ISPRS Journal of Photogrammetry and Remote Sensing 202 (2023) 303–313

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

Automated glacier extraction using a Transformer based deep learning approach from multi-sensor remote sensing imagery

Yanfei Peng [a], Jiang He [a], Qiangqiang Yuan [a,b,*], Shouxing Wang [a], Xinde Chu [c], Liangpei Zhang [d]

[a] School of Geodesy and Geomatics, Wuhan University, Wuhan, Hubei 430079, China
[b] Hubei Luojia Laboratory, Wuhan, Hubei 430079, China
[c] College of Geography and Environmental Science, Northwest Normal University, Lanzhou 730070, China
[d] State Key Laboratory of Information Engineering, Survey Mapping and Remote Sensing, Wuhan University, Wuhan, Hubei 430079, China

ARTICLE INFO

Keywords:
Glacier
Deep learning
Transformer
Multi-source data
Qilian mountain

ABSTRACT

Glaciers serve as sensitive indicators of climate change, making accurate glacier boundary delineation crucial for understanding their response to environmental and local factors. However, traditional semi-automatic remote sensing methods for glacier extraction lack precision and fail to fully leverage multi-source data. In this study, we propose a Transformer-based deep learning approach to address these limitations. Our method employs a U-Net architecture with a Local-Global Transformer (LGT) encoder and multiple Local-Global CNN Blocks (LGCB) in the decoder. The model design aims to integrate both global and local information. Training data for the model were generated using Sentinel-1 Synthetic Aperture Radar (SAR) data, Sentinel-2 multispectral data, High Mountain Asia (HMA) Digital Elevation Model (DEM), and Shuttle Radar Topography Mission(SRTM) DEM. The ground truth was obtained for a glaciated area of 1498.06 km^2 in the Qilian mountains using classic band ratio and manual delineation based on 2 m resolution GaoFen (GF) imagery. A series of experiments including the comparison between different models, model modules and data combinations were conducted to evaluate the model accuracy. The best overall accuracy achieved was 0.972. Additionally, our findings highlight the significant contribution of Sentinel-2 data to glacier extraction.

1. Introduction

Glaciers are widely recognized as critical indicators of climate change, given their remarkable sensitivity to even minor climatic fluctuations (Lemke et al., 2007). Consequently, various regions across the globe are expected to experience shifts in water availability, both on a seasonal and long-term basis (Huss and Hock, 2018; Milillo et al., 2022), with arid and semi-arid environments particularly vulnerable to these changes (Immerzeel et al., 2010; Pritchard, 2019; Bhattacharya et al., 2021). Accurately delineating glacier boundaries plays a vital role in assessing the extent of glacier areas (Racoviteanu et al., 2015) and comprehending the sensitivity of glacier change to environmental and local factors (Catania et al., 2018; Sun et al., 2019), which makes fast and accurate method of wide-range glacier extraction becomes very necessary (Paul et al., 2020).

Remote sensing technology plays a crucial role in glacier extraction due to the remote locations and large spatial extents of glaciers. The choice of data source for glacier extraction is based on the distinct properties of different land covers on the glacier surface, and can be categorized into optical, Synthetic Aperture Radar (SAR), and multi-source datasets. Optical imagery serves as the fundamental data source for glacier extraction. The principle is based on the very low spectral reflectance of ice and snow in the shortwave infrared versus the high reflectance in the visible spectrum (Paul et al., 2015). Specifically, the approaches mainly include the thresholding method and the index-based method (Bolch et al., 2010). According to the Spatial-temporal resolution, the data source contains Landsat, Sentinel-2, Hexagon KH-9, and Pléiades et al. (Bolch et al., 2010; Holzer et al., 2015). However, the effectiveness of optical imagery is limited by weather conditions and challenges in distinguishing debris-covered glaciers from surrounding bedrock due to similar spectral characteristics. To overcome these limitations, SAR data has been employed in glacier extraction (Frey et al., 2012; Malenovský et al., 2012; Zhou and Zheng, 2017). SAR-based methods rely on two key theories. The first theory revolves around the low coherence exhibited by glaciers(both clean and debris-covered) due to their dynamic nature compared to the higher coherence of

