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

Meheruba Hasin Alif, Marzanul Momenine, Syeda Mahmuda, Amitabha Chakrabarty, Sanjida Tasnim, Azwad Aziz

***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. As a computationally efficient and scalable framework for cryospheric monitoring, Glacier-Seg can become capable of real-time glacier boundary tracking and change detection.**

***Index Terms*—Glacier Inventory, Glacier Monitoring, Glacier Mapping, Machine Learning (ML), Cryosphere, Remote Sensing, Classification, Vision Transformers , Artificial Intelligence, Computer Vision, Deep Learning Algorithms.**

I. Introduction

A. 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river basins of the Indus, where the water from melted glaciers contributes to around one-fourth of the annual water withdrawn for irrigation [1].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.

However, rising global temperatures have triggered widespread glacier retreat and thinning across multiple regions [4]. This loss disrupts water availability in regions dependent on seasonal meltwater for drinking, agriculture, and hydropower, threatening billions of livelihoods [5]. Furthermore, glacier degradation destabilizes alpine ecosystems by altering microbial and species diversity, potentially leading to local extinctions [4]. Assessing and monitoring these changes is essential for predicting future environmental conditions in glacier-dependent regions. Glacier mapping enables the identification of long-term trends in ice loss, retreat, and mass balance, providing critical insights for water resource management and early hazard prediction, particularly glacial lake outburst floods in mountain and polar regions.

Patterns of glacial changes are not always apparent due to variability in glacier subsystem responses to climate change [6]. Historically, scientists relied on manual field surveys for accurate glacial boundary mapping, which is extremely challenging due to harsh mountain climates [7]. As a result, digital mapping using remote sensing and Artificial Intelligence (AI) has gained traction. Computer vision methods, especially CNNs, have been employed for glacier analysis, but challenges remain due to clouds, debris, and small datasets focusing mostly on naked ice glaciers, which reduces accuracy for debris-covered glaciers [8].

B. Problem Statement

Glaciers play a critical role in climate regulation, water supply, and ecosystem services, yet automated mapping of their dynamics remains challenging due to sparse, noisy, and temporally limited datasets [9]. Many calving fronts are recorded only a few times per year, and high-quality annotations are scarce, making deep learning models prone to overfitting and limiting their generalisation to unseen regions. Temporal sequences are often unavailable, further restricting models' ability to capture seasonal or inter-annual glacier variations.

Remote sensing imagery adds additional complexity: cloud cover, seasonal snow, shadows, debris, and variable surface conditions distort glacier boundaries [10]. Atmospheric

All the authors are with the Department of Computer Science and Engineering, BRAC University, Dhaka, Bangladesh e-mail: meheruba.hasin.alif@g.bracu.ac.bd; syeda.mahmuda@g.bracu.ac.bd; marzanul.momenine@g.bracu.ac.bd

Manuscript received October 14, 2025; revised October 14, 2025.

variability and ground changes exacerbate these challenges, particularly for mapping complex topographies and calving fronts. While high-resolution images and larger batch sizes can improve accuracy, they demand substantial computational resources, which are often unavailable in practical research settings. As a result, convolutional neural networks (CNNs) frequently struggle to capture fine-scale glacier front details consistently, highlighting the need for hybrid or multi-modal architectures that can integrate spatial, spectral, and contextual cues efficiently [8].

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 multimodal 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** managed to 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.

C. 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:

- **Benchmarking:** Compare state-of-the-art segmentation models such as U-Net, EfficientNet, ResNet50, DeepLabV3, and SegFormer to identify the most effective architectures for glacier mapping under diverse environmental conditions.
- **Data Modality Analysis:** Assess the impact of single-modal and multimodal inputs (SAR, optical, DEM) on segmentation performance and evaluate how feature fusion enhances accuracy and robustness.
- **Hybrid Architecture Design:** Develop lightweight hybrid CNN–Transformer or distilled ViT architectures that maintain spatial fidelity while reducing computational overhead.
- **Cryospheric Forecasting:** Enable glacier retreat tracking and hazard forecasting (e.g., glacial lake outburst floods), supporting water resource management and future extensions to broader satellite datasets.

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.

II. Related Work

There are nineteen glacier regions worldwide [11], four of which are in Asia, with the most prominent located 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).

Remote sensing datasets form the foundation of glacier segmentation research, enabling models to learn glacier morphology, delineate boundaries, and classify ice types in regions where in-situ data collection is limited. Multispectral imagery from satellites such as *Landsat* and *Sentinel*, available via Google Earth Engine (GEE) and the USGS, provides spectral, textural, and topographical information across visible and infrared bands [12]. Despite their richness, spectral similarities between ice, snow, and debris often hinder accurate classification. To overcome atmospheric and cloud interference, Synthetic Aperture Radar (SAR) data are frequently employed for their all-weather imaging capabilities [13], though they remain sensitive to noise and shadowing effects.

