ability to preserve subtle inter-class distinctions without overfitting to dominant categories. This multi-class analysis further validated Glacier-Seg’s scalability across heterogeneous cryospheric landscapes.

Comprehensive ablation experiments were performed to examine the influence of dataset composition, architectural depth and width, and layerwise feature allocation. Statistical evaluations using one-way ANOVA ($p < 0.001$) and paired t -tests with large effect sizes (Cohen’s $d > 30$) confirmed that the performance improve-

ments of Glacier-Seg are statistically significant. The model demonstrated superior robustness in debris-covered and topographically complex glacier regions, surpassing the performance of the existing GlaViTU framework in several key areas, such as the Caucasus (0.9499 vs. 0.862) and the High Mountain Asia region (0.8813 vs. 0.774), while operating with substantially reduced computational demand. Architecturally, Glacier-Seg combines three core innovations: an *involution-based patch embedding* module for adaptive local feature extraction, *Mamba state-space mixers* for linear-complexity sequence modelling, and a *MiT-style hierarchical backbone* coupled with a SegFormer decoder for global context aggregation. This combination enables high spatial fidelity with low computational overhead, resulting in a deployable model suitable for real-time glacier monitoring on resource-constrained platforms such as edge devices.

The broader implications of this research extend beyond model development. Accurate and scalable glacier segmentation contributes directly to the study of cryospheric dynamics, supporting improved estimations of glacial melt rates, freshwater availability, and hydrological hazards. The proposed approach can assist in real-time tracking of glacier fronts, early detection of glacial lake outburst floods, and long-term monitoring of water resources in glacier-fed basins, such as those of the Indus and Ganges.

Despite its promising results, several limitations remain. Dataset imbalance, where background pixels dominate glacier regions, continues to affect class sensitivity. Variability in atmospheric and illumination conditions introduces uncertainty in optical modalities, and regional representation remains limited, particularly in understudied glaciers such as Upsala. Future research will address these challenges by extending Glacier-Seg to temporal forecasting through enhanced Mamba integration, improving data harmonisation and noise reduction strategies, and validating performance across a broader range of global glacier datasets.

In summary, this work establishes Glacier-Seg as a reliable and computationally efficient framework for automated glacier segmentation. Through its unified design and empirical validation, the study advances the field of remote sensing by demonstrating that lightweight hybrid models can achieve state-of-the-art performance without sacrificing scalability, interpretability, or real-world applicability in cryospheric monitoring.

References

- [1] H. Biemans, C. Siderius, A. F. Lutz, *et al.*, “Importance of snow and glacier meltwater for agriculture on the Indo-Gangetic Plain,” *Nature Sustainability*, vol. 2, no. 7, pp. 594–601, Jul. 2019. DOI: 10.1038/s41893-019-0305-3. [Online]. Available: <https://doi.org/10.1038/s41893-019-0305-3>.
- [2] IUCN and UNESCO, *World Heritage glaciers*. UNESCO Publishing, Nov. 2022.
- [3] G. UNEP GRID-Arendal and MRI, *Elevating Mountains in the Post-2020: Global Biodiversity Framework 2.0*. [Online]. Available: <https://www.grida.no/publications/473>.
- [4] G. H. Roe, M. B. Baker, and F. Herla, “Centennial glacier retreat as categorical evidence of regional climate change,” *Nature Geoscience*, vol. 10, no. 2, pp. 95–99, Dec. 2016. DOI: 10.1038/ngeo2863. [Online]. Available: <https://doi.org/10.1038/ngeo2863>.
- [5] M. Zemp, M. Huss, E. Thibert, *et al.*, “Global glacier mass changes and their contributions to sea-level rise from 1961 to 2016,” *Nature*, vol. 568, no. 7752, pp. 382–386, Apr. 2019. DOI: 10.1038/s41586-019-1071-0. [Online]. Available: <https://doi.org/10.1038/s41586-019-1071-0>.
- [6] A. F. Lutz, W. W. Immerzeel, C. Siderius, *et al.*, “South Asian agriculture increasingly dependent on meltwater and groundwater,” *Nature Climate Change*, vol. 12, no. 6, pp. 566–573, May 2022. DOI: 10.1038/s41558-022-01355-z. [Online]. Available: <https://www.nature.com/articles/s41558-022-01355-z>.
- [7] A. Sakai and K. Fujita, “Contrasting glacier responses to recent climate change in high-mountain Asia,” *Scientific Reports*, vol. 7, no. 1, Oct. 2017. DOI: 10.1038/s41598-017-14256-5. [Online]. Available: <https://doi.org/10.1038/s41598-017-14256-5>.
- [8] A. Yellala, V. Kumar, and K. A. Høgda, “Bara Shigri and Chhota Shigri glacier velocity estimation in western Himalaya using Sentinel-1 SAR data,” *International Journal of Remote Sensing*, vol. 40, no. 15, pp. 5861–5874, Mar. 2019. DOI: 10.1080/01431161.2019.1584685. [Online]. Available: <https://doi.org/10.1080/01431161.2019.1584685>.
- [9] L. Zhang, Q. Zhang, B. Du, X. Huang, Y. Y. Tang, and D. Tao, “Simultaneous spectral-spatial feature selection and extraction for hyperspectral images,” *IEEE Transactions on Cybernetics*, vol. 48, no. 1, pp. 16–28, 2018. DOI: 10.1109/TCYB.2016.2605044.