* Corresponding author at: School of Geodesy and Geomatics, Wuhan University, Wuhan, Hubei 430079, China.
E-mail address: qqyuan@sgg.whu.edu.cn (Q. Yuan).

https://doi.org/10.1016/j.isprsjprs.2023.06.015
Received 16 February 2023; Received in revised form 28 May 2023; Accepted 28 June 2023
Available online 2 July 2023

adjacent bedrock. Common data sources for this theory include Sentinel-1 and Advanced Land Observing Satellite Phased Array Synthetic Aperture Radar (ALOS PALSAR) (Frey et al., 2012; Malenovský et al., 2012; Zhou and Zheng, 2017). However, the processing of SAR coherence is intricate and constrained by non-steady deformation processes (Rosen et al., 2000). The second theory is based on the distinctive backscattering properties of glaciers and other land cover types. Sentinel-1 Ground Range Detected (GRD) data is widely employed within this framework (Peng et al., 2021; Wang et al., 2022c). Following the diversification characteristics of glaciers, the integration of multi-source data has emerged as a mainstream approach (Frey et al., 2012; Robson et al., 2015). Multi-source data typically includes optical imagery, SAR data, and Digital Elevation Models (DEM). DEMs are useful for glacier extraction due to the elevation differences between glaciers and adjacent bedrock (Bolch et al., 2010). Object-Based Image Analysis (OBIA) is a commonly employed method in this context, wherein specific surface features are captured using algorithms tailored to the respective data types (Blaschke, 2010). However, most existing approaches for combining multi-source data focus on extracting distinct glacier parameters from different datasets (Holzer et al., 2015; Zhang et al., 2020; Zhao et al., 2020). Consequently, a more efficient approach for integrating multi-source data is necessary (Paul et al., 2013).

However, the aforementioned methods rely on differentiating glaciers from surrounding land objects based on various characteristics such as spectral differences, coherence differences resulting from glacier changes, and elevation changes, among others. This approach to glacier identification gives rise to two challenges. Firstly, it struggles to accurately identify land features with subtle differences, such as glacier debris. Secondly, it requires the setting of different thresholds for glaciers with different characteristics. These limitations highlight the inherent inaccuracies and result variations in traditional methods due to differences in datasets and glacier geometries. Deep learning, on the other hand, offers a promising alternative by leveraging Convolutional Neural Networks (CNNs) and multi-layer learning to extract spatial information of glacier outlines from remote sensing images (LeCun et al., 2015; Xiao et al., 2021). Deep learning frameworks have provided effective solutions for visual object extraction tasks in remote sensing, including land cover mapping and the extraction of natural features such as lakes, as well as urban features such as buildings and roads, achieving state-of-the-art reconstruction results (Yuan et al., 2020).

In the field of glacier extraction, deep learning technology has found applications. One of the early models is U-Net (Ronneberger et al., 2015), which utilizes skip connections and a decoder with multi-upsampling layers to capture both low and high-level information. U-Net has been employed to segment ice and ocean in individual glaciers in Greenland and Antarctica (Baumhoer et al., 2019; Mohajerani et al., 2019; Zhang et al., 2019). Another notable model is DeepLab V3+ (Chen et al., 2018), which incorporates the Atrous Spatial Pyramid Pooling (ASPP) module to expand the receptive field. This model enhancement has been applied to identify extensive and long-term glaciers, as demonstrated in the construction of the glacier termini dataset for multiple glaciers in Greenland (Cheng et al., 2021; Zhang et al., 2021). However, CNN-based models fail to consider the varying importance of image features. This deficiency has been effectively addressed by the attention mechanism, which dynamically weighs the significance of features based on their relationships and interactions within the input data (Ba et al., 2014; Wang et al., 2018; Woo et al., 2018). Chu et al. (Chu et al., 2022) have integrated the Convolutional Block Attention Module (CBAM) with the ASPP module of DeepLab V3 + to identify glacier within complex mountainous environments, achieving state-of-the-art results. To overcome this limitation, the Vision Transformer (ViT) (Dosovitskiy et al., 2020) has been employed to extract more cohesive glacier outlines across larger areas. In the context of image segmentation based on remote sensing data, most Transformer-based models adopt a hybrid design of CNNs and Transformers to capture both local and global information (Strudel et al., 2021; Wang et al., 2022a; Zhang et al., 2022; Xiao et al., 2022). However, these combinations typically use a Transformer as the encoder and a CNN or Transformer as the decoder. This architecture posing challenges in simultaneously capturing spatial and semantic information for decoder. Regarding the encoder, the high computational cost of Transformers has spurred research on efficiently capturing global information, including several approaches that combine local and global Transformers using shifted and sub-window operations (Chu et al., 2021; Liu et al., 2021).