Traditional machine learning methods established the groundwork for automated glacier monitoring through image processing, spatial analysis, and statistical modelling. Integration of remote sensing, GIS, and DEMs enabled accurate geolocation and morphological assessment, achieving optical accuracies of 70–90% and SAR accuracies around 75% [14]. Techniques such as thresholding, region growing, and fuzzy logic were later replaced by ensemble classifiers like Random Forests, which fused spectral and topographic features from Landsat-8 and DEMs, achieving accuracies above 97% [15].

With the advent of deep learning, Convolutional Neural Networks (CNNs) revolutionised glacier mapping by learning spatial and spectral hierarchies from multi-sensor inputs. U-Net variants demonstrated precise front extraction from Landsat imagery, outperforming edge-based methods with mean deviations under 100 m [9]. Ensemble and multi-phase frameworks, including DeepLabV3+ ensembles and VGG16-based convolutional sparse coding, further improved segmentation robustness across alpine and marine glaciers [16], [17].

Recent advances in Vision Transformers (ViTs) and hybrid architectures have extended these capabilities by incorporating global attention and efficient representation learning. Models such as SegFormer achieved competitive mIoU scores across standard benchmarks while maintaining compact designs [18]. Hybrid CNN–Transformer frameworks like GlaViTU integrated optical, DEM, and SAR modalities to achieve an average IoU of 0.894 globally [19]. Collectively, these innovations mark a transition from pixel-based classification toward scalable, multi-modal, and transformer-driven architectures for robust and efficient glacier segmentation.

A. Research Gap

Despite advances in deep learning, there remains a lack of lightweight, efficient hybrid architectures specifically op-

timised for glacier segmentation under real-time and edge deployment constraints.

- **Limitations of Existing Models:** Transformer-based U-Nets, though accurate, often incur high FLOPs and parameter counts, making them unsuitable for field monitoring. Additionally, Mamba-based models remain underexplored.
- **Underexplored Potential:** Mamba-based state-space models, known for their linear complexity and long-sequence modelling ability, remain largely untapped in cryospheric research and are yet to be integrated into ViT hybrids for temporal glacier dynamics.
- **Persistent CNN Reliance:** Literature shows a continued dependence on CNN-based frameworks in glacier mapping, while ViTs—though emerging—are not yet optimised for sparse, noisy, or multimodal satellite data in real-time applications.

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.

III. Datasets

To evaluate the performance and generalisation of the proposed glacier segmentation model, five publicly available datasets were used, encompassing both cryospheric and non-cryospheric domains.

a) Dataset 1: Calving Front Definition Dataset (CaFFe): The CaFFe dataset [20], obtained from PANGAEA, provides SAR imagery with annotated calving fronts from seven marine-terminating glaciers across Antarctica, Greenland, and Alaska. Spanning 1995–2020, it includes imagery from multiple radar missions (Sentinel-1, TerraSAR-X, ALOS PALSAR, etc.), offering diverse temporal and spatial resolutions. Each sample contains SAR images, binary front masks, multi-class zone masks, and bounding boxes. A total of 681 tiles were used, split by glacier for training and testing. 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.

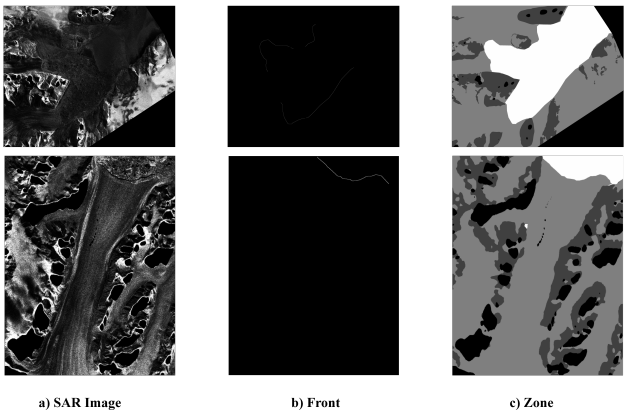

Figure 1: Sample SAR tiles from Sjörgen and Crane glaciers [20]

b) Dataset 2: NIRD Dataset: The NIRD dataset [21] integrates optical (Landsat, Sentinel-2), SAR (ENVISAT, Sentinel-1), and DEM data across six regions: the Alps, HMA, Indonesia, New Zealand, Andes, and Scandinavia. Each 10 m resolution tile includes six modalities (DEM, co- and cross-polarised SAR, optical outlines, InSAR, and bright-dark outlines). The dataset covers ∼9% of global glaciers (19k total) and spans 1988–2020. It was divided into 60% training, 20% validation, and 20% testing sets (1,027 tiles total). As many tiles from the second dataset [21] 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.

