International Journal of Applied Earth Observation and Geoinformation 136 (2025) 104371

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag

Glacial lake mapping using remote sensing Geo-Foundation Model

Di Jiang [a,b,c], Shiyi Li [c,*], Irena Hajnsek [c,d], Muhammad Adnan Siddique [e], Wen Hong [a,b], Yirong Wu [a,b]

[a] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, 100101, China
[b] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing, 100049, China
[c] Institute of Environmental Engineering, ETH Zurich, Zurich, 8093, Switzerland
[d] Microwave and Radar Institute, German Aerospace Center (DLR), Wessling, Germany
[e] Information Technology University, Lahore, Pakistan

ARTICLE INFO

Dataset link: https://gitlab.ethz.ch/dijiang/transformer

Keywords:
Glacial lake
Geospatial Foundation Model
Sentinel-1 and 2
GaoFen-3

ABSTRACT

Glacial lakes are vital indicators of climate change, offering insights into glacier dynamics, mass balance, and sea-level rise. However, accurate mapping remains challenging due to the detection of small lakes, shadow interference, and complex terrain conditions. This study introduces the U-ViT model, a novel deep learning framework leveraging the IBM-NASA Prithvi Geo-Foundation Model (GFM) to address these issues. U-ViT employs a U-shaped encoder–decoder architecture featuring enhanced multi-channel data fusion and global-local feature extraction. It integrates an Enhanced Squeeze-Excitation block for flexible fine-tuning across various input dimensions and combines Inverted Bottleneck Blocks to improve local feature representation. The model was trained on two datasets: a Sentinel-1&2 fusion dataset from North Pakistan (NPK) and a Gaofen-3 SAR dataset from West Greenland (WGL). Experimental results highlight the U-ViT model's effectiveness, achieving an F1 score of 0.894 on the NPK dataset, significantly outperforming traditional CNN-based models with scores below 0.8. It excelled in detecting small lakes, segmenting boundaries precisely, and handling cloud-shadowed features compared to public datasets. Notably, the U-ViT demonstrated robust performance with a 50% reduction in training data, underscoring its potential for efficient learning in data-scarce tasks. However, its performance on the WGL dataset did not surpass that of DeepLabV3+, revealing limitations stemming from differences between pre-training and input data modalities. The code supporting this study is available online. This research sets the stage for advancing large-scale glacial lake mapping through the application of GFMs.

Contents

* Corresponding author.
E-mail address: shiyi.li@ifu.baug.ethz.ch (S. Li).

https://doi.org/10.1016/j.jag.2025.104371
Received 21 August 2024; Received in revised form 16 December 2024; Accepted 10 January 2025
Available online 21 January 2025

1. Introduction

Glaciers are highly sensitive to climate change, resulting in rapid shrinkage and retreat in the 21st century (Lee et al., 2023). This accelerated glacier mass loss has led to the formation and growth of various glacial lakes, such as ice-marginal, moraine-dammed, and supraglacial lakes (Hugonnet et al., 2021; Ke et al., 2024). The development of these lakes significantly influences glacier dynamics through processes of melting, infilling, and drainage. As glacial lakes expand, they reduce surface albedo, which increases melting and creates a feedback loop that further accelerates glacier melting and lake expansion, directly affecting glacier mass balance. When lake water infiltrates the ice-bed interface, it reduces friction and increases basal sliding, causing faster glacier movement and potential surges (Turton et al., 2021; Stevens et al., 2022). Sudden drainage events can trigger rapid hydrological and geomorphological changes, often leading to the collapse and retreat of glacier fronts (Shugar et al., 2020). Moreover, the outburst of lakes confined by ice or moraine dams can result in Glacial Lake Outburst Floods (GLOFs), which pose severe risks to downstream communities and infrastructure (Zhang et al., 2023b). Consequently, accurate mapping and monitoring of glacial lakes is essential for understanding these dynamics, assessing hydrological stability, and mitigating associated hazards.

Advancements in remote sensing platforms, such as Sentinel, Landsat, MODIS, GaoFen and ALOS, have made high-resolution, multi-sensor imagery available for glacial lake mapping (Arthur et al., 2020; Zheng et al., 2023). Although extensive datasets have been developed for glacial lake mapping in regions such as Greenland, High Mountain Asia (HMA), and the Alps, most rely exclusively on either optical or Synthetic Aperture Radar (SAR) data, with limited use of multi-sensor approaches (Shugar et al., 2020; Veh et al., 2023; Taylor et al., 2023; Nie et al., 2017; Chen et al., 2021). Multi-sensor datasets, which integrate SAR and optical imagery, have been shown to significantly enhance mapping accuracy by combining the strengths of both data types. These datasets enable all-weather, all-day detection of glacial lakes while reducing the effects of layover, shadows, and interference from surface features resembling lakes (Schröder et al., 2020; Jiang et al., 2022). Despite these advantages, the application of multi-sensor datasets is still in its early stages, primarily due to the scarcity of labeled multi-sensor datasets, which are critical for training and validating automated models (Xu et al., 2024). Furthermore, limited research has been conducted to explore how combining different data types can enhance extraction accuracy and reliability (Mustafa et al., 2024).

Traditional methods for glacial lake mapping often use manually engineered indices, such as the Normalized Difference Water Index (NDWI), Normalized Difference Snow Index (NDSI), and radar backscatter intensity (Wang and Sugiyama, 2024). These indices are typically applied in semi-automatic approaches with empirically determined thresholds or automatic methods like Random Forest, Support Vector Machines (Wangchuk and Bolch, 2020). Although effective, these methods depend on fixed thresholds and manual adjustments. Deep learning provides a transformative approach in this field by offering fully automated and scalable solutions. Current state-of-the-art models primarily utilize U-Net, DeepLabV3+, and Attention U-Net, which are designed to capture both global and local features of the imagery (Dirscherl et al., 2021; Kaushik et al., 2022; Walther et al., 2023). By training on large datasets, these models can be generalized more effectively to different regions and conditions. However, challenges persist due to the complexity of glacial environments. False predictions often arise near cloud edges, shadows, mixed pixels at lake boundaries, and wet ice surfaces, where distinguishing water from other features is difficult. Additionally, small lakes, especially those with muddy or frozen surfaces, are frequently misclassified. Lastly, the availability of labeled multi-sensor datasets, particularly contains specific regions or sensor types, remains limited (Xu et al., 2024).

To address these challenges, the new wave of remote sensing geo-foundation models (GFM) provides a promising solution. These models, pre-trained on large and diverse datasets, have the ability to capture a wide range of features, which can then be transferred to various downstream tasks (Li et al., 2021; He et al., 2022). Through fine-tuning, GFMs can be effectively applied with limited task-specific data, significantly reducing the dependence on large labeled datasets while achieving high generalizability across different geographical regions and environmental conditions (Hong et al., 2024; Zhang et al., 2023a). Notable examples include Meta's Segment Anything Model (SAM) and IBM-NASA's Prithvi (Kirillov et al., 2023; Jakubik et al., 2023). Recently there is growing interest in adopting GFMs across various fields, their direct adaptation for glacial lake mapping presents unique challenges. Critical considerations include how these models accommodate different input data modalities, (e.g., multispectral and SAR imagery) and how they can be fine-tuned specifically for the segmentation task to optimize accuracy. This is especially important in complex terrains or regions with cloud cover, where distinguishing lakes from other features becomes challenging.

In this work, we introduce a novel model architecture U-ViT based on the Prithvi for precise and efficient glacial lake mapping using multi-sensor data. The model is designed to operate across diverse regions and data types, even with limited training data. To our knowledge, this is the first effort to apply a foundation model to this specific downstream task. The U-ViT model adopts a U-shaped architecture, combining the strengths of a Vision Transformer (ViT) encoder with a CNN-based decoder to capture both global and local features. To optimize multi-sensor data fusion and ensure dimensional alignment for different inputs, we introduced an Enhanced Squeeze-Excitation (ESE) block. Additionally, an improved Inverted Bottleneck Block (IBB) enhances the model's ability to extract local features. This flexible

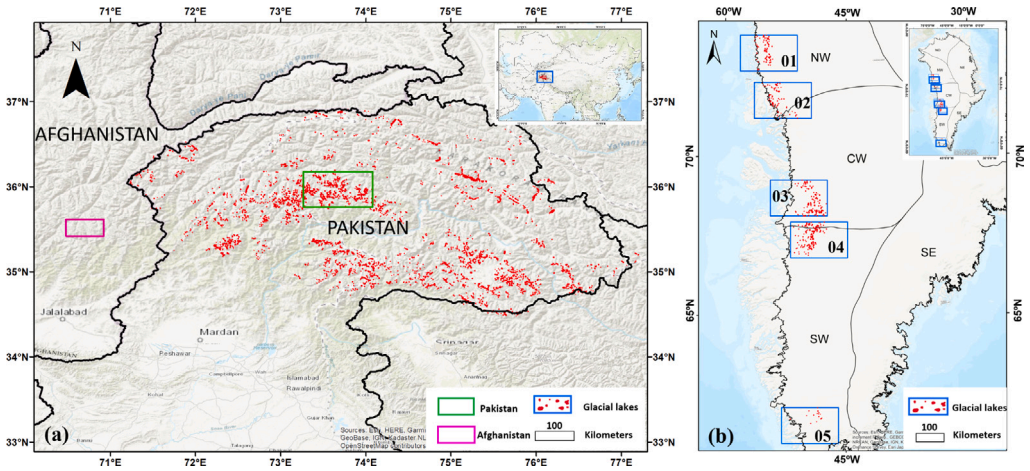

Fig. 1. Location of the study area for glacial lake mapping. The study areas include training datasets in North Pakistan (NPK, panel a) and West Greenland (WGL, panel b), along with two sites for comparison with public datasets. Glacial lakes within each region are highlighted by red polygons. (a) The NPK dataset contains 5,216 glacial lakes, covering a total area of 119.7 km^2 (red polygons). Two comparison sites from public datasets are highlighted: Pakistan (green rectangle) and Afghanistan (purple rectangle).(b) The WGL dataset is divided into five major regions (01–05, shown in the inset) and includes 1,605 glacial lakes with a total area of 111.66 km^2 (red polygons). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Remote sensing data used in this work.