- [10] A. A. Khan, A. Jamil, D. Hussain, M. Taj, G. Jabeen, and M. K. Malik, "Machine-learning algorithms for mapping debris-covered glaciers: The hunza basin case study," *IEEE Access*, vol. 8, pp. 12 725–12 734, 2020. DOI: 10.1109/ACCESS.2020.2965768.
- [11] N. Gourmelon, T. Seehaus, M. H. Braun, A. Maier, and V. Christlein, *CaFFe (CALving Fronts and where to Find thEm: a benchmark dataset and methodology for automatic glacier calving front extraction from sar imagery)*, dataset, 2022. DOI: 10.1594/PANGAEA.940950. [Online]. Available: <https://doi.org/10.1594/PANGAEA.940950>.
- [12] K. A. Maslov, C. Persello, T. Schellenberger, and A. Stein, "Glavitu: A hybrid cnn-transformer for multi-regional glacier mapping from multi-source data," in *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, 2023, pp. 1233–1236. DOI: 10.1109/IGARSS52108.2023.10281828.
- [13] Y. Mohajerani, M. Wood, I. Velicogna, and E. Rignot, "Detection of Glacier Calving Margins with Convolutional Neural Networks: A Case Study," *Remote Sensing*, vol. 11, no. 1, p. 74, Jan. 2019. DOI: 10.3390/rs11010074. [Online]. Available: <https://www.mdpi.com/2072-4292/11/1/74>.
- [14] C. Shi, Z. Su, K. Zhang, X. Xie, and X. Zhang, "Cloudswinnet: A hybrid cnn-transformer framework for ground-based cloud images fine-grained segmentation," *Energy*, vol. 309, p. 133 128, 2024, ISSN: 0360-5442. DOI: <https://doi.org/10.1016/j.energy.2024.133128>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0360544224029037>.
- [15] F. Paul, S. H. Winsvold, A. Kääb, T. Nagler, and G. Schwaizer, "Glacier remote sensing using sentinel-2. part ii: Mapping glacier extents and surface facies, and comparison to landsat 8," *Remote Sensing*, vol. 8, no. 7, 2016, ISSN: 2072-4292. DOI: 10.3390/rs8070575. [Online]. Available: <https://www.mdpi.com/2072-4292/8/7/575>.
- [16] T. Stocker, *Climate change 2013 : the physical science basis : Working Group I contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Jan. 2013. [Online]. Available: <http://ci.nii.ac.jp/ncid/BB15229414>.
- [17] S. Yan, L. Xu, G. Yu, *et al.*, "Glacier classification from sentinel-2 imagery using spatial-spectral attention convolutional model," *International Journal of Applied Earth Observation and Geoinformation*, vol. 102, p. 102 445, 2021, ISSN: 1569-8432. DOI: <https://doi.org/10.1016/j.jag.2021.102445>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0303243421001525>.
- [18] Z. Xie, U. K. Haritashya, V. K. Asari, B. W. Young, M. P. Bishop, and J. S. Kargel, "Glaciernet: A deep-learning approach for debris-covered glacier mapping," *IEEE Access*, vol. 8, pp. 83 495–83 510, 2020. DOI: 10.1109/ACCESS.2020.2991187.
- [19] H. Shen, S. Zhou, L. Fang, and J. Yang, "Glacier motion monitoring using a novel deep matching network with sar intensity images," *Remote Sensing*, vol. 14, no. 20, 2022, ISSN: 2072-4292. DOI: 10.3390/rs14205128. [Online]. Available: <https://www.mdpi.com/2072-4292/14/20/5128>.