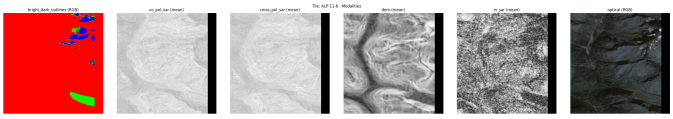

Figure 2: Modalities of the NIRD Dataset.

c) Dataset 3: HKH Glacier Mapping Dataset: The HKH dataset [22] focuses on the Hindu Kush Himalaya ("Third Pole") region across eight countries, providing 14,190 Landsat 7-based image patches (512×512×15) annotated with clean and debris-covered glacier masks. Each patch includes 15 channels—spectral bands, spectral indices (NDVI, NDSI, NDWI), and topography (SRTM elevation/slope). The dataset serves as a high-quality benchmark for training and evaluating glacier delineation models under diverse terrain and climate conditions.

IV. Methodology

This section outlines the complete methodology employed in our study, including augmentation strategies, model selection, training configurations, and evaluation protocols.

A. Data Preprocessing

To increase the diversity and robustness of the training data, this pipeline was designed to simulate real-world imaging

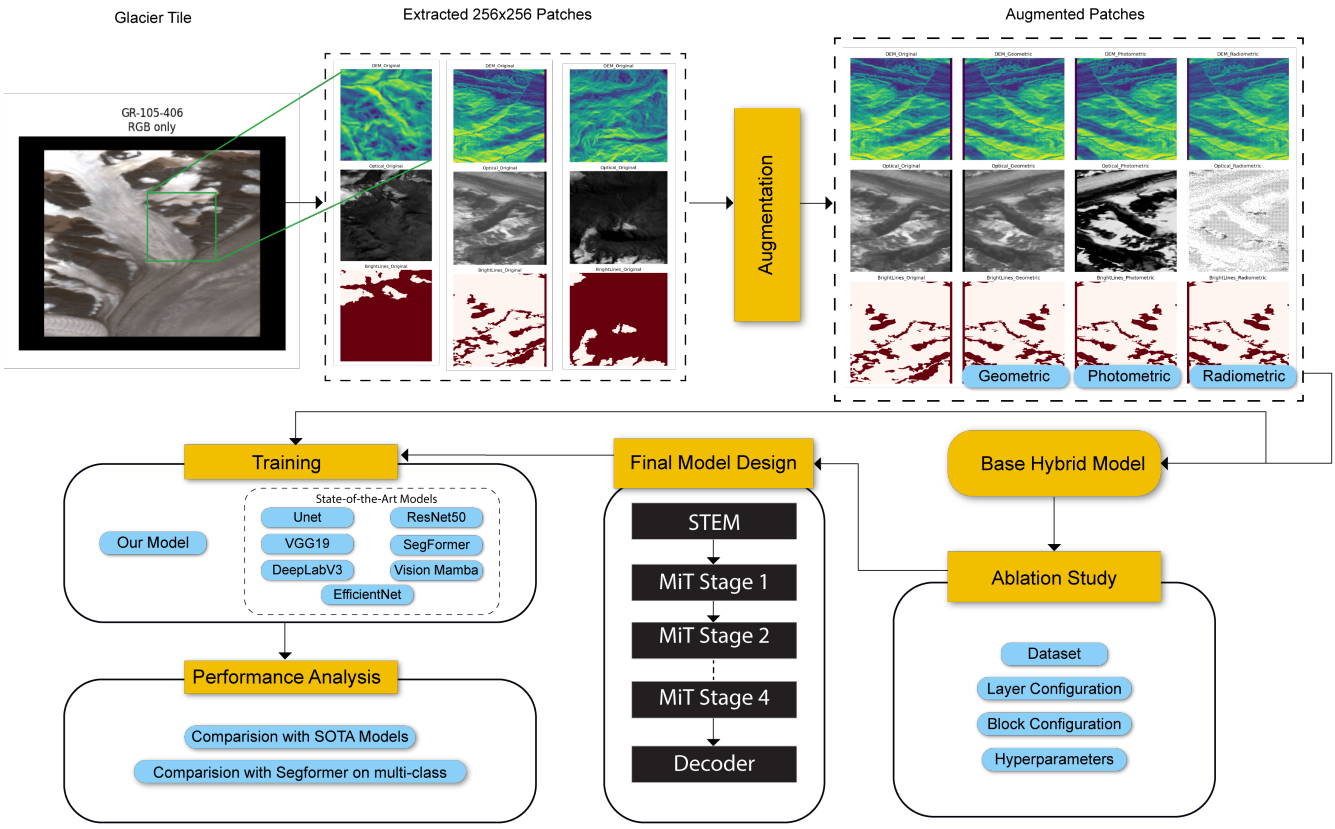

Figure 3: Overview of the methodology for comparison and our model finalisation.

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.