Data source	Band	Resolution	Date	Region
Sentinel-1	VV	10 m	07/2022	NPK
Sentinel-2	B2,B3,B4,B8	10 m	07/2022-09/2022	NPK
GaoFen-3	HH, HV	10 m	02/2022-03/2022	WGL

architecture supports various ViT-based foundation models as backbones, allowing it to adapt across diverse datasets. We validated the model using our datasets from North Pakistan and West Greenland, and further tested its performance on public datasets, demonstrating its superior generalization across diverse environments. Extensive ablation and comparative experiments highlight the model's strengths in both accuracy and adaptability. The results demonstrate that the model performs well even in challenging conditions and minimizes the reliance on large labeled datasets, making it an efficient solution for glacial lake mapping.

2. Datasets

2.1. Training datasets

This work comprises two training datasets, the North Pakistan (NPK) dataset and the West Greenland (WGL) dataset. The NPK dataset is a multi-sensor fused dataset, composed of SAR imagery from Sentinel-1 and multispectral imagery from Sentinel-2. The WGL dataset, on the other hand, is a SAR-only dataset, constructed using dual-polarization SAR imagery acquired by the GaoFen-3 (GF-3) satellite. The selection of NPK and WGL datasets allows for a comprehensive examination of glacial lakes across different environments, including high-altitude mountainous terrain and ice sheets. Besides that, these two datasets help evaluate the proposed model's ability to manage different sensor inputs. A summary of the two datasets can be found in Table 1. Detailed descriptions of the two datasets are given below.

The NPK dataset covers the Hindu Kush, Karakoram, and Himalayan regions (Fig. 1(a)), featuring complex terrain such as mountains, deep valleys, and extensive glacier systems. It contains 5216 manually delineated glacial lakes, covering approximately 119.7 km^2, and includes various lake types like moraine-dammed, ice-dammed, proglacial, and supraglacial lakes. Most lakes are small, with over half measuring less than 0.001 km^2, leading to an imbalanced sample ratio of 0.006:0.994 (positive-to-negative) (Fig. 2(a)). In this dataset, we used the Sentinel-1 Ground Range Detected (GRD) product with VV polarization and Sentinel-2's RGB bands (Bands 2, 3, 4) along with the near-infrared (NIR) band (Band 8) (Torres et al., 2012; Drusch et al., 2012). All data bands are of the same spatial resolution of 10 m. VV polarization minimizes multipath effects, penetrates ice and snow, and improves the signal-to-noise ratio, making it suitable for glacial lake mapping. The RGB bands provide detailed surface information, distinguishing between ice, water, and land, while the NIR band is effective for detecting glacial lakes due to its high water absorption properties. The Sentinel-1 data were acquired in July 2022, and Sentinel-2 data were extended from July to September 2022 to ensure complete coverage and minimize cloud interference. Data were sourced from the Alaska Satellite Facility (ASF) and USGS Earth Explorer, respectively.

The WGL dataset focuses on the West Greenland region, characterized by vast ice sheets and dynamic glacial systems (Fig. 1(b)). It primarily features supraglacial lakes formed on the ice sheet surface under harsh polar conditions dominated by ice, snow, and meltwater, which is significantly different from the NPK region. The availability of optical data is limited due to the polar night and uniform white surfaces outside of the melt season, which makes it challenging to derive useful information from visible spectrum imagery. Therefore, the WGL dataset is based exclusively on GF-3 SAR imagery, which is sourced from the Open Spatial Data Sharing Project of the Aerospace Information Research Institute, Chinese Academy of Sciences. The dataset includes key regions such as the Upernavik Isstrøm system (Regions 01 and 02), Jakobshavn Glacier (Regions 03 and 04), and the Sermiligaarsuk Glacier in Southwest Greenland (Region 05). These areas are important for understanding glacier behavior, ice dynamics, and sea level rise contributions. The GF-3 imagery is acquired in Fine Strip II (FSII) mode with 10-meter resolution and dual-polarization (HH and HV) (Zhang and Liu, 2017). In total, the WGL dataset contains 1,605 glacial lakes, covering approximately 111.66 km^2. While lakes in WGL are slightly larger on average compared to NPK (Fig. 2(b)), the dataset still faces class imbalance, with a positive-to-negative sample ratio of 10:1.5, which affects model training and generalization.

2.2. Comparison dataset

We utilized an existing public glacial lake inventory dataset (Wang et al., 2020) to validate the efficiency and accuracy of the proposed U-ViT model. This dataset integrates glacier inventory data and Landsat imagery, applying the NDWI threshold along with manual visual

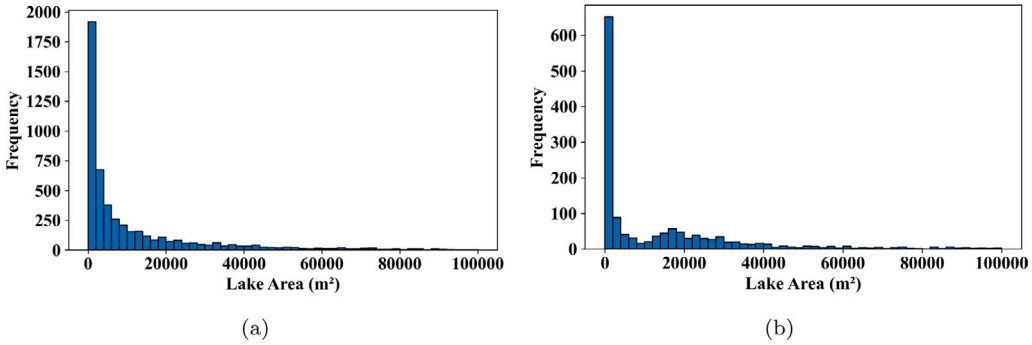

Fig. 2. The frequency distribution of glacial lake areas: (a) North Pakistan dataset and (b) West Greenland dataset.

interpretation to delineate glacial lakes. It identifies 27205 glacial lakes covering $1,806.47 \pm 2.11$ km^2 in 1990 and 30121 lakes covering $2,080.12 \pm 2.28$ km^2 in 2018 within the HMA, with lake sizes ranging from 0.0054 to 6.46 km^2. The dataset is available through the National Special Environment and Function of Observation and Research Stations Shared Service Platform. We selected this dataset for its monthly intervals, which provide greater temporal precision than annual datasets, minimizing uncertainties from image variability and seasonal changes. To assess the model's generalization and mapping accuracy, we applied it to two test sites from September 2018: one in Pakistan (3362 km^2) and another in Afghanistan (894 km^2), as shown in Fig. 1(a). These sites contain 426 and 174 glacial lakes, respectively.

3. Method

This section presents a detailed description of the data preprocessing, the proposed U-ViT model, and the experiments conducted for comparative and ablation studies.

3.1. Data preprocessing

For the NPK dataset, before applying data fusion, both SAR and multi-spectral images were pre-processed to further improve the data quality. For the Sentinel-1 VV GRD images, we applied the precise orbit file for accurate geolocation, removed thermal noise to enhance signal quality, eliminated border noise artifacts to generate clearer images, and performed radiometric calibration to convert raw SAR data to backscatter coefficients. The backscattering images were further filtered using the Lee filter with a 3×3 window to reduce speckle noise while preserving edges and processed with terrain correction using the SRTM 1-sec HGT product to correct geometric distortions. The corrected SAR backscattering images are finally converted to decibels (dB) for easier interpretation. The Sentinel-2 images used in this study were selected from the Level-2 Bottom-Of-Atmosphere (BOA) corrected reflectance product with a threshold on cloud cover less than 15%. Both Sentinel-1 and Sentinel-2 images were finally geocoded on a common coordinates grid and applied amplitude normalization. In the end, we concatenate together the processed data, resulting in a five-channel multi-sensor dataset, including B4 (red), B3 (green), B2 (blue), B8 (NIR), and VV. The dataset comprises 1,055 images in uint16 format, originally cropped to 320×320 pixels and subsequently re-scaled to 224×224 pixels.

For the WGL dataset, the preprocessing steps include radiometric calibration to correct the GF-3 SAR dual-polarization data for an accurate representation of the radar backscatter coefficient, formation of the C2 matrix containing the backscatter information, and the application of a 9×9 boxcar filter to reduce speckle noise. Following the preprocessing, we employed a model-based dual-polarization decomposition method to construct an RGB false-color image using Eq. (1):

$$\begin{aligned} R &= 10\log_{10}(m_s) \\ G &= 10\log_{10}(m_v) \\ B &= 10\log_{10}(m_s/m_v) \end{aligned} \tag{1}$$

where m_s the Bragg surface scattering component and m_v the random volume scattering component extracted using the decomposition model (Mascolo et al., 2022). In this false-color composite, the R,G,B bands represented the surface scattering component, the volume scattering component, and the difference between the surface and volume scattering component, respectively. This RGB synthesis enabled us to assess the dominant scattering mechanism, thereby providing valuable insights into the surface and structural properties of the observed objects. Finally, the WGL dataset comprised a total of 1799 images in float32, each with a crop size of 256×256 pixels, which were rescaled to 224×224.

In both the NPK and WGL datasets, glacial lakes are defined as natural water bodies formed directly or indirectly by glaciers or glaciation processes within the cryosphere (Dou et al., 2023; Zhang et al., 2022). This encompasses lakes created by glacier movement and those primarily fed by glacial meltwater, mainly including ice-marginal lakes, moraine-dammed lakes, ice-dammed lakes, and supraglacial lakes. During manual vectorization, glacial lakes were distinguished from ordinary lakes based on their geographical location in glaciated regions, primary water source, and distinctive morphological characteristics, such as texture and color, as observed in SAR and optical imagery (Wang et al., 2020). This process relied on expert knowledge to ensure accurate classification.

3.2. U-ViT architecture

The proposed U-ViT architecture is shown in Fig. 3. It is composed of an ESE block, a U-shaped encoder–decoder structure, and a lightweight CNN module composed of three IBBs. In the encoder–decoder part, the encoder takes the ViT backbone as a feature extractor, and the decoder is composed of three up-sampling and CNN layers. The U-ViT model introduces several innovative features aimed at addressing key challenges in glacial lake mapping and improving segmentation accuracy across various data types. The main contributions of the U-ViT model are as follows:

Handling diverse input data types: The U-ViT integrates an ESE block to address the challenge of processing multi-sensor data, such as SAR and optical imagery. The ESE block aligns input features from different data types, allowing the model to effectively handle various input bands without requiring major modifications. This flexibility eliminates the model's dependence on specific data types, enabling it to be trained and applied using a wide range of input sources.