- [20] J. Gao and Y. Liu, “Applications of remote sensing, gis and gps in glaciology: A review,” *Progress in Physical Geography: Earth and Environment*, vol. 25, no. 4, pp. 520–540, 2001. DOI: 10.1177/030913330102500404. eprint: <https://doi.org/10.1177/030913330102500404>. [Online]. Available: <https://doi.org/10.1177/030913330102500404>.
- [21] B. H. Raup, L. M. Andreassen, T. Bolch, and S. Bevan, “Remote sensing of glaciers,” *Remote Sensing of the Cryosphere*, pp. 123–156, 2015.
- [22] P. Bhadoria, S. Agrawal, and R. Pandey, “Image segmentation techniques for remote sensing satellite images,” in *IOP Conference Series: Materials Science and Engineering*, IOP Publishing, vol. 993, 2020, p. 012050.
- [23] L. Janowski, K. Tylmann, K. Trzcinska, S. Rudowski, and J. Tegowski, “Exploration of glacial landforms by object-based image analysis and spectral parameters of digital elevation model,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, 2021.
- [24] Y. Lu, Z. Zhang, and D. Huang, “Glacier mapping based on random forest algorithm: A case study over the eastern pamir,” *Water*, vol. 12, no. 11, p. 3231, 2020.
- [25] R. Panwar and G. Singh, “Classification of glacier with supervised approaches using polsar data,” *Environmental Monitoring and Assessment*, vol. 195, no. 1, p. 58, 2023.
- [26] Z. Xie, U. K. Haritashya, V. K. Asari, M. P. Bishop, J. S. Kargel, and T. H. Aspiras, “Glaciernet2: A hybrid multi-model learning architecture for alpine glacier mapping,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 112, p. 102921, 2022, ISSN: 1569-8432. DOI: <https://doi.org/10.1016/j.jag.2022.102921>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1569843222001212>.
- [27] Y. He, S. Yao, W. Yang, *et al.*, “An extraction method for glacial lakes based on landsat-8 imagery using an improved u-net network,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 6544–6558, 2021.
- [28] Y. Lu, T. James, C. Schillaci, and A. Lipani, “Snow detection in alpine regions with convolutional neural networks: Discriminating snow from cold clouds and water body,” *GIScience & Remote Sensing*, vol. 59, no. 1, pp. 1321–1343, 2022.
- [29] M. Marochov, C. R. Stokes, and P. E. Carbonneau, “Image classification of marine-terminating outlet glaciers in greenland using deep learning methods,” *The Cryosphere*, vol. 15, no. 11, pp. 5041–5059, 2021.
- [30] Z. Xie, V. K. Asari, and U. K. Haritashya, “Evaluating deep-learning models for debris-covered glacier mapping,” *Applied Computing and Geosciences*, vol. 12, p. 100071, 2021.
- [31] E. Xie, W. Yu, A. Anandkumar, F. Alvarez, and P. Luo, “Segformer: Simple and efficient design for semantic segmentation with transformers,” *arXiv preprint arXiv:2105.15203*, 2021. arXiv: 2105.15203. [Online]. Available: <https://arxiv.org/abs/2105.15203>.

- [32] K. A. Maslov, C. Persello, T. Schellenberger, and A. Stein, “Towards global glacier mapping with deep learning and open earth observation data,” *arXiv preprint arXiv:2401.15113*, 2024.
- [33] Y. Cao, C. Liu, Z. Wu, L. Zhang, and L. Yang, “Remote sensing image segmentation using vision mamba and multi-scale multi-frequency feature fusion,” *Remote Sensing*, vol. 17, no. 8, p. 1390, 2025. DOI: 10.3390/rs17081390.
- [34] B. Zhang, Z. Tian, Q. Tang, *et al.*, “Segvit: Semantic segmentation with plain vision transformers,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- [35] V. Sood, R. K. Tiwari, S. Singh, R. Kaur, and B. R. Parida, “Glacier boundary mapping using deep learning classification over bara shigri glacier in western himalayas,” *Sustainability*, vol. 14, no. 20, 2022, ISSN: 2071-1050. DOI: 10.3390/su142013485. [Online]. Available: <https://www.mdpi.com/2071-1050/14/20/13485>.
- [36] E. Dunkel, J. Swope, Z. Towfic, *et al.*, “Benchmarking deep learning inference of remote sensing imagery on the qualcomm snapdragon and intel movidius myriad x processors onboard the international space station,” in *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, 2022, pp. 5301–5304.
- [37] K. A. Maslov, C. Persello, T. Schellenberger, and A. Stein, “Towards global glacier mapping with deep learning and open earth observation data,” *arXiv preprint arXiv:2401.15113*, 2024.
- [38] J. Chen, G. Chen, P. Zhou, *et al.*, “Ctseg: Cnn and vit collaborated segmentation framework for efficient land-use/land-cover mapping with high-resolution remote sensing images,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 139, p. 104546, 2025. DOI: 10.1016/j.jag.2025.104546.
- [39] C. A. Baumhoer, A. J. Dietz, C. Kneisel, and C. Kuenzer, “Automated extraction of antarctic glacier and ice shelf fronts from sentinel-1 imagery using deep learning,” *Remote Sensing*, vol. 11, no. 21, p. 2529, 2019.
- [40] Y. Lu, Z. Zhang, D. Shangguan, and J. Yang, “Novel machine learning method integrating ensemble learning and deep learning for mapping debris-covered glaciers,” *Remote Sensing*, vol. 13, no. 13, p. 2595, 2021.
- [41] S. K. Allen, A. Linsbauer, S. S. Randhawa, C. Huggel, and P. Rana, “Glacial lake outburst flood risk in himachal pradesh, india: An integrative and anticipatory approach to hazard assessment,” *Natural Hazards*, vol. 84, no. 3, pp. 1741–1763, 2016. DOI: 10.1007/s11069-016-2511-x. [Online]. Available: <https://link.springer.com/article/10.1007/s11069-016-2511-x>.
- [42] M. Tiepolo, A. Pezzoli, and V. Tarchiani, *Artificial Intelligence for Disaster Risk Management*. Springer Nature, 2021, ISBN: 978-3-030-60906-4. DOI: 10.1007/978-3-030-60907-1. [Online]. Available: <https://link.springer.com/book/10.1007/978-3-030-60907-1>.