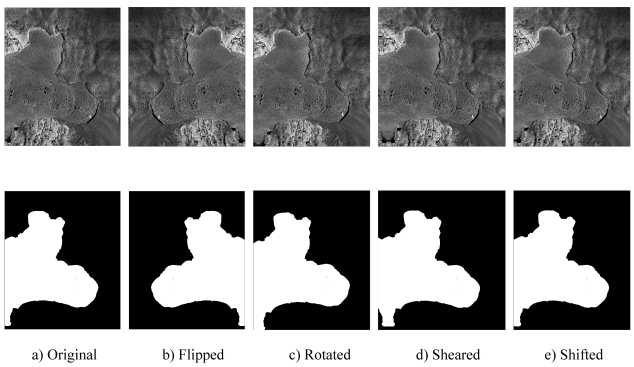

Figure 4: Geometric augmentations performed on the dataset 1.

a) Geometric Augmentation: Spatial diversity was introduced by randomly applying one transformation per image-mask pair, adjusting bounding boxes accordingly:

- **Rotation:** Randomly between -30° and +30°; bilinear interpolation for images, nearest-neighbour for masks.
- **Flipping:** Horizontal, vertical, or combined.
- **Shifting:** Pixel-level translation ±40 pixels along x and y axes.
- **Shearing:** Horizontal shear factor [-0.2, 0.2].

Bounding boxes were recalculated and clipped to remain within image boundaries. Augmented data were saved alongside originals to increase training diversity (Figure 4).

b) Radiometric Augmentation: Applied only to SAR images, leaving masks and bounding boxes unchanged: **Speckle Noise:** Modelled with multiplicative Gamma distribution for 60% of images:

$$I' = I \cdot 10^{\Delta_{dB}/10}, \quad \Delta_{dB} \sim \mathcal{N}(0, \sigma^2) \tag{1}$$

Radiometric Gain Shift: Simulated sensor gain variations:

$$I' = I \cdot G, \quad G \sim \text{Gamma}(L, 1/L) \tag{2}$$

c) Photometric Augmentation: Brightness and contrast adjustments were applied to all images to simulate varying exposure and lighting conditions:

$$I' = \alpha \cdot I + \beta, \quad \alpha \sim U(\alpha_{\min}, \alpha_{\max}), \ \beta \sim U(\beta_{\min}, \beta_{\max}) \tag{3}$$

Where I and I' are original and transformed pixel intensities, α is the contrast factor, and β the brightness offset. This ensures the model generalises across diverse visual environments.

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.

B. Design Specification

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.

1) Stem: Overlap Patch Embedding (Involution2D): The stem module downsamples the input (e.g., multimodal 11-channel imagery) while projecting it into a compact embedding using **Involution2D** [23], which generates spatially varying kernels for each location. This pixel-adaptive mechanism captures local variability—such as clean ice, debris, water, and rock—preserving fine glacier boundaries and texture transitions better than standard convolutions. A subsequent 3×3 convolution with stride 2 and BatchNorm expands the feature map to 16 channels (128×128) while keeping the model lightweight. By tailoring kernels to each pixel's neighbourhood, involution enhances detail capture in noisy, heterogeneous glacier imagery without the computational overhead of full attention mechanisms, striking a balance between expressiveness and efficiency for cryospheric applications.

2) Mamba Block (State-Space Mixer): The Mamba block replaces traditional self-attention in selected encoder stages. It is based on a **state-space model (SSM)** formulation:

$$h_t = Ah_{t-1} + Bx_t, \quad y_t = Ch_t, \tag{4}$$

where A, B, and C are learnable matrices that govern temporal or sequential feature propagation. The block includes RMSNorm (or LayerNorm), a feed-forward MLP (ratio 4.0), and DropPath regularisation. It processes flattened features of shape (B, N, C) to model long-range dependencies in linear time.

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 [24]. Glacier morphology requires global context understanding—capturing long flowlines and valley structures. Mamba achieves this with linear computational complexity ($O(NC)$), offering efficiency advantages over quadratic self-attention ($O(N^2)$). Its recurrent state formulation provides smooth global reasoning, which improves robustness to SAR noise and multimodal inconsistencies, making it well-suited for resource-constrained inference.

3) MixVisionTransformer (MiT) Stages: The encoder is divided into four hierarchical stages, each consisting of patch embedding and one or more Mamba or Attention blocks. The embedding dimensions are $[16, 32, 64, 128]$ with depths $[1, 1, 1, 1]$. Each stage progressively reduces spatial resolution while increasing channel width, enabling both local and global feature abstraction.

Glacier segmentation demands multi-scale feature extraction—fine scales capture edges of narrow termini, while coarse scales represent basin-level context. The pyramid structure balances representation richness and computational cost, ensuring a lightweight design (~ 0.68M parameters) that remains effective across different glacier sizes and terrains.

- **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.

4) Decoder: SegFormerHead (All-Scale Fusion): The decoder adopts the **SegFormerHead** structure, which linearly projects each stage's output to a common dimension (128), upsamples lower-resolution maps, concatenates them, and applies lightweight fusion convolutions before final classification.