Channel-wise attention for multi-band data: For multi-band input data, the ESE block dynamically assigns different weights to each

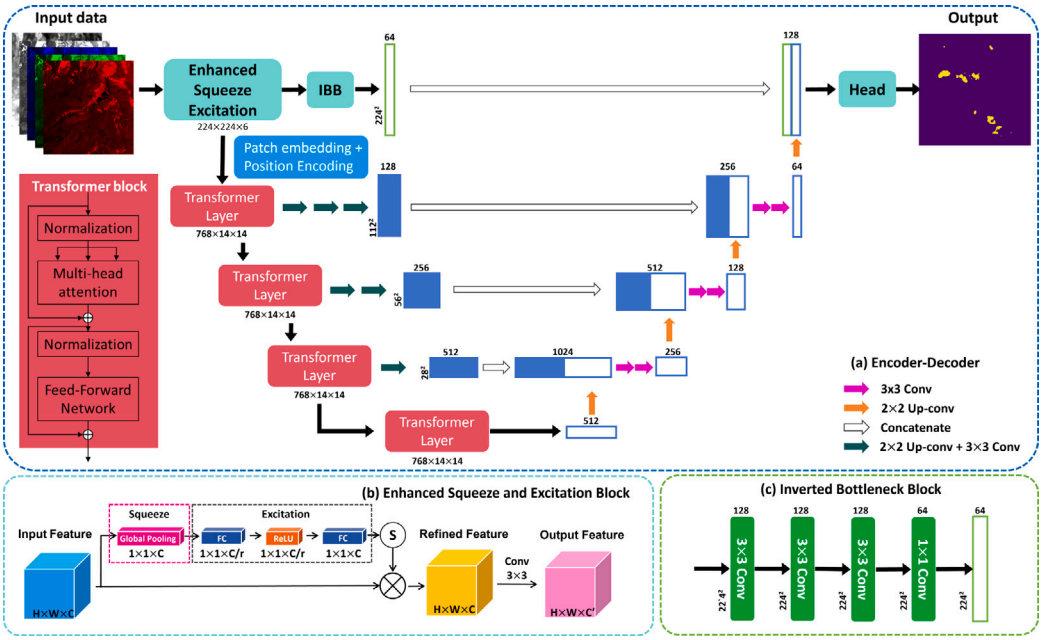

Fig. 3. Hybrid U-ViT model architecture for multi-sensor glacial lake extraction. The main structure (a) of the proposed model shown in the middle is composed of an ESE block (b), a U-Net like encoder–decoder (a), and a lightweight IBB block (c). Feature maps generated from the Encoder–Decoder and the IBB are concatenated together and fed into a segmentation head to predict the final glacial lake maps. In (a), the encoder uses the ViT backbone, which has 12 transformer blocks that are selected into four layers.

channel, allowing the model to focus on the most informative bands. This capability ensures that the model prioritizes the channels that contribute the most useful information for segmentation, improving accuracy by reducing the influence of less relevant data.

Transfer learning and efficient training on small datasets: The U-ViT model benefits from the pre-trained Prithvi GFM, which was trained on harmonized large-scale Landsat and Sentinel-2 datasets. This pre-training allows U-ViT to be fine-tuned with relatively small labeled datasets, significantly improving its transferability across various segmentation tasks, such as flood mapping and glacier extraction. This transfer learning capability reduces the need for extensive labeled data and makes the model suitable for deployment in regions with limited training data.

Combining CNNs and ViTs for superior feature extraction: The hybrid architecture of U-ViT leverages the strengths of both CNNs and ViTs. The CNN layers like IBBs excel at capturing local details, while the ViT layers are adept at modeling long-range dependencies and global context. By combining these two approaches, U-ViT achieves a balanced capability to extract both local and global features, leading to higher segmentation accuracy and better generalization across different environments and data types.

Through these innovations, the U-ViT model addresses critical challenges in multi-source remote sensing data processing, channel prioritization, efficient training on small datasets, and comprehensive feature extraction. These advancements enable the model to perform robustly across diverse segmentation tasks, making it a versatile tool for applications such as glacial lake mapping, monitoring, and other geospatial analysis tasks.

3.2.1. Enhanced squeeze-excitation block

To improve the data fusion of optical and SAR bands and align the input bands with the pre-trained ViT backbone parameters, we designed the ESE block as shown in Fig. 3(b).

In the ESE block, an input tensor is processed through two phases. The squeeze phase first applies global average pooling to capture channel-wise global information. The excitation phase subsequently processes the squeezed output with two fully connected layers including ReLU and sigmoid activation functions to generate per-channel weights that highlight important features while suppressing less relevant ones (Hu et al., 2018). The weights are further used to scale input channels to generate refined features, and a convolution layer is followed by the feature scaling to finally produce the fused features that match with the encoder backbone requirements. With the ESE block, the model is able to take input of arbitrary depth and dynamically adjust the input data fusion through training. The whole processing steps of the ESE block can be summarized as Eq. (2):

$$\begin{aligned} \text{Squeeze}: \quad & Z = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_{ij} \\ \text{Excitation}: \quad & S = \sigma(W_2 \cdot \text{ReLU}(W_1 \cdot Z)) \\ \text{Scaling}: \quad & X_{\text{scaled}} = X \cdot \text{view}(1, 1, C) \\ \text{Convolution}: \quad & Y_{\text{output}} = \text{Conv}(X_{\text{scaled}}) \end{aligned} \tag{2}$$

where $X \in R^{H \times W \times C}$ in height H, width W and depth C the input tensor, $\sigma(\cdot)$ the sigmoid activation function, W_1 and W_2 the weights in size (C, S) and (S, C) respectively, and $\text{Conv}(\cdot)$ the convolution kernel with kernel size 3 and padding of 1.

3.2.2. Encoder–decoder structure

In the encoder–decoder structure, the ViT based encoder takes the output of the ESE block as input to generate latent feature maps and passes both the final and intermediate feature maps to the decoder via skip connections (Fig. 3(a)).

In the encoder, the ViT backbone is initialized using pre-trained weights from the Prithvi GFM. The output of the ESE block is firstly partitioned and flattened into a sequence of small non-overlapping patches and sinusoidal position encoding is added to the patch embedding to keep the position information of patches. We used a patch size of 14×14 to align with the pre-training parameter of Prithvi. After positional encoding, 12 transformer blocks are sequentially connected to process the patch tokens. Each transformer block is composed of alternating layers of multi-headed self-attention and feed-forward network, with Layer Norm (LN) and residual connections applied before and after every block respectively. In ViT, transformer blocks of different levels have varying receptive fields. Shallow-level transformer blocks mainly focus on low-level features such as edges and textures, whereas deeper-level blocks gradually capture more global and high-level semantic

features. To better leverage the rich latent features learned by ViT and meanwhile reduce the computational cost, we grouped the transformer blocks into four layers, then extracted the feature maps generated by blocks 3, 6, 9, 12 as intermediate and final feature maps to pass to the decoder (Hatamizadeh et al., 2022).

The decoder is an expansive path that has a symmetrical layered structure as the encoder. Within each decoder layer, feature maps generated from the ViT encoder are concatenated with features from the last layer. To unify the feature map dimension of different encoder layers, the feature maps from the last layer are up-sampled using 2×2 up-convolution at the respective sampling rate (i.e. [1, 2, 4, 8]) and processed with a convolution block containing a 3×3 convolution kernel, batch normalization (BN) and ReLU activation has been applied before concatenation. The concatenated feature maps of each layer are again processed twice with the 3×3 convolution block and are passed to the upper layer by a 2×2 up-convolution kennel that halves the feature channels, until the top layer in which the output feature is concatenated with the IBB feature for segmentation.

3.2.3. Inverted bottleneck blocks

In contrast to traditional bottleneck residual block (He et al., 2016) with a wide-narrow-wide structure in depth, the IBB follows a narrow-wide-narrow structure, containing a projection layer, a depth-wise convolution layer, and an expansion projection layer. It is a lightweight CNN block that enables efficiently extracting local features. We used IBB to construct the lightweight CNN block to incorporate the CNN inductive bias into the model (Fig. 3(c)).

Following the IBB block, features from the encoder–decoder and IBB are concatenated and fed into the segmentation head for glacial lake mapping. In this way, both global and local information from the input multi-sensor data are effectively integrated.

3.3. Experiments design

In this study, all experiments are conducted on a single NVIDIA GeForce RTX A6000 GPU. For both the NPK and WGL datasets, we have randomly split our training datasets into two portions for training (90%) and validation (10%) according to common practice (Chicco, 2017; Xu et al., 2024). We used the AdamW (Loshchilov, 2017) optimizer and a one-cycle cosine annealing scheduler during training, with a maximum learning rate set to 10^{-3} and a batch size of 16. Additionally, we conducted validation using the public glacial lake inventory dataset (Wang et al., 2020), including those not used in model training, to further ensure the model is not overfitting to the training data and effectively generalized to unseen data.

3.3.1. Loss function

During the training, we used the focal loss function to force the model to focus on hard-to-classify samples and to address the imbalance issue within the dataset (Lin et al., 2017). The focal loss is defined as Eq. (3):

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \tag{3}$$

where

$$p_t = \begin{cases} p & \text{positive sample } (y=1) \\ 1-p & \text{negative sample } (y=0) \end{cases}$$

the model's predicted probability for the true class, α_t the class weighting factor chosen based on positive–negative sample imbalances, and γ the focusing parameter that attenuates the model's focus on easy-to-classify samples. In our experiments, we chose $\alpha_t = 0.03$ and $\gamma = 2$ for NPK datasets and $\alpha_t = 0.8$ and $\gamma = 2$ for WGL datasets.