To overcome these limitations, this paper aims at developing a novel algorithm that can obtained the local–global information of glacier feature based on the fusion of Transformer and CNN. We introduce the Transformer based model to extract local–global information of glacier outline, which is an early attempt of Transformer application in the glacier extraction. Unlike the standard Transformer with the pure CNN and Transformer as the decoder, we designed a decoder with an attention layer and CNN to extract 1D and 2D feature information. In the meantime, the locally-grouped Transformer and a global sub-sampled Transformer were employed as encoder to capture the multi-scale long-distance information. Besides, in order to capture the spatial location information of the input sequence, we adopted Conditional Position Encoding(CPE) instead of the original position encoding of ViT so that the token's length can vary with the input length instead of fixed. A series of experiments are performed to validate the applicability of the proposed method with the different combinations of Sentinel-1, Sentinel-2, three land indexes and elevation data. The quantitative evaluation and visual effects of the proposed method are verified through different models and ablation study.

The rest of this paper is structured as follows. Section 2 introduces the methodology including the training dataset, deep learning model construction and their combined process. Section 3 presents experiment results and discussion. The experiments result contains assessments between different models, ablation study and contribution between different data combinations. The discussion is focus on the model's response to the heterogeneous data and model efficiency. Finally, Section 4 provides the conclusions.

2. Methodology

In this study, we present a Transformer based model for glacier area extraction. The model consists of an encoder based on Local-Global Transformer (LGT) and a decoder composed of several Local-Global CNN Blocks (LGCB). Four types of remote sensing data are utilized as the training dataset. The ground truth for evaluation is the glaciated area covering 1498.06 km^2 in the Qilian mountain region. The following section will provide a detailed introduction of datasets, construction of the model, and training process.

2.1. Data source

2.1.1. Remote sensing datasets

This work utilized four types of datasets, namely SAR (Sentinel-1), optical (Sentinel-2), image band indices, and DEM. SAR and optical data have been widely used in glacier extraction due to their complementary nature in capturing land information. In addition, the normalized-difference snow index (NDSI) (Salomonson and Appel, 2004), normalized difference water index (NDWI) (Xu, 2006) and normalized difference vegetation index (NDVI) were jointly employed to distinguish the glacier from the surrounding glacier lake, streams, and sparse vegetation (Wang et al., 2022c; Zhang et al., 2021). The DEM played a crucial role in identifying mountain glaciers from surrounding ridges and distinguishing debris-covered glacier termini from valleys. Specifically, the datasets were divided into four parts, including: (1) two Sentinel-1 GRD bands of "VV" and "VH", (2) five Sentinel-2 SR bands of "B2", "B3", "B4", "B8" and "B11", (3) three indexes bands conduct from sentinel-2 bands, (4) one elevation band from High Mountain Asia (HMA) or SRTM DEM. All bands were resampled to a 10 m resolution. The details

of each band (Table 1) are as follows.

Sentinel-1 acquires SAR imagery at the C-band (5.405 GHz) with varying polarizations and resolutions. In this work, the Sentinel-1 Level-1 GRD product is utilized, which consists of the vertical transmit/horizontal receive bands "VV" and "VH" (Torres et al., 2012). The resolution is 10 m. The pre-processed data for the 2018 summer was accessed from Google Earth Engine (GEE) platform.