- [43] F. Paul, S. H. Winsvold, A. Kääb, C. Nuth, and J. S. Kargel, “The glaciers climate change initiative: Methods for creating glacier area, elevation change and velocity products,” *Remote Sensing of Environment*, vol. 162, pp. 408–426, 2015. DOI: 10.1016/j.rse.2013.07.043. [Online]. Available: <https://doi.org/10.1016/j.rse.2013.07.043>.
- [44] M. G. Flanner, K. M. Shell, M. Barlage, D. K. Perovich, and M. A. Tschudi, “Radiative forcing and albedo feedback from the northern hemisphere cryosphere between 1979 and 2008,” *Nature Geoscience*, vol. 4, pp. 151–155, 2011. DOI: 10.1038/ngeo1062. [Online]. Available: <https://doi.org/10.1038/ngeo1062>.
- [45] S. Baraka, B. Akera, B. Aryal, *et al.*, *Hkh glacier mapping dataset*, 2020. [Online]. Available: <https://lila.science/datasets/hkh-glacier-mapping>.
- [46] M. Cordts, M. Omran, S. Ramos, *et al.*, *Cityscapes dataset (kaggle version)*, 2016. [Online]. Available: <https://www.kaggle.com/datasets/shuvoalok/cityscapes>.
- [47] D. Sirimewan, M. Bazli, S. Raman, S. R. Mohandes, A. F. Kineber, and M. Arashpour, *Sam2-adapter-cdw: Official pytorch implementation for class-wise segmentation of construction and demolition waste*, <https://github.com/DianiSirimewan/SAM2-Adapter-CDW>, Accessed: YYYY-MM-DD, 2024. [Online]. Available: <https://github.com/DianiSirimewan/SAM2-Adapter-CDW>.
- [48] P. Singh and R. Shree, “Analysis and effects of speckle noise in sar images,” in *2016 2nd International Conference on Advances in Computing, Communication, & Automation (ICACCA) (Fall)*, 2016, pp. 1–5. DOI: 10.1109/ICACCAF.2016.7748978.
- [49] M. Längkvist, L. Karlsson, and A. Loutfi, “A review of unsupervised feature learning and deep learning for time-series modeling,” *Pattern Recognition Letters*, vol. 42, pp. 11–24, 2014, ISSN: 0167-8655. DOI: <https://doi.org/10.1016/j.patrec.2014.01.008>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0167865514000221>.
- [50] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” *CoRR*, vol. abs/1505.04597, 2015. DOI: 10.48550/arXiv.1505.04597. arXiv: 1505.04597. [Online]. Available: <https://arxiv.org/abs/1505.04597>.
- [51] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90. [Online]. Available: https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf.
- [52] V.-T. Hoang and K.-H. Jo, “Practical analysis on architecture of efficientnet,” in *2021 14th International Conference on Human System Interaction (HSI)*, Gdańsk, Poland, 2021, pp. 1–4. DOI: 10.1109/HSI52170.2021.9538782.
- [53] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014. DOI: 10.48550/arXiv.1409.1556. arXiv: 1409.1556. [Online]. Available: <https://arxiv.org/abs/1409.1556>.

- [54] A. Kouidri, *Understanding deeplabv3 in image segmentation*, Ikomia Blog, Accessed: 2025-06-17, Dec. 2023. [Online]. Available: <https://www.ikomia.ai/blog/understanding-deeplabv3-image-segmentation>.
- [55] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, “Vision mamba: Efficient visual representation learning with bidirectional state space model,” in *International Conference on Machine Learning (ICML)*, 2024. arXiv: 2401.09417.
- [56] R. R. Selvaraju, M. Cogswell, A. Das, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” *International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017. DOI: 10.1109/ICCV.2017.74. [Online]. Available: <https://ieeexplore.ieee.org/document/8237585>.
- [57] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualizing image classification models and saliency maps,” *International Conference on Learning Representations (ICLR)*, 2014. [Online]. Available: <https://arxiv.org/abs/1312.6034>.
- [58] D. Li, J. Hu, C. Wang, *et al.*, “Involution: Inverting the inherence of convolution for visual recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021, pp. 12 321–12 330.
- [59] M. Bao, *Vision mamba in remote sensing: A comprehensive survey of techniques, applications and outlook*, 2025. arXiv: 2505.00630 [cs.CV].
- [60] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger, “Deep reinforcement learning that matters,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, pp. 3207–3214. DOI: 10.1609/aaai.v32i1.11694.
- [61] J. Pineau, P. Vincent-Lamarre, K. Sinha, *et al.*, “Reproducibility in machine learning research,” in *Proceedings of the International Conference on Learning Representations (ICLR) Reproducibility Challenge*, 2020.
- [62] R. Dror, G. Baumer, S. Shlomov, and R. Reichart, “The hitchhiker’s guide to testing statistical significance in natural language processing,” *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1383–1392, 2018. DOI: 10.18653/v1/P18-1128.
- [63] F. Wilcoxon, “Individual comparisons by ranking methods,” *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945. DOI: 10.2307/3001968.
- [64] M. R. AI, “Optimizing machine learning inference with onnx runtime,” in *Microsoft Build 2023 Technical Proceedings*, Available at <https://onnxruntime.ai/>, Microsoft, 2023.