3.3.2. Evaluation metrics

The evaluation metrics used in this work include overall accuracy (OA), mean intersection over Union (mIoU), mean F1 score (F1), Precision, and Recall with the formulation (4) (5) (6) (7) respectively. The overall accuracy measures the correctness of the model's predictions as the average accuracy across all classes.

$$\mathrm{OA} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{4}$$

The mean IoU measures the overlap between the predicted and actual regions, averaged across all classes.

$$\mathrm{mIoU} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} \tag{5}$$

The F1 score balances precision and recall through their harmonic mean, averaged across all classes.

$$\mathrm{F1} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}} \tag{6}$$

Here the precision is a measure of the accuracy of the positive predictions. Recall represents the ability of the model to identify all the actual positive cases.

$$\begin{aligned} \mathrm{Precision} &= \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FP})} \\ \mathrm{Recall} &= \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FN})} \end{aligned} \tag{7}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

3.3.3. Model comparison

We compared the proposed U-ViT architecture against the state-of-the-art remote sensing image segmentation methods on the glacial lake mapping task, including U-Net, DeepLab V3, and MAnet.

- U-Net is one of the most popular models in glacial lake extraction, utilizing a symmetrical encoder–decoder architecture with skip connections to capture both low-level and high-level semantic information (Ronneberger et al., 2015).
- DeepLabV3+ enhances lake extraction in complex backgrounds by combining ASPP to capture multi-scale contextual information, offering improved accuracy compared to simpler architectures like U-Net (Chen et al., 2018).
- MAnet improves upon U-Net models by introducing Position-wise Attention Block and Multi-scale Fusion Attention Block, to capture both the spatial and channel dependencies between the feature map. This reduces the impact of irrelevant information or noise, enhancing the accuracy and robustness of lake classification (Fan et al., 2020).

Table 2
Performance comparison of different models.

Dataset	Index	U-Net	DeepLabV3+	MAnet	Proposed
NPK	Precision	0.6521	0.6617	0.6318	**0.8707**
	Recall	0.8669	0.8647	0.8503	**0.9219**
	F1-score	0.714	0.7232	0.6902	**0.8946**
	mIoU	0.6335	0.6413	0.6143	**0.8256**
	OA	0.9911	0.9918	0.9897	**0.998**
WGL	Precision	0.8079	**0.8610**	0.7589	0.7396
	Recall	0.8186	**0.8537**	0.7584	0.6986
	F1-score	0.8152	**0.8572**	0.7587	0.7116
	mIoU	0.6888	**0.7545**	0.6205	0.5696
	OA	0.8381	**0.8801**	0.7955	0.7747

4. Results

4.1. Model comparison

The performance of different models on the NPK and WGL datasets is summarized in Table 2.

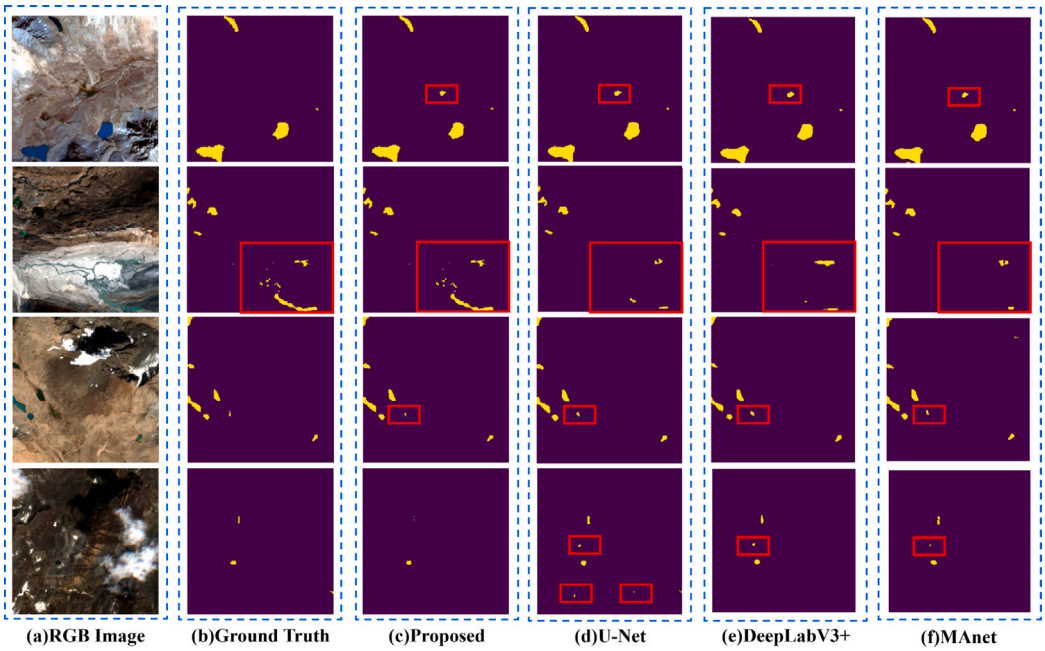

Fig. 4. The visualization results for the proposed model, U-Net, DeepLab V3+, and MAnet on the NPK dataset are displayed from (c) to (f), respectively. Each of the four rows represents a different region. (a) shows the color composition using the red, green, and blue bands from Sentinel-2 imagery. (b) presents the corresponding manually labeled ground truth. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

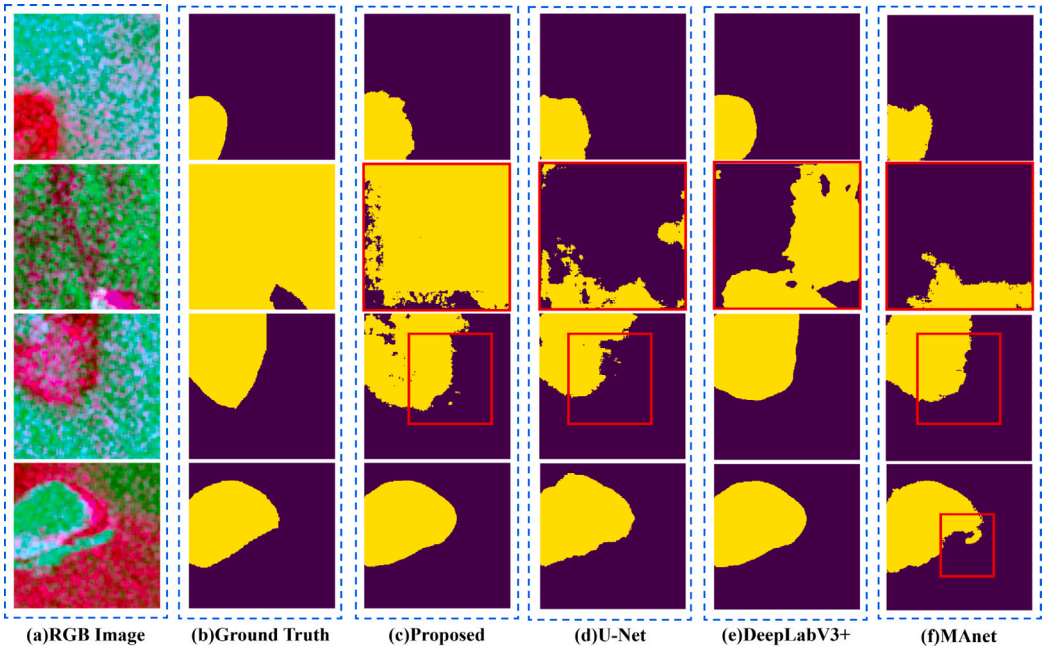

Fig. 5. The visualization results of the proposed model, U-Net, DeepLab V3+, and MAnet on WGL dataset are shown from (c) to (f), respectively. (a) depicts the false color composition based on Eq. (1) from Sentinel-1 dual-polarization imagery, while (b) presents the corresponding manually delineated labels considered as ground truth.

On the NPK dataset, the proposed model outperformed all compared CNN-based models across every evaluation metric and achieved an F1 score of 0.8946 and an mIoU of 0.8256, showing a significant improvement compared to the best CNN-based model, DeepLabV3+, which had an F1 score of 0.7232 and an mIoU of 0.6413. Moreover, the proposed model exhibited a balanced precision and recall of 0.8707 and 0.9219, respectively, whereas the CNN-based models demonstrated high recall and low precision with a difference of more than 0.2. This indicates that the proposed model achieves a good trade-off between precision and recall, while the CNN-based models are prone to higher false positives, leading to higher recall and lower precision.

In contrast, the best performance on the WGL dataset was achieved by DeepLabV3+, with an F1 score of 0.8572 and an mIoU of 0.7545. The proposed model only achieved an F1 score of 0.7116 and an mIoU of 0.5696. In this case, the CNN-based models no longer showed imbalanced precision and recall. Instead, the proposed model exhibited a slightly imbalanced precision of 0.7396 and recall of 0.6986, with a difference of about 0.04. It was also noted that the mIoU of all models was generally lower than 0.8, likely due to insufficient training data.

Visual comparisons of the segmentation results from different models are shown in Figs. 4 and 5 for the NPK and WGL datasets, respectively.