This head efficiently fuses global and local features without heavy attention mechanisms, ensuring both speed and accuracy. It enhances boundary recovery by combining fine-scale details with coarse contextual cues—essential for tracing thin or fragmented glacier fronts. The simplicity of the design improves generalisability and reduces overfitting on small datasets.

5) Regularisation and Training Strategy: Additional mechanisms include RMSNorm/LayerNorm for stable gradient flow, DropPath (0.2) for regularisation, and an MLP ratio of 4.0 for sufficient non-linearity. The model is trained using the **BCE + Dice loss** to balance class imbalance and boundary precision, with the AdamW optimiser (lr $= 1 \times 10^{-4}$) and batch size of 30 for 50 epochs.

Remote-sensing glacier datasets often suffer from class imbalance and limited annotations. The combined BCE–Dice loss mitigates the dominance of background pixels while improving overlap accuracy. DropPath and AdamW provide better generalisation and weight stability in small-batch training typical of high-resolution tiles.

6) Overall Design Rationale: The architecture integrates the following principles:

- **Local adaptivity:** Involution-based stem preserves sub-pixel glacier boundaries.
- **Global context:** Mamba block captures long-range spatial continuity efficiently.

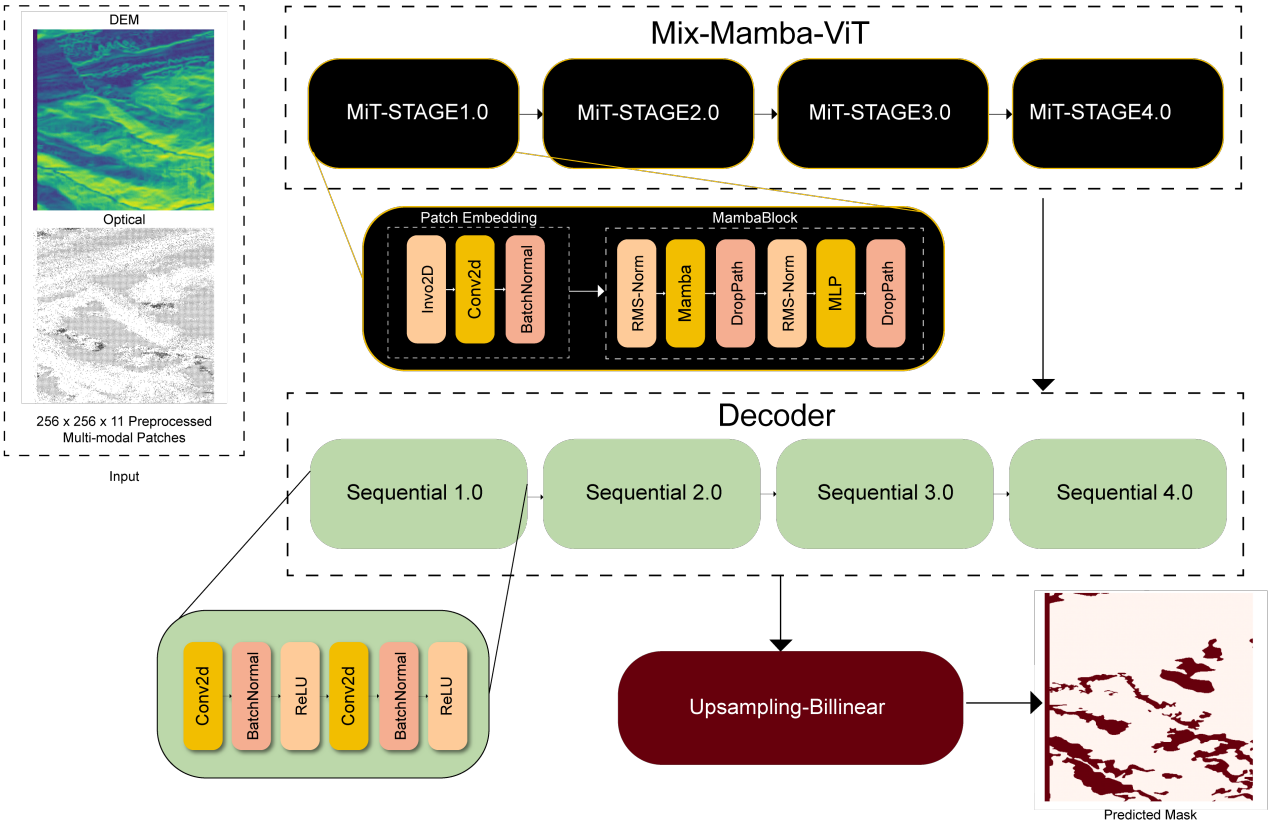

Figure 5: Architecture of Proposed Glacier-Seg

- **Multi-scale hierarchy:** MiT encoder enables both structural and boundary-level understanding.
- **Efficient fusion:** SegFormerHead provides lightweight yet accurate decoding.
- **Regularisation:** DropPath and composite loss improve generalisation under data scarcity.