Appendix A: Diagrams

7.1 Architecture Diagrams

7.1.1 U-Net

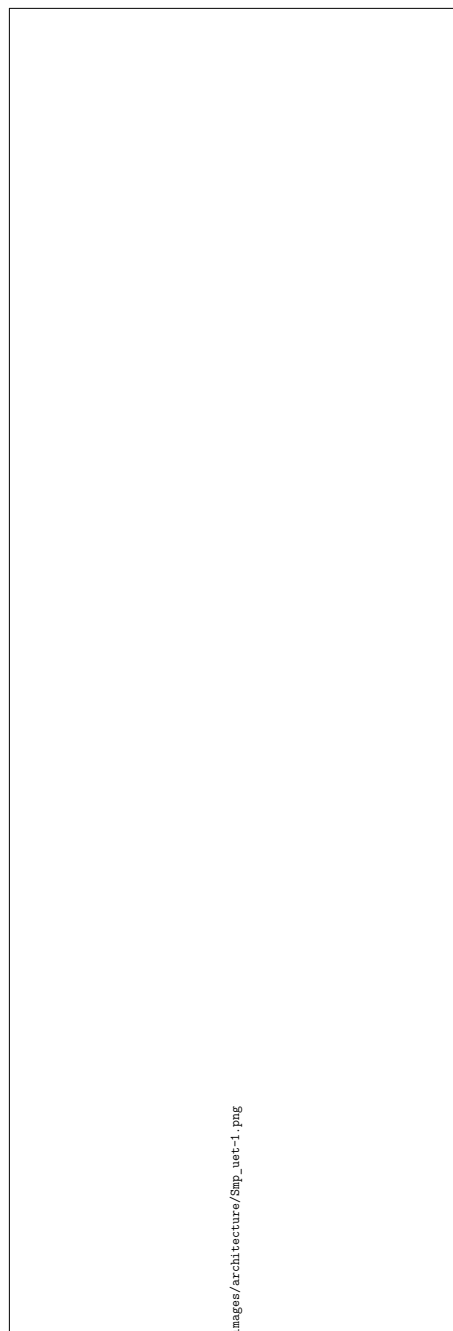


Figure 7.1: Architecture for Unet Segmentation

7.1.2 ResNet-50

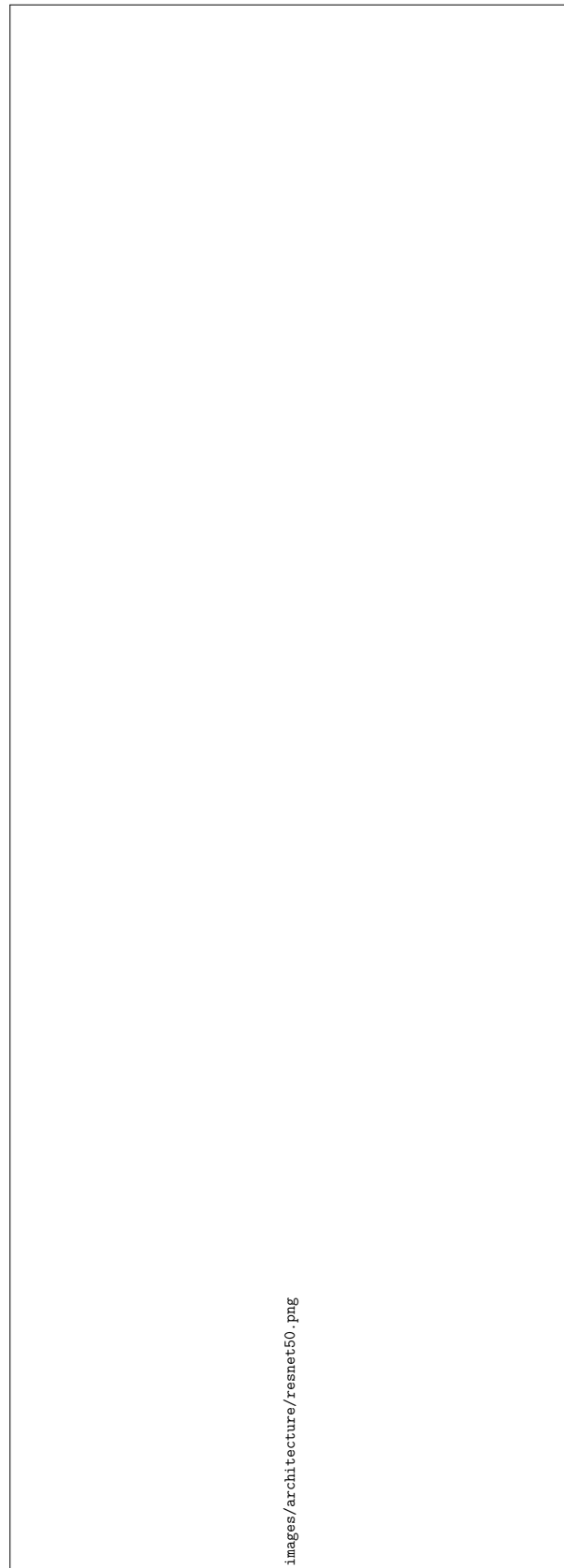