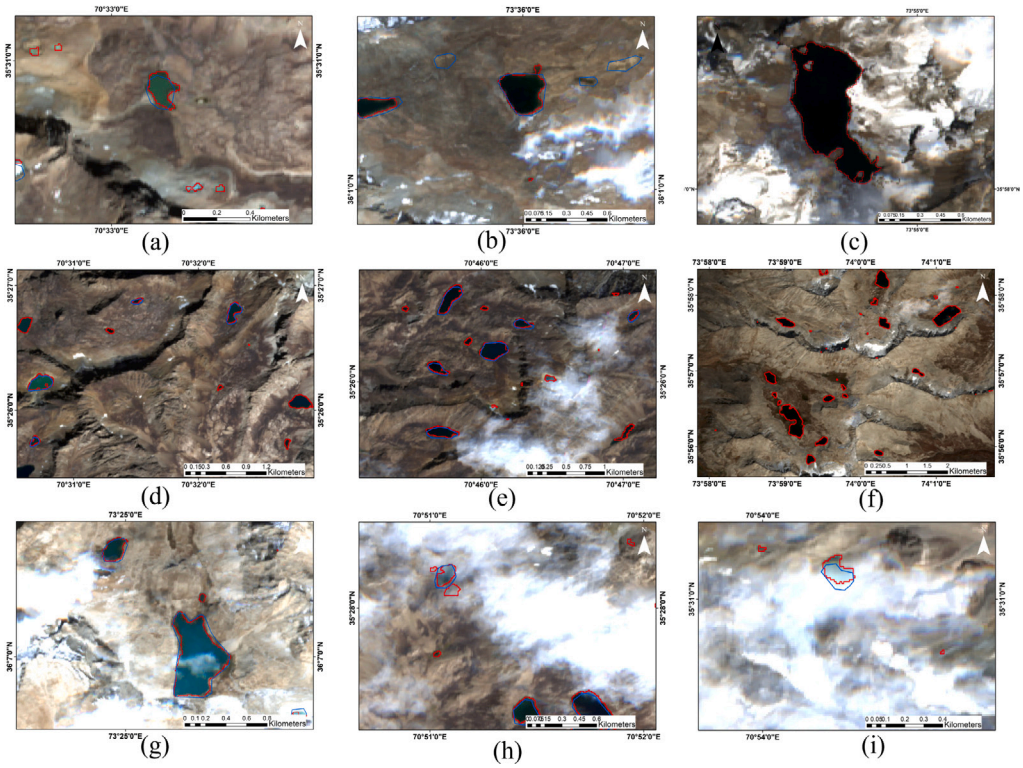

Fig. 6. Comparison of proposed model mapped glacial lake boundaries (red lines) with the public dataset (blue lines) in North Pakistan and Afghanistan. The proposed model produces more detailed and precise boundaries (a–d), particularly in regions where the public dataset lacks accuracy, such as with small lakes (< 0.01 km^2) that were previously ignored (d–f). Additionally, U-ViT demonstrates strong performance in detecting lakes in challenging conditions (f,c), effectively mapping lakes under cloud shadows (c) and in areas with cloud cover (g–i). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

In Fig. 4, the RGB images in column (a) were generated using Sentinel-2 RGB bands, and the ground truth are shown in column (b). In the first and fourth rows, false positives presented in the segmentation results are highlighted by the red boxes, showcasing the relatively higher occurrence of false positives in CNN-based models (column (d) to (f)) compared to the proposed model (column (c)). In the second and third row, the spatial discontinuity and unclear lake boundaries are emphasized by the red boxes in the CNN-based models results. In contrast, the proposed model demonstrated better segmentation of lakes given irregular and challenging lake shapes.

In Fig. 5, the false RGB image in column (a) was generated with the SAR polarimetric decomposition as per Eq. (1). In the first row, all models successfully detected the lake in the scene but showed different edge extraction results. The second row represents a challenging scene where the lake exhibits a complex scattering mechanism across almost the entire image. In this case, the proposed model successfully captured most of the ground measurements, whereas the CNN-based models failed in most of the scenes. In the third and fourth rows, blurring and jagged edge extraction was observed in the red boxes across all tested models except for DeepLabV3+.

4.2. Comparison with public dataset

Based on the mapping results in Pakistan and Afghanistan, the proposed model shows significant improvements in automated glacial lake mapping over the public dataset, demonstrating strong spatiotemporal transferability across large regions.

As shown in Fig. 6, the proposed method offers several advantages when working with multi-source data and effectively manages diverse and complex terrains, including normal conditions (Fig. 6(a, b, d)), cloud cover (Fig. 6(g, h, j)), cloud shadows (Fig. 6(e, g, h, i)), and glacial lakes on islands within larger water bodies (Fig. 6(c)). The model accurately maps glacial lakes of various sizes, from large (> 0.1 km^2) to small (< 0.01 km^2), and is particularly effective at detecting very small lakes (as small as 0.000164 km^2) (Wangchuk and Bolch, 2020), outperforming the public dataset in these cases (Fig. 6(b, e)). For example, several glacial lakes, particularly in northeastern Pakistan, that were absent in public datasets were accurately detected by our model (Fig. 6(f)). Additionally, the model excels in detecting lakes with deeper water and darker appearances (Fig. 6(c, d, g)), consistently producing more detailed and precise lake boundaries than the public dataset in most cases.

Notably, compared to public datasets using NDWI thresholding and manual methods, our model offers greater accuracy with fewer false positives, significantly reducing human bias in glacial lake inventories. For instance, glacial lakes covered by ice, snow and cloud, often undetectable in optical images—were successfully identified using our multi-sensor approach, which leverages the benefits of SAR imagery (Fig. 7(a,c)). Furthermore, the model effectively distinguishes false positives: regions that appear lake-like in optical images (blue polygons in Fig. 7 (b, d)) are correctly identified as non-lakes when cross-checked with SAR data, as shown in Fig. 7 (f, h).

Despite the public dataset's false predictions, we used it to calculate standard evaluation metrics to assess our model's accuracy quantitatively. In Pakistan, the model achieved a precision of 0.8091, a recall of 0.8323, and an F1 score of 0.8184, indicating a strong balance between precision and recall. Similarly, in Afghanistan, the model obtained a precision of 0.7659, a recall of 0.874, and an F1 score of 0.8103, demonstrating the model's ability to maintain high recall while providing balanced overall accuracy across different regions.

4.3. Ablation study

To evaluate the impact of each component of the proposed model on its performance, we conducted a series of ablation experiments using the NPK dataset. We maintained the same hyper-parameters, including

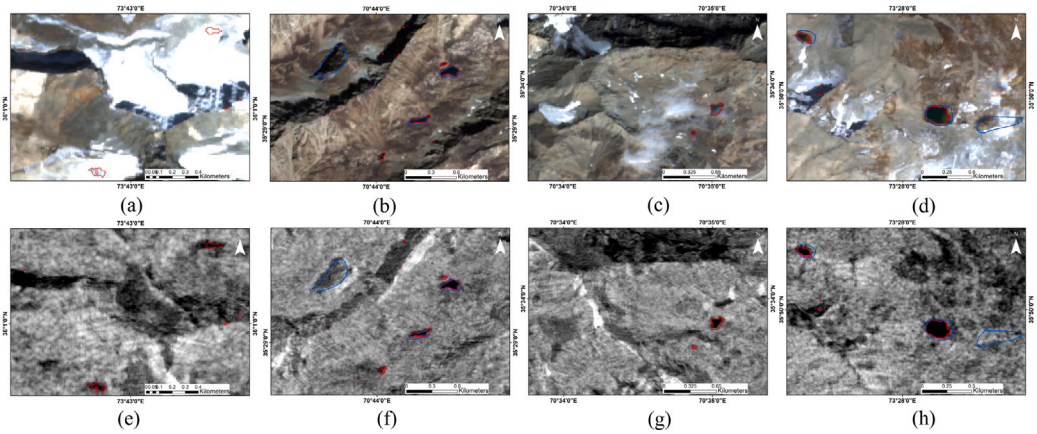

Fig. 7. Comparison of glacial lake boundaries mapped using the proposed multi-source method with those from public datasets based solely on Landsat imagery. The red lines represent the results extracted by the proposed model using multi-source data, while the blue lines indicate the lake boundaries derived from the existing datasets using NDWI thresholding and manual interpretation. Panels (a), (c), (e), and (f) show RGB composite images, while panels (b), (d), (g), and (h) correspond to VV polarization SAR images. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Results of ablation study. The ESE block was replaced by duplicating the green band, the IBB block was replaced by a simple Conv block, and the ViT-encoder was replaced by the vanilla U-Net encoder.

Module	Precision	Recall	F1	mIoU
Enhanced SE	0.7809	0.9226	0.8373	0.7537
IBB	0.8139	0.9155	0.8575	0.7777
ViT-encoder	0.7336	0.8763	0.7880	0.7010

learning rate and batch size, and tested the Enhanced SE block, IBB block, and Transformer Encoder, as detailed in Table 3.

First, we replaced the Enhanced SE block with band replication to produce a 6-band image input for the ViT encoder. The NIR band, which yielded the highest F1 score when duplicated, was concatenated with the raw input. This band replication approach achieved an F1 score of 0.8373 and an mIoU of 0.7535, both lower than those of the original model.

Next, we replaced the IBB block with a simple convolution block consisting of a Conv kernel (3 × 3, 64 depth), ReLU activation, and Batch Normalization. This modification resulted in an F1 score of 0.8575 and an mIoU of 0.7777, again with lower performance than the original model using the IBB block.

Additionally, we substituted the ViT encoder with a traditional CNN-based U-Net encoder while retaining the Enhanced SE block and IBB block at the top layer of the U-structure. This comparison aimed to evaluate the feature extraction capabilities of the transformer encoder versus the CNN. The U-Net encoder achieved an F1 score of 0.7880 and an mIoU of 0.7010, indicating a significant performance drop compared to the ViT encoder. Notably, the U-Net with the Enhanced SE block and IBB block outperformed the vanilla U-Net shown in Table 2, suggesting that these added components effectively improved model performance regardless of the encoder type.

4.4. Model performance using different backbones

To further evaluate the influence of the backbone on the performance of the proposed model in the glacial lake mapping task, we compared the Prithvi and ViT-MAE-Base backbones using the NPK and WGL datasets. Both Prithvi and ViT-MAE-Base share the same ViT structure and MAE pretraining method. Prithvi was pre-trained on HLS datasets and has a parameter size of about 127MB, while ViT-MAE-Base was pre-trained on the ImageNet-1K dataset and has a parameter size of about 86MB. The model performances using the two backbones are shown in Fig. 8.

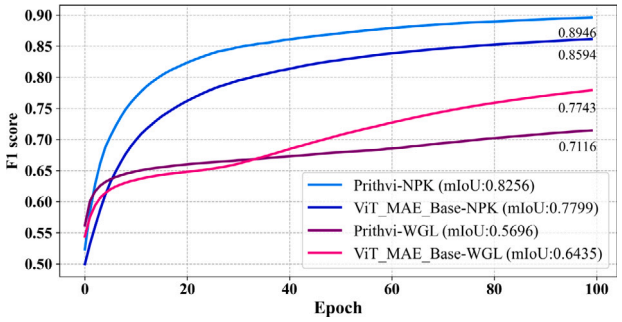

Fig. 8. F1 score on different large pre-trained models. The light blue and dark blue represent the F1 scores obtained from the transformer encoder pre-trained with Prithvi and ImageNet-1K, on the NPK dataset respectively. Pink and purple represent the F1 scores obtained from the transformer encoder pre-trained with Prithvi and ImageNet-1K on the WGL dataset respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

On the NPK dataset, Prithvi achieved an F1 score of 0.8946 and an mIOU of 0.8256, surpassing ViT-MAE-Base by 0.8594 in the F1 score and 0.7799 in mIoU. This superior performance can be attributed to the NPK dataset's multispectral channels from Sentinel-2, which are similar to the modality of Prithvi's pretraining dataset. This similarity likely helped Prithvi to quickly converge and achieve better segmentation results on the NPK dataset.