This modular design ensures that the proposed hybrid network maintains high segmentation accuracy, low computational complexity, and strong generalisability across multi-modal glacier datasets.

C. Baselines

This research benchmarks several state-of-the-art segmentation models. U-Net, a symmetric 23-layer encoder-decoder CNN, preserves spatial information via convolutions, pooling, up-convolutions, and skip connections. ResNet-50, a 50-layer CNN with residual bottleneck blocks, mitigates vanishing gradients while extracting hierarchical features. EfficientNet balances depth, width, and resolution using inverted bottlenecks with depthwise and pointwise convolutions. VGG-19 employs 16 convolutional layers with consistent 3×3 kernels and max-pooling for fine feature extraction. DeepLabV3 leverages atrous convolutions and ASPP for multi-scale context and improved boundary segmentation. SegFormer, a hierarchical transformer, uses self-attention and a lightweight MLP decoder, while Vision Mamba applies hierarchical state-space modeling with overlapping patch embeddings for efficient high-resolution processing.

D. Evaluation Metrics

After training and fine-tuning our segmentation models on glacier datasets, we assess their performance using widely adopted semantic segmentation metrics. The Intersection over Union (IoU) and Dice coefficient quantify spatial overlap between predicted and true glacier regions, offering balanced measures of segmentation accuracy. Pixel Accuracy provides an overall correctness ratio but can be skewed by large non-glacier areas. Precision and Recall capture complementary aspects—precision reflects the reliability of glacier predictions, while recall measures completeness of glacier detection. Finally, the AUC-ROC score evaluates the model's discriminative ability across thresholds, ensuring robustness under class imbalance. These metrics provide insights into how well the models delineate glacier boundaries and segment different classes, which is critical for accurate glacier mapping in satellite imagery.

V. Results and Discussion

A. Dataset Analysis

Training Dataset 1 used 20 epochs with 449–552 images per epoch. U-Net (mIoU 0.390, Dice 0.402) showed limited discriminative capability, while VGG19 improved mIoU (0.531)

but had lower Dice (0.381), which indicates inconsistent segmentation. EfficientNet achieved the highest performance (mIoU 0.775, Dice 0.864) due to compound scaling that balances depth, width, and resolution. ResNet50 (mIoU 0.490, Dice 0.622) and DeepLabV3 (mIoU 0.541, Dice 0.650) benefited from residual connections and atrous convolutions, respectively. SegFormer reached mIoU 0.620, Dice 0.591. The proposed **Glacier-Seg** attained mIoU 0.7109 and Dice 0.8310, outperforming CNN baselines while remaining more compact than transformer architectures. Fig. 7 illustrates its sharper boundary recovery and noise robustness, highlighting the hybrid CNN–Mamba–Transformer synergy that balances spatial precision and efficiency.

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. SegFormer achieved the best results (mIoU 0.965, Dice 0.970) due to global attention, while ResNet50 underperformed (mIoU 0.383, Dice 0.453), likely reflecting gradient degradation in deeper layers. Once again, Glacier-Seg matched SegFormer (mIoU 0.956, Dice 0.977) with fewer parameters, demonstrating efficient feature aggregation and stable convergence. These results highlight that multimodal integration is critical for high-resolution glacier delineation, particularly in heterogeneous terrains. Building upon these findings, SegFormer and EfficientNet were selected as the main baseline for Dataset 3.

Figure 8: Comparison of Model images Performance on Dataset 2

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. On Dataset-3 (HKH region), single-class results in Table I show that Glacier-Seg outperformed SegFormer across all metrics (Dice 0.8749 vs. 0.8555; mIoU 0.7828 vs. 0.7546), validating its higher boundary precision and balanced Precision–Recall profile. For the multi-class case (Table II), SegFormer attained the best aggregate scores

Table I: 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

(mIoU ≈0.66, Dice ≈0.79), yet Glacier-Seg remained competitive (mIoU ≈0.61, Dice ≈0.71) with notably higher Recall (≈0.85), capturing minority classes such as debris or meltwater more reliably. EfficientNet followed closely, confirming the adaptability of compound-scaled CNNs to multimodal inputs.

Overall, Glacier-Seg achieves a strong accuracy–efficiency trade-off: outperforming heavier models in single-class tasks and maintaining comparable multi-class performance with 2–5× fewer parameters. These results emphasise its suitability for operational, real-time glacier mapping, offering robust segmentation under limited data and compute constraints.