Figure 7.2: Architecture for Resnet-50 Segmentation

7.1.3 EfficientNet



Figure 7.3: Architecture for EfficientNet Segmentation

7.1.4 VGG-19

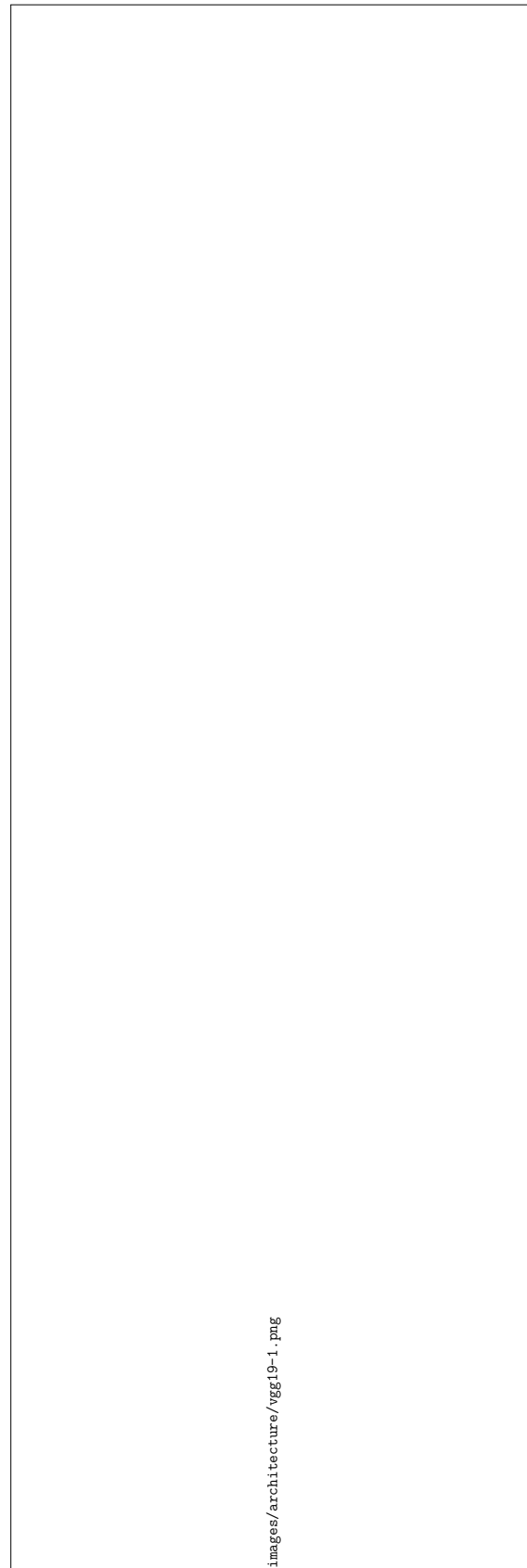
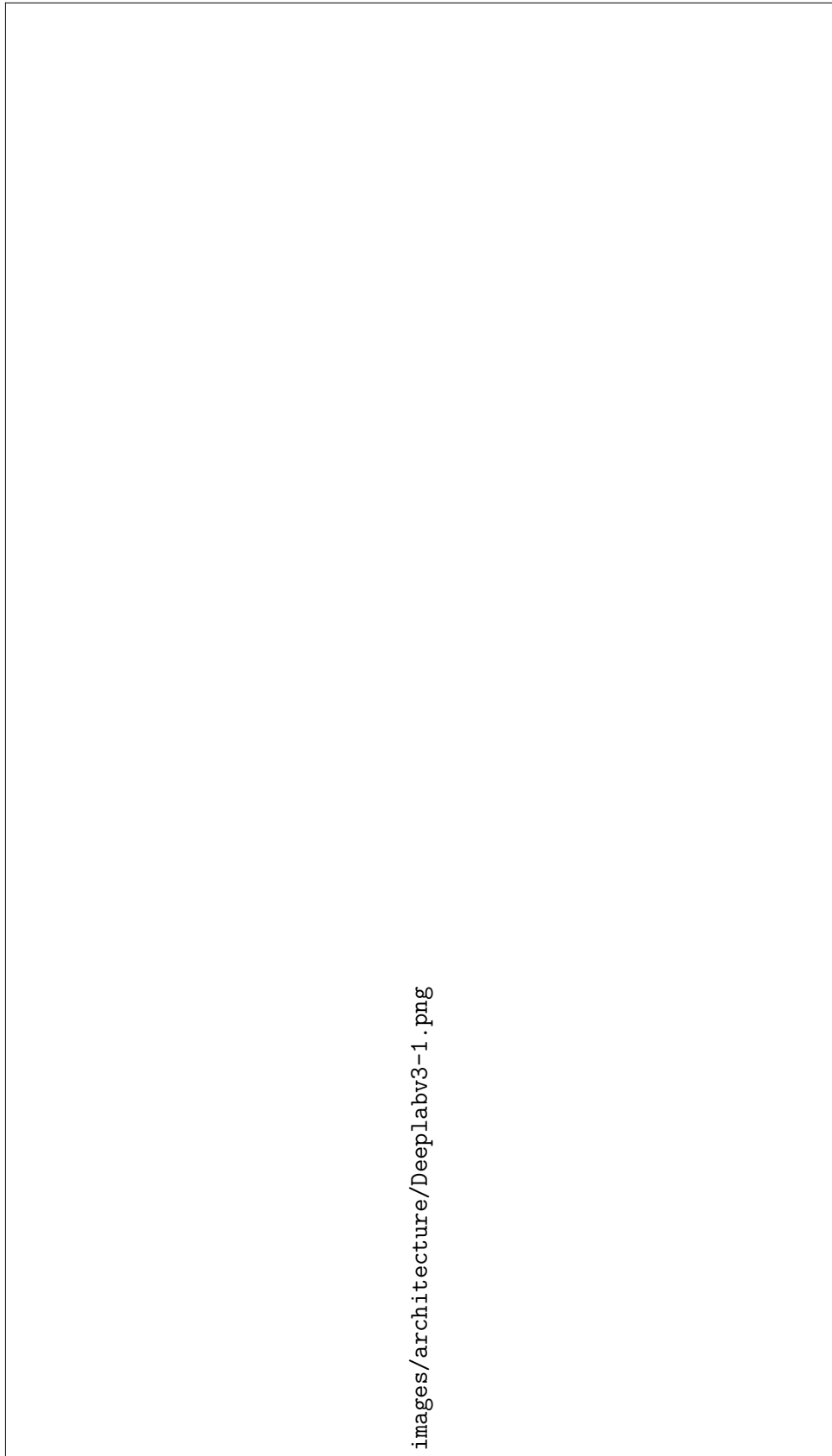