In contrast, on the WGL dataset, Prithvi was underperformed by ViT-MAE-Base. Prithvi achieved an F1 score of 0.7116 (0.7743 in ViT-MAE-Base) and an mIOU of 0.5696 (0.6435 in ViT-MAE-Base). The WGL dataset features a false RGB composition created with SAR polarimetric decomposition, differing significantly in modality and input channels from Prithvi's pretraining dataset. While the WGL dataset's closer alignment with standard RGB input types in ViT-MAE-Base's pretraining, enables it to achieve superior transfer learning performance on this dataset.

4.5. Contribution of different bands

We conducted a comparative study on different band combinations using the NPK dataset to quantitatively analyze the importance of different spectral bands in glacial lake segmentation. The study aimed to evaluate the contribution of various bands in the multi-sensor fusion data. The results are presented in Table 4.

Compared to the all-band input, the removal of the NIR band resulted in the most significant performance degradation, with the F1

Table 4
Experiments results for different band combinations of the proposed model. Where each curve shows the F1 score without the R, G, B, NIR, and VV bands separately.

Band combination	Precision	Recall	F1	mIoU
B,G,NIR,VV	0.8425	0.9168	0.8759	0.8008
R,B,NIR,VV	0.8287	0.8976	0.8599	0.7806
R,G,NIR,VV	0.8278	0.9104	0.8644	0.7862
R,G,B,VV	0.8242	0.8898	0.8540	0.7734
R,G,B,NIR	0.8331	0.9109	0.8679	0.7905

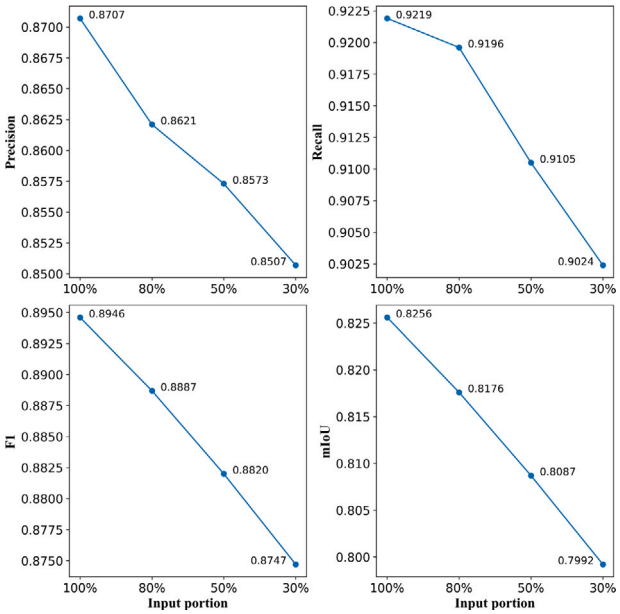

Fig. 9. Performance of the model trained with different proportions of the total sample size.

score dropping from 0.8946 to 0.8540 (a decrease of 0.0406) and the mIoU dropping from 0.8256 to 0.7734 (a decrease of 0.0522). This highlighted that the NIR band had the most impact on model performance. The NIR band's high sensitivity to water absorption makes it crucial for accurate glacial lake segmentation.

Following the NIR band, the removal of the green, blue, and VV bands sequentially resulted in decreasing impacts on model performance with a significant increase of the F1 scores and the mIoU values that approach the scores of the all-band inputs. The green and blue bands are effective due to water's high reflectance in these wavelengths. The VV channel, which helps delineate water bodies with lower backscattering intensities on water surfaces, is less impactful for glacial lakes due to interference from floating objects such as ice and rocks.

Lastly, the removal of the R band had the least impact on performance. This minimal effect is primarily due to the similarity between water and ice in the R band, along with the presence of vegetation and shadows on lake boundaries, which reduces the R band's distinctiveness in identifying glacial lakes.

4.6. Sensitivity on dataset sizes

To evaluate the generalization capability of the proposed model and its advantage on small sample datasets, we conducted experiments using different portions of the training data from the NPK dataset. Specifically, we trained the model with 80%, 50%, and 30% of the training set (total size 941) and used the same test set for evaluation. The results are detailed in Fig. 9.

With 80% of the training data, there was a slight decrease in all metrics, but performance remained relatively high. This suggests that a modest reduction in training data does not significantly impair the model's effectiveness. With 50% of the training data, precision and recall continued to decline, but the mIOU remained unchanged compared to using the complete training data. This indicates that the model still effectively delineates regions despite having less training data. However, with only 30% of the training data, all metrics reached their lowest values, with the mIOU dropping below 0.8.

These results demonstrate the model's robustness in handling data imbalance and variations in data volume. Additionally, the model's ability to maintain high accuracy with only 50% of the training data highlighted its potential for scenarios where only limited labeled data is available.

5. Discussions

5.1. Model comparison analysis

In this work, the proposed model improves the accuracy of glacial lake detection by combining advanced architectural features, robust transfer learning, and multi-sensor data fusion. The U-ViT architecture enhances feature extraction by capturing both global and local spatial details, enabling precise boundary delineation, especially for small or irregularly shaped lakes. Transfer learning with the Prithvi backbone, pretrained on domain-specific geospatial data, ensures strong performance even with limited labeled datasets. Multi-sensor fusion of SAR and optical data mitigates the effects of clouds, shadows, and terrain occlusions, creating a more complete and reliable input for segmentation. Additionally, the model's ability to minimize noise and reduce false positives produces cleaner segmentation outputs, resulting in more accurate and consistent mapping of glacial lakes.

However, when evaluating the proposed model on two different datasets. It outperformed CNN-based models on the NPK dataset but was underperformed by them on the WGL dataset. These contrasting results suggest that the similarity between the input data modality and the data used for backbone pretraining can significantly impact a model's performance when employing transformers as encoders. In detail, on the NPK dataset, the input images share the same four multispectral channels as the Prithvi pretraining dataset, including the RGB bands and the NIR band. This alignment likely contributed to the model's superior performance on the NPK dataset. In contrast, the WGL dataset was created using solely SAR polarimetric decomposition and does not share any similar channels with Prithvi. Consequently, the model's performance deteriorated on the WGL dataset, which was expected given the lack of similarity between the input data and the pretraining data.

This observation is further supported by the comparison of different backbones on the two datasets. As shown in Section 4.4, Prithvi excels in environments similar to its pretraining data on the NPK dataset, while ViT-MAE-Base performs better with RGB-like inputs on the WGL dataset. This highlights the importance of selecting an appropriate backbone based on the specific characteristics of the target dataset. These findings underscore the necessity of considering data modality alignment in model training and the potential benefits of multi-sensor data fusion for improving segmentation accuracy. It also implied that pre-training a foundation model with SAR data included can bring stronger generalization capabilities and further expand the range of applications of foundation models.

We also tested the model performance on dataset of reduced sizes, which demonstrated robust performance when reducing the training dataset to only 50% of the original sizes. This can be primarily attributed to the pretraining of Prithvi on domain-specific geospatial knowledge, which makes it less sensitive to the dataset size used in the fine-tuning stage. For glacial lake mapping tasks, the reduction of required labeled data can be very advantageous, given that the labeling effort of glacial lake samples can be very time-consuming and often requires expert knowledge.

5.2. Accuracy on public dataset

When evaluated against the public dataset, the model's performance metrics are slightly lower than those achieved on the NPK dataset. These discrepancies can be attributed to several factors, including differences in image resolution, datatype, minimum lake area thresholds, and annotation standards.

The NPK dataset uses 10 m resolution Sentinel-1 and Sentinel-2 imagery, which provides higher spatial detail compared to the 30 m resolution Landsat −8 imagery used in the public dataset. This higher resolution allows the proposed model to detect smaller glacial lakes that are often missed in the public dataset. This difference is further amplified by the datasets' minimum lake area thresholds: the proposed model detects lakes as small as 100 m^2, whereas the public dataset only includes lakes larger than 5,400 m^2. Additionally, the inclusion of multi-sensor data in the NPK dataset, particularly the integration of SAR imagery, enhances detection accuracy in regions obscured by snow or shadows, where optical imagery alone is less effective.

Another source of discrepancy arises from differences in annotation standards between the two datasets. In the NPK dataset, all water bodies associated with glaciers or glaciation in the alpine cryosphere were classified as glacial lakes. However, in certain cases, it was challenging to determine whether a water body qualified as a glacial lake or not. Examples include long, narrow water bodies along rivers and instances where the number of pure water body pixels was minimal, making classification uncertain. These differences in annotation criteria lead to variations in the results, as some lakes detected in the NPK dataset may not be included in the public dataset.

To further understand the discrepancies between the two results, we carried out lake area uncertainty analysis across Pakistan and Afghanistan, focusing on the effects of varying image resolutions and minimum area thresholds (Dou et al., 2023). For large lakes (> 0.1 km^2) and medium-sized lakes (0.01–0.1 km^2), the uncertainty levels of the proposed model were comparable to those of public datasets. Specifically, the proposed model yielded average uncertainties of 1.08% and 3.83% for large and medium lakes in Pakistan, and 0.96% and 4.17% in Afghanistan. The public datasets showed similar results, with uncertainties of 0.78% and 3.13% for Pakistan, and 0.81% and 3.39% for Afghanistan.

However, for small lakes (< 0.01 km^2), the proposed model showed significantly higher uncertainty than public datasets. In Pakistan, the uncertainty reached 74.91% in the proposed model compared to 7.57% in the public dataset, while in Afghanistan, these values were 69.46% and 7.81%, respectively. This discrepancy arises from differences in minimum area thresholds: the proposed model detects lakes as small as 0.0001 km^2, whereas the public dataset uses a higher threshold of 0.0054 km^2. Since uncertainty is directly related to lake size and shape, smaller lakes are more difficult to delineate accurately, contributing to the higher uncertainty in the proposed model.