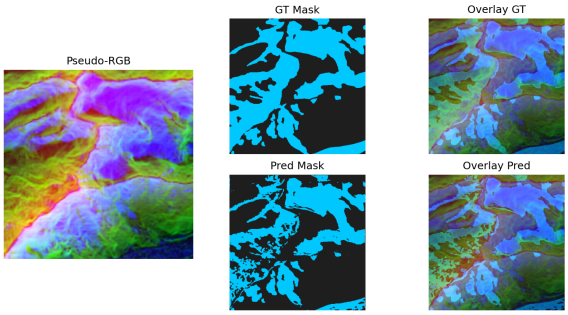

Figure 9: Predicted Mask by Glacier-Seg on HKH Dataset

Table II: 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

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),

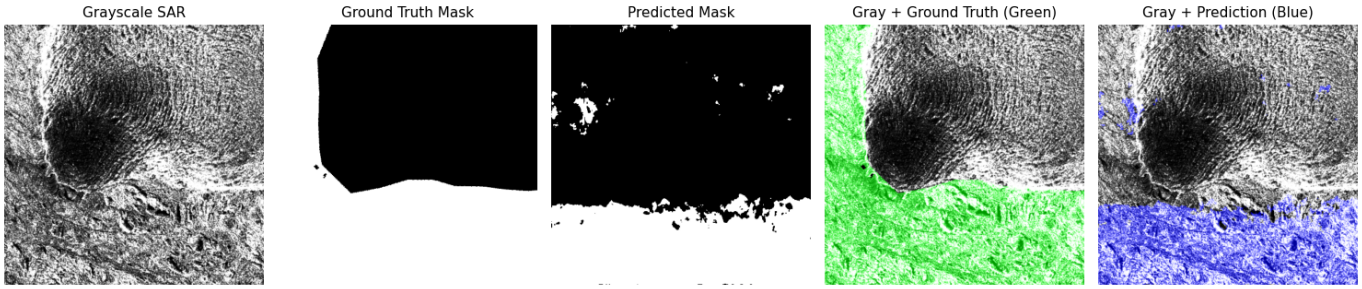

Figure 6: Predicted Mask by Glacier-Seg on CaFFe Dataset

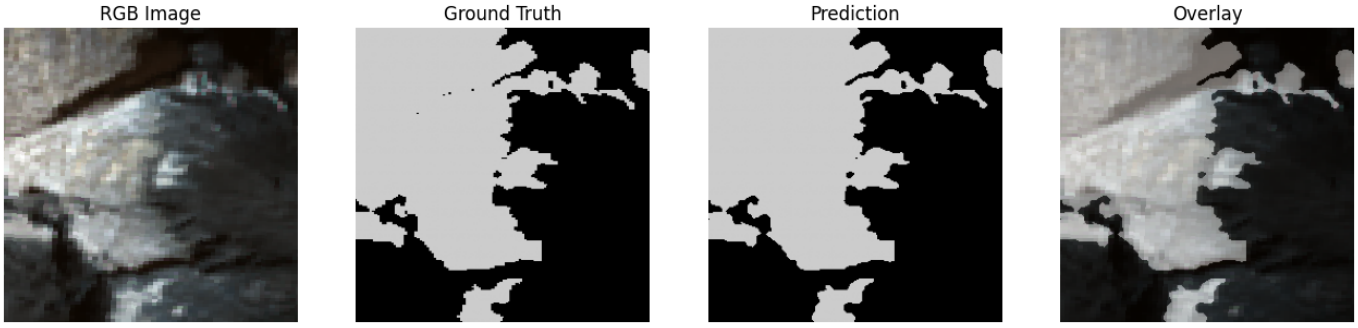

Figure 7: Predicted Mask by Glacier-Seg on NIRD Dataset

and Precision ($\approx$ 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 $\approx$ 0.58, Dice $\approx$ 0.69), 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 $\approx$ 0.61, Dice $\approx$ 0.71), with a comparable Final Loss ($\approx$ 0.30) to SegFormer ($\approx$ 0.28), evidencing stable convergence despite its compact design. However, Glacier-Seg excels in Recall ($\approx$ 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.

B. ONNX Deployment and CPU Inference

The Open Neural Network Exchange (ONNX) [25] is an open-source, standardised format for representing and deploying machine-learning models across frameworks. Developed jointly by Microsoft and Meta in 2017, with continuing support from NVIDIA, Intel, and AMD, ONNX has evolved into the de-facto standard for model interoperability. For this research, ONNX was used to deploy the lightweight Glacier-Seg model (2.59 M parameters, $\sim$10.7 MB) on resource-constrained systems relevant to cryospheric monitoring. It supports optimisations such as quantisation (FP32 $\rightarrow$ INT8) and graph fusion, reducing both memory footprint and latency.