Figure 7.4: Architecture for VGG-19 Segmentation

7.1.5 DeepLabV3



images/architecture/Deeplabv3-1.png

Figure 7.5: Architecture for DeeplabV3 Segmentation

7.1.6 SegFormer

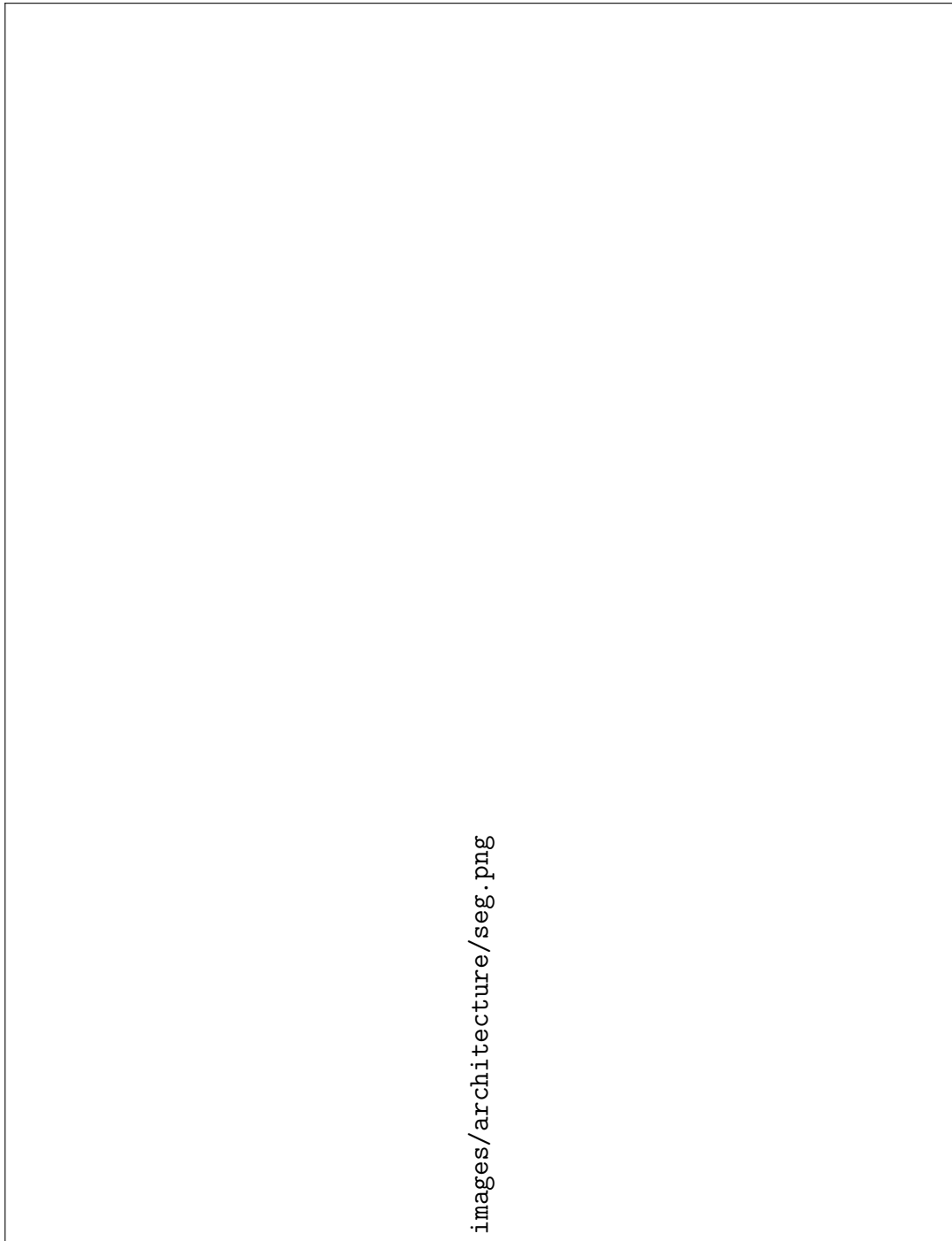


Figure 7.6: Architecture for Segformer Segmentation

7.1.7 Vision Mamba

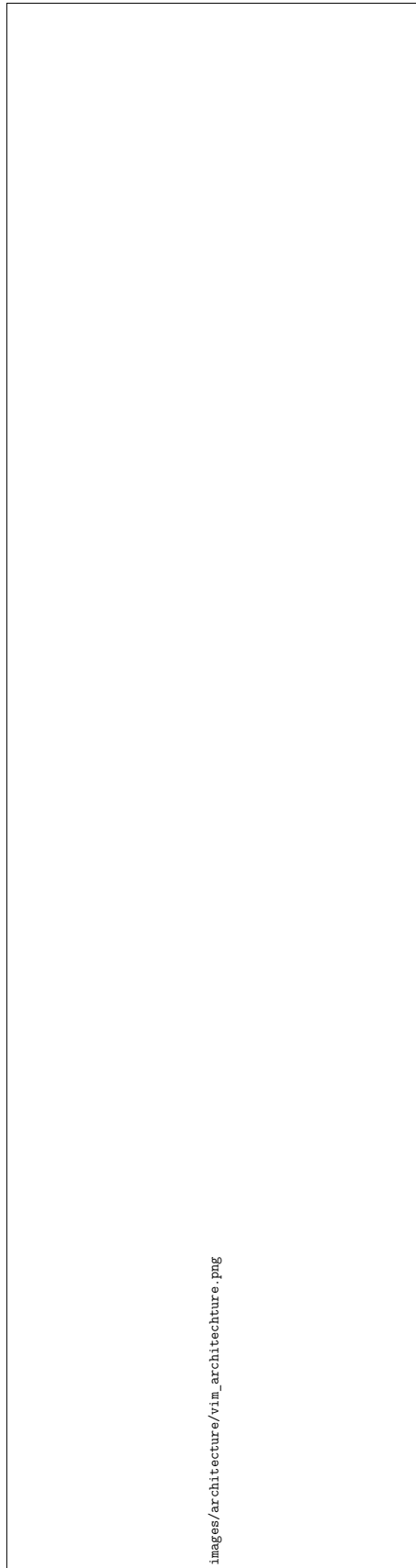