While the lower threshold in the proposed model allows for a more comprehensive representation of small lakes, the higher uncertainty reflects the challenges in comparing these lakes in different datasets, particularly due to their small size and different resolutions. Additionally, differences in glacial lake distribution extents and minimum area thresholds make direct quantitative comparisons between the two datasets complex. Despite these challenges, the proposed model offers notable advantages, including improved detection of small lakes, reduced human bias, and better handling of cloud and shadow interference, as demonstrated in Section 4.2.

5.3. Transformer layers used in encoder

The ViT backbone consists of 12 transformer layers, each capturing different levels of representation, from local to global contexts. It is common to select layers of varying depths to leverage information at multiple levels (Si et al., 2022). In our model, we selected layers 3, 6, 9, and 12 to enrich feature diversity and complexity, following a structure similar to UNETR.

To evaluate whether these layer choices are optimal for segmentation, we conducted an ablation experiment, testing each of the 12 transformer layers in the encoder individually. This reduced the U-structure to a single layer for the encoder–decoder, while retaining the top-level skip connection with the IBB block. The experiment was performed on the WGL dataset using both the Prithvi and ViT-MAE-Base backbones, and the results are shown in Fig. 10.

For both backbones, deeper layers (9–12) consistently achieved higher F1 scores compared to shallower layers. However, the first six layers displayed different trends between the two backbones: Prithvi's layers 1–6 showed no significant variations in F1 score, while ViT-MAE-Base demonstrated a clear increase in F1 score from layers 1 to 5. This comparison highlights the complexity of how intermediate transformer layers interact with input data to generate feature embeddings. Determining the optimal layer selection requires a comprehensive study that rigorously evaluates the impact of different layers on performance, which is beyond the scope of this study. Additionally, there is a risk of overfitting when fine-tuning layer selection, which must be carefully managed.

5.4. Limitations and outlook

In this study, the proposed model was evaluated for glacial lake mapping using the GFM, providing a comprehensive assessment of its strengths and limitations. However, further improvements can enhance the model's performance and generalizability.

Firstly, in our comparisons, we only tested against CNN-based models since they are commonly used for glacial lake extraction. We did not compare against ViT-based models, such as Segformer (Xie et al., 2021) and Mask2Former (Cheng et al., 2022), or traditional methods like random forest and object-based image analysis (Morgan et al., 2024; Wangchuk and Bolch, 2020). Future work could incorporate these models for a more comprehensive evaluation, potentially improving and benchmarking against our proposed approach.

Secondly, although multi-sensor data offers clear advantages, the NPK and WGL datasets have constraints. The NPK dataset, for instance, primarily covers the summer season, leading to a lack of winter data, which reduces the model's effectiveness in detecting glacial lakes during winter. This seasonal bias limits the model's generalizability, as it struggles to recognize seasonal variations in glacial lakes caused by snow and ice cover. Expanding the datasets to include winter imagery could improve the model's robustness year-round.

Thirdly, analysis shows that the SAR VV polarization channel is less effective compared to the NIR, G, and B bands. This limitation is especially evident when cloud cover obscures multispectral imagery, rendering the SAR data insufficient for accurate lake extraction. Incorporating additional polarizations and exploring other spectral bands could provide richer information, enhancing the model's performance under such conditions.

Fourthly, the 100 m^2 minimum lake area threshold is chosen to ensure the detection of very small glacial lakes, some spanning just a few meters. While this approach enhances mapping sensitivity, it also introduces greater uncertainty and a higher potential for misclassification. To address this, we integrated multi-sensor datasets into our model, which impose stricter criteria on what qualifies as a water body through cross-referencing optical and SAR data. Nevertheless, additional post-processing steps could further reduce noise and improve the accuracy of the glacial lake inventories, such as temporal consistency checks, morphological filtering, connectivity analysis, topographic validation using ancillary data, and expert verification through manual or semi-automated methods (Shugar et al., 2020; Mustafa et al., 2024; Nie et al., 2017).

Finally, the absence of long-term time-series analysis limits our understanding of glacial lake dynamics and their impact on glacier

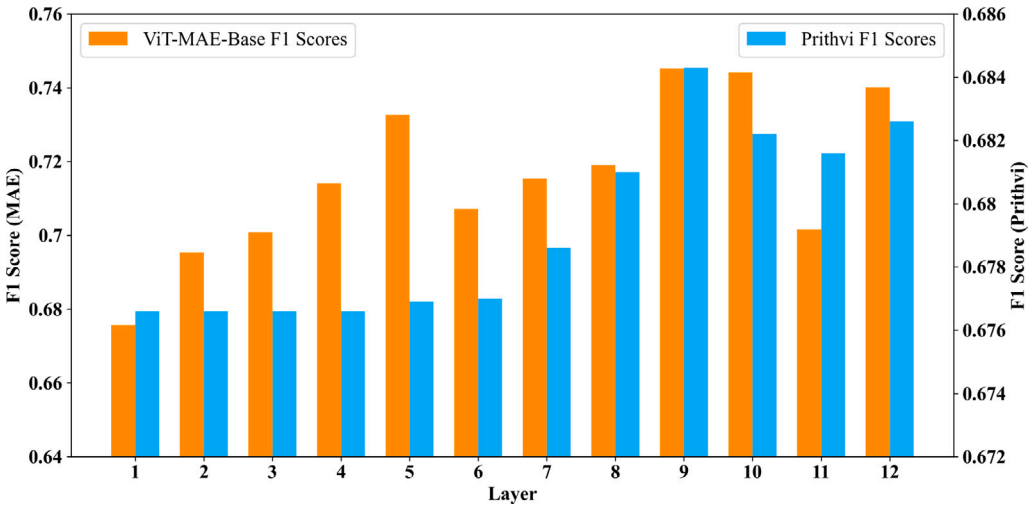

Fig. 10. Comparison of each transformer layer importance in different pre-trained models.F1 scores are obtained by inputting only intermediate layers 1, 2, 3, 4, 5, 6, 7, 8, 9, and 12 into the encoder. The orange bars represent each layer pre-trained on ImageNet-1K, with corresponding F1 scores on the left axis. The blue bars represent each layer pre-trained on Prithvi, with corresponding F1 scores on the right axis. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

mass balance and sea-level rise. To address this, future research should aim to develop high-precision, short-interval glacial lake datasets using multi-source remote sensing data. Such datasets would enable long-term monitoring and a deeper analysis of glacial lake evolution, providing valuable insights into their role in glacier dynamics and climate change.

6. Conclusion

In this study, we introduced a hybrid U-ViT model for precise and efficient glacial lake mapping, combining a Vision Transformer (ViT)-based encoder with a CNN-based decoder within a U-shaped architecture. The Prithvi Geo-Foundation Model (GFM) served as the encoder's backbone, enhancing performance for geoscience tasks. By connecting intermediate encoder layers to the decoder via CNN-based skip connections, the model effectively integrates features at both local and global scales. This design presents a novel method for fine-tuning GFMs with multi-sensor data, demonstrating flexibility in input dimensions and strong overall performance.

We validated the model through comparative experiments and ablation studies using datasets from two distinct regions with varying data modalities. The U-ViT model achieved the highest accuracy on the NPK dataset, which aligns closely with Prithvi's pre-training data. However, on the WGL dataset, DeepLabV3+ outperformed the U-ViT model, highlighting the impact of input modality differences between polarimetric SAR and multispectral data. The ablation study confirmed that integrating the Enhanced Squeeze-Excitation (ESE) block and the Inverted Bottleneck Block (IBB) significantly improved segmentation accuracy, demonstrating the value of CNN-based local feature extraction. Our experiments with different spectral bands revealed that the NIR and Green bands were most influential in glacial lake detection, while the SAR VV band was complementary but impacted by interference from floating objects in lakes.

The model also demonstrated robustness when trained with smaller datasets, maintaining high performance even with a 50% reduction in training size, showcasing its potential for use in data-scarce scenarios. Furthermore, when compared with public datasets in Pakistan and Afghanistan, the U-ViT model excelled in detecting small lakes and refining boundaries, effectively mapping lakes under cloud cover, shadow, and challenging terrain.

Despite these advancements, limitations persist. Incorporating SAR data into the GFM's pre-training process could enhance fine-tuning for multi-sensor applications. Further investigation into model architectures and learning strategies is needed to optimize the model's capabilities. Future work should also focus on developing a comprehensive benchmark dataset for glacial lake mapping that encompasses diverse regions and temporal spans, facilitating systematic evaluations and the creation of more generalized models.

In conclusion, the U-ViT model provides an effective approach for leveraging GFMs in large-scale glacial lake mapping, and the Prithvi GFM shows considerable potential for advancing this field. This study contributes new insights into the use of advanced deep learning techniques for automatic glacial lake mapping and establishes a foundation for broader applications of ViTs and GFMs in cryosphere research.

CRediT authorship contribution statement

Di Jiang: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Data curation, Conceptualization. **Shiyi Li:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Irena Hajnsek:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Muhammad Adnan Siddique:** Writing – review & editing, Data curation. **Wen Hong:** Supervision, Resources, Project administration, Formal analysis. **Yirong Wu:** Supervision, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The author Jiang Di is jointly supported by the Chinese Scholarship Council and National Natural Science Foundation of China, Grant No. 42276252 as visiting Ph.D. student in ETH Zurich. The authors also acknowledge IBM and NASA for providing the Prithvi Geospatial Foundation Model to support this work.

Data availability

The data supporting this study is available upon request. The code for this study is accessible at the following link: https://gitlab.ethz.ch/dijiang/transformer.

References

Arthur, J.F., Stokes, C., Jamieson, S.S., Carr, J.R., Leeson, A.A., 2020. Recent understanding of Antarctic supraglacial lakes using satellite remote sensing. Prog. Phys. Geography: Earth Environ. 44 (6), 837–869.