Exporting Vision-Mamba (ViM-Seg) to ONNX was unsuccessful due to unsupported state-space operators (`SelectiveScan`, `BlockSSM`) that depend on Triton/CUDA kernels without ONNX symbolic definitions. During `torch.onnx.export()`, these dynamic scan operations could not be serialised into the static computation graph, resulting in incomplete conversion. Glacier-Seg demonstrates the lowest inference time (14.9 ms), outperforming SegFormer-B0 (21.8 ms) and running $\sim$4–5$\times$ faster than DeepLabV3 and ResNet50. Its 2.59 M parameters make it $\sim$11$\times$ smaller than ResNet50 and 4$\times$ smaller than EfficientNet-B3, validating the efficiency of its hybrid involution–Mamba–Transformer design. The ONNX-exported Glacier-Seg occupies only 10.68 MB, roughly 21$\times$ smaller than ResNet50-UNet, enabling deployment on embedded and UAV platforms with tight memory and power budgets. Its 14.9 ms latency corresponds to $\sim$67 frames/s on CPU—satisfying real-time requirements for glacier-front detection.

Even though we had significant reduction in all the parameters, we could not yet lower the Giga Floating Point

Table III: Comprehensive comparison of model complexity and ONNX CPU inference performance.

Model	Params (M)	Size (MB)	Inference (ms)	MACs (G)	FLOPs (G)	ONNX Time (ms)
Glacier-Seg	**0.68**	**2.64**	**2.58**	**3.76**	**7.52**	**14.91**
U-Net	31.0	∼120	4–6	7.6	15.2	99.6
ResNet-50	25.6	∼98	3–4	4.1	8.2	75.39
EfficientNet-B3	12.0	∼45	2–3	0.9	1.8	18.73
VGG-19	143.0	∼575	7–9	9.8	19.6	129.2
DeepLabV3	41.0	∼160	9–11	8.1	16.2	59.49
Vision Mamba	7.50	28.74	10.24	0.52	1.04	N/A.
SegFormer-B0	3.73	14.28	3.52	1.74	3.48	21.82

Operations (GFLOPs) count compared to the SOTA models. GFLOPs is a standard metric for assessing model efficiency, particularly relevant for deployment in resource-constrained environments like remote sensing applications for glacier mapping. This effectively highlights Glacier-Seg's position as a balanced, lightweight hybrid (Mamba–Transformer–CNN) in terms of efficiency for cryospheric monitoring. 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.

C. Limitations of the study

While this study provides valuable insights into glacier segmentation, several factors constrained its scope and depth:

- Limited access to high-performance GPUs allowed only a few full training runs per model, restricting extensive hyperparameter tuning, cross-validation, and ablation studies. Long training times for deep CNN and Transformer models also limited post-training analyses such as error inspection, threshold optimisation, and uncertainty estimation.
- Dataset 1 suffered from extreme class imbalance and noise due to exclusive SAR usage, while overall scarcity of large, well-annotated glacier datasets reduced generalisability. Multi-class performance was lower, with minority classes underrepresented, highlighting the need for improved sampling or adaptive loss strategies.
- On Dataset 2, unusually fast convergence within 3–4 epochs required learning-rate reduction (1×10^{-6}), suggesting overfitting to specific multimodal patterns.
- Limited time-series imagery (e.g., CaFFe with only four-year intervals) hindered evaluation of seasonal melt, calving, or mass-balance variations, restricting assessment of model performance in tracking glacier retreat dynamics.

VI. Conclusion and Future Work

This research introduces **Glacier-Seg**, a lightweight hybrid model that integrates convolutional inductive biases, transformer-based global reasoning, and Mamba-inspired state-space modelling for efficient multimodal glacier segmentation. Evaluated across SAR-only (CaFFe) and multimodal (NIRD, HKH) datasets, Glacier-Seg achieved accuracy comparable to or exceeding larger models such as SegFormer, DeepLabV3, and EfficientNet-UNet, while maintaining a compact 10.68 MB footprint and 14.9 ms CPU inference time. These results highlight its effectiveness for scalable and real-time cryospheric observation on resource-limited platforms.

Future developments will extend Glacier-Seg toward dynamic and predictive modelling. Incorporating recurrent and state-space components will enable temporal forecasting of glacier retreat and mass loss using multi-year Sentinel and Landsat archives. Expanding segmentation to multiple surface classes—debris, moraine, ice, water, and shadow—will improve understanding of glacier facies and meltwater pathways. Broader validation on datasets such as GLIMS and RGI 7.0, alongside cross-sensor calibration, will strengthen generalisability. Additionally, deployment through ONNX/TensorRT conversion and optimisation via pruning or quantisation will further reduce latency for drone and CubeSat applications.

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] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] S. Baraka, B. Akera, B. Aryal, *et al.*, *Hkh glacier mapping dataset*, 2020. [Online]. Available: https://lila.science/datasets/hkh-glacier-mapping.

[23] 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.

[24] M. Bao, *Vision mamba in remote sensing: A comprehensive survey of techniques, applications and outlook*, 2025. arXiv: 2505.00630 [cs.CV].

[25] M. R. AI, "Optimizing machine learning inference with onnx runtime," in *Microsoft Build 2023 Technical Proceedings*, Available at https://onnxruntime.ai/, Microsoft, 2023.