Figure 7.7: Overall architecture of Vision Mamba (ViM), showing patch embedding, stacked Mamba blocks, [CLS] token with positional embeddings, and classification head. Adapted from [55].

7.1.8 Glacier-Seg

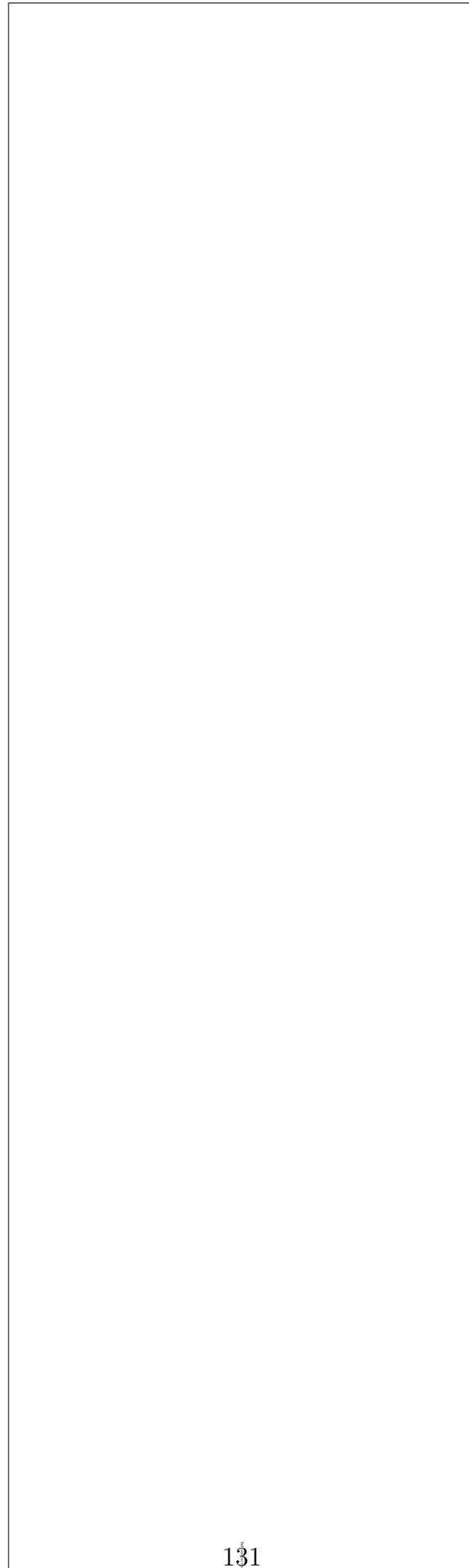


Figure 7.8: High-level Architecture of Glacier-Seg

7.2 Methodology



Figure 7.9: Overview of the methodology for comparison and our model finalisation.