Chen, F., Zhang, M., Guo, H., Allen, S., Kargel, J.S., Haritashya, U.K., Watson, C.S., 2021. Annual 30 m dataset for glacial lakes in High Mountain Asia from 2008 to 2017. Earth Syst. Sci. Data 13 (2), 741–766. http://dx.doi.org/10.5194/essd-13-741-2021, URL https://essd.copernicus.org/articles/13/741/2021/.

Chen, L., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. CoRR arXiv:1802.02611, URL http://arxiv.org/abs/1802.02611.

Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R., 2022. Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1290–1299.

Chicco, D., 2017. Ten quick tips for machine learning in computational biology. BioData Min. 10 (1), 35.

Dirscherl, M., Dietz, A.J., Kneisel, C., Kuenzer, C., 2021. A novel method for automated supraglacial lake mapping in Antarctica using Sentinel-1 SAR imagery and deep learning. Remote Sens. 13 (2), 197.

Dou, X., Fan, X., Wang, X., Yunus, A.P., Xiong, J., Tang, R., Lovati, M., van Westen, C., Xu, Q., 2023. Spatio-temporal evolution of glacial lakes in the Tibetan Plateau over the past 30 years. Remote Sens. 15 (2), 416.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., et al., 2012. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. Remote Sens. Environ. 120, 25–36.

Fan, T., Wang, G., Li, Y., Wang, H., 2020. MA-net: A multi-scale attention network for liver and tumor segmentation. IEEE Access 8, 179656–179665. http://dx.doi.org/10.1109/ACCESS.2020.3025372.

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D., 2022. Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 574–584.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2022. Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Hong, D., Zhang, B., Li, X., Li, Y., Li, C., Yao, J., Yokoya, N., Li, H., Ghamisi, P., Jia, X., Plaza, A., Gamba, P., Benediktsson, J.A., Chanussot, J., 2024. SpectralGPT: Spectral remote sensing foundation model. IEEE Trans. Pattern Anal. Mach. Intell. 46 (8), 5227–5244. http://dx.doi.org/10.1109/TPAMI.2024.3362475.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7132–7141.

Hugonnet, R., McNabb, R., Berthier, E., Menounos, B., Nuth, C., Girod, L., Farinotti, D., Huss, M., Dussaillant, I., Brun, F., et al., 2021. Accelerated global glacier mass loss in the early twenty-first century. Nature 592 (7856), 726–731.

Jakubik, J., Roy, S., Phillips, C.E., Fraccaro, P., Godwin, D., Zadrozny, B., Szwarcman, D., Gomes, C., Nyirjesy, G., Edwards, B., Kimura, D., Simumba, N., Chu, L., Mukkavilli, S.K., Lambhate, D., Das, K., Bangalore, R., Oliveira, D., Muszynski, M., Ankur, K., Ramasubramanian, M., Gurung, I., Khallaghi, S., Hanxi, Li, Cecil, M., Ahmadi, M., Kordi, F., Alemohammad, H., Maskey, M., Ganti, R., Weldemariam, K., Ramachandran, R., 2023. Foundation models for generalist geospatial artificial intelligence. arXiv:2310.18660, URL https://arxiv.org/abs/2310.18660.

Jiang, D., Li, X., Zhang, K., Marinsek, S., Hong, W., Wu, Y., 2022. Automatic supraglacial lake extraction in greenland using sentinel-1 SAR images and attention-based U-net. Remote Sens. 14 (19), 4998.

Kaushik, S., Singh, T., Joshi, P.K., Dietz, A.J., 2022. Automated mapping of glacial lakes using multisource remote sensing data and deep convolutional neural network. Int. J. Appl. Earth Obs. Geoinf. 115, 103085.

Ke, L., Ding, X., Ning, Y., Liao, Y., Song, C., 2024. Annual trajectory of global glacial lake variations and the interactions with glacier mass balance during 2013–2022. CATENA 245, 108280. http://dx.doi.org/10.1016/j.catena.2024.108280.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., et al., 2023. Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026.

Lee, H., Calvin, K., Dasgupta, D., Krinner, G., Mukherji, A., Thorne, P., Trisos, C., Romero, J., Aldunce, P., Barret, K., et al., 2023. In: Lee, H., Romero, J. (Eds.), IPCC, 2023: Climate Change 2023: Synthesis Report, Summary for Policymakers. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change Core Writing Team. IPCC, Geneva, Switzerland.

Li, Y., Xie, S., Chen, X., Dollar, P., He, K., Girshick, R., 2021. Benchmarking detection transfer learning with vision transformers. arXiv preprint arXiv:2111.11429.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2980–2988.

Loshchilov, I., 2017. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

Mascolo, L., Cloude, S.R., Lopez-Sanchez, J.M., 2022. Model-based decomposition of dual-pol SAR data: Application to sentinel-1. IEEE Trans. Geosci. Remote Sens. 60, 1–19. http://dx.doi.org/10.1109/TGRS.2021.3137588.

Morgan, T., McNabb, R., Dunlop, P., 2024. Monitoring the changes in glacial lakes in the Southern Alps, New Zealand from 2000–2023 using an Object-Based Image Analysis (OBIA) approach in Google Earth Engine (GEE). Technical Report, Copernicus Meetings.

Mustafa, H., Tariq, A., Shu, H., ul Hassan, S.N., Khan, G., Brian, J.D., Almutairi, K.F., Soufan, W., 2024. Integrating multisource data and machine learning for supraglacial lake detection: Implications for environmental management and sustainable development goals in high mountainous regions. J. Environ. Manag. 370, 122490.

Nie, Y., Sheng, Y., Liu, Q., Liu, L., Liu, S., Zhang, Y., Song, C., 2017. A regional-scale assessment of Himalayan glacial lake changes using satellite observations from 1990 to 2015. Remote Sens. Environ. 189, 1–13.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. CoRR arXiv:1505.04597, URL http://arxiv.org/abs/1505.04597.

Schröder, L., Neckel, N., Zindler, R., Humbert, A., 2020. Perennial supraglacial lakes in Northeast Greenland observed by polarimetric SAR. Remote Sens. 12 (17), 2798.

Shugar, D.H., Burr, A., Haritashya, U.K., Kargel, J.S., Watson, C.S., Kennedy, M.C., Bevington, A.R., Betts, R.A., Harrison, S., Strattman, K., 2020. Rapid worldwide growth of glacial lakes since 1990. Nature Clim. Change 10 (10), 939–945.

Si, C., Yu, W., Zhou, P., Zhou, Y., Wang, X., Yan, S., 2022. Inception transformer. Adv. Neural Inf. Process. Syst. 35, 23495–23509.

Stevens, L.A., Nettles, M., Davis, J.L., Creyts, T.T., Kingslake, J., Hewitt, I.J., Stubblefield, A., 2022. Tidewater-glacier response to supraglacial lake drainage. Nat. Commun. 13 (1), 6065.

Taylor, C., Robinson, T.R., Dunning, S., Rachel Carr, J., Westoby, M., 2023. Glacial lake outburst floods threaten millions globally. Nat. Commun. 14 (1), 487.

Torres, R., Snoeij, P., Geudtner, D., Bibby, D., Davidson, M., Attema, E., Potin, P., Rommen, B., Floury, N., Brown, M., et al., 2012. GMES sentinel-1 mission. Remote Sens. Environ. 120, 9–24.

Turton, J.V., Hochreuther, P., Reimann, N., Blau, M.T., 2021. The distribution and evolution of supraglacial lakes on 79°N Glacier (north-eastern Greenland) and interannual climatic controls. Cryosphere 15 (8), 3877–3896. http://dx.doi.org/10.5194/tc-15-3877-2021.

Veh, G., Lützow, N., Tamm, J., Luna, L.V., Hugonnet, R., Vogel, K., Geertsema, M., Clague, J.J., Korup, O., 2023. Less extreme and earlier outbursts of ice-dammed lakes since 1900. Nature 614 (7949), 701–707.

Walther, S., Cseres, L., Marquis, R., Chapuis, B., Perez-Uribe, A., 2023. Diving into supraglacial lakes detection: a deep semantic segmentation approach. In: Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems. pp. 1–4.

Wang, X., Guo, X., Yang, C., Liu, Q., Wei, J., Zhang, Y., Liu, S., Zhang, Y., Jiang, Z., Tang, Z., 2020. Glacial lake inventory of high-mountain Asia in 1990 and 2018 derived from Landsat images. Earth Syst. Sci. Data 12 (3), 2169–2182.

Wang, Y., Sugiyama, S., 2024. Supraglacial lake evolution on Tracy and Heilprin Glaciers in northwestern Greenland from 2014 to 2021. Remote Sens. Environ. 303, 114006.

Wangchuk, S., Bolch, T., 2020. Mapping of glacial lakes using Sentinel-1 and Sentinel-2 data and a random forest classifier: Strengths and challenges. Sci. Remote. Sens. 2, 100008.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P., 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. Adv. Neural Inf. Process. Syst. 34, 12077–12090.

Xu, X., Liu, L., Huang, L., Hu, Y., 2024. Combined use of multi-source satellite imagery and deep learning for automated mapping of glacial lakes in the Bhutan Himalaya. Sci. Remote. Sens. 10, 100157. http://dx.doi.org/10.1016/j.srs.2024.100157, URL https://www.sciencedirect.com/science/article/pii/S2666017224000415.

Zhang, M., Chen, F., Guo, H., Yi, L., Zeng, J., Li, B., 2022. Glacial lake area changes in high mountain asia during 1990–2020 using satellite remote sensing. Research.

Zhang, R., Hu, X., Li, B., Huang, S., Deng, H., Qiao, Y., Gao, P., Li, H., 2023a. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15211–15222.

Zhang, Q., Liu, Y., 2017. Overview of Chinese first C band multi-polarization SAR satellite GF-3. Aerosp. China 3, 22–31.

Zhang, T., Wang, W., An, B., Wei, L., 2023b. Enhanced glacial lake activity threatens numerous communities and infrastructure in the third pole. Nat. Commun. 14 (1), 8250.

Zheng, L., Li, L., Chen, Z., He, Y., Mo, L., Chen, D., Hu, Q., Wang, L., Liang, Q., Cheng, X., 2023. Multi-sensor imaging of winter buried lakes in the Greenland Ice Sheet. Remote Sens. Environ. 295, 113688.