Sentinel-2 is a wide-swath, high-resolution, multi-spectral imaging mission(Drusch et al., 2012). We utilized the Level-2 Bottom-Of-Atmosphere (BOA) corrected reflectance product, specifically incorporating bands B2 (496.6 nm), B3 (560 nm), B4 (664.5 nm), B8 (835.1 nm), and B11 (1613.7 nm). These bands are available as the Sentinel-2 Surface Reflectance (SR) product on the GEE platform. Expect B11(resolution: 20 m), the resolution of all the other bands is 10 m. We acquired the summer 2018 product with a cloud cover percentage of less than 5%.

NDSI, NDWI and NDVI were calculated through the Sentinel-2 bands mentioned in the previous paragraph with the formulation (1) (2) and (3) respectively.

$$NDVI = \frac{B4 - B8}{B4 + B8} \tag{1}$$

$$NDSI = \frac{B3 - B11}{B3 + B11} \tag{2}$$

$$NDWI = \frac{B3 - B8}{B3 + B8} \tag{3}$$

The HMA DEM product (Shean, 2017) consists of 8-meter DEM mosaics of glacier and snow regions in the high mountain Asia (HMA) area, generated from very-high-resolution (VHR) commercial optical satellite imagery, including QuickBird (Toutin and Cheng, 2002; Shean et al., 2016). The dataset was obtained from the National Snow and Ice Data Center(NSIDC)(https://nsidc.org/data/hma_dem8m_mos/versions/1). However, HMA DEM can't cover the whole study region. To address this, we used the SRTM DEM as a complementary dataset for an area of approximately 3978 km2. The SRTM V3 product provided by GEE platform was utilized (Farr et al., 2007).

2.1.2. Ground truth

The training label used in this study is glacier area at the Qilian mountains in 2018 (Fig. 1). These outlines were obtained through a combination of classical band ratio criterion and manual mediation, based on the 2 m resolution Gaofen(GF) imagery (Li, 2022).

The Qilian Mountains (39.7° – 37.3° N, 93.4° – 102.8° E) encompass various types of glaciers, including continental glaciers in the central and eastern regions and polar glaciers in the western region (Shi and Liu, 2000). Analysis of data from the First Chinese Glacier Inventory (FCGI) and Second Chinese Glacier Inventory (SCGI) revealed an area reduction of 420.81 km^2 in the Qilian Mountains between 1956 and 2010 (Sun et al., 2018).

According to the dataset, the Qilian Mountains in 2018 were home to a total of 2,740 glaciers covering an area of 1,514.01 km^2. Among these glaciers, the glaciers with areas 1—–10 km^2 accounted for the largest glacierized area (832.52 km^2). In our study, we adopt the 2072 glaciers (area $>$ 0.05 km^2) with the total area of 1498.06 km^2.

Table 1
Remote sensing data source used in this study.

Data source	Band	Resolution(m)	Date
Sentinel-1 GRD	VV, VH	10	2018.7
Sentinel-2 SR	B2, B3, B4, B8, B11	10	2018.7–9
Indices	NDVI, NDWI, NDSI	10	2018.7–9
HMA DEM	elevation	8	2017

2.2. Data pre-processing

Data pre-processing was performed using GEE platform. The pre-processing steps involved:(1) filtering of Sentinel-1 GRD and Sentinel-2 SR image; (2) calculation of image band indices; (3) normalization of each band; (4) integration of 11 bands. The ground truth data is generated by converting the vector shapefiles into binary raster images. To enhance the training dataset, data augmentation techniques were applied, including rotation by 90°, 180°, and 270°, as well as image flipping and mirroring. The flip operation vertically reorients the raster by flipping it from top to bottom along the horizontal axis through the center, while the mirror operation horizontally flips the raster from left to right along the vertical axis through the center. For training and validation, we utilized 70% and 30% of the data, respectively.

After extracting the glacier area, the obtained results can be further divided into individual glaciers through a three-step process. Firstly, the patches are merged by averaging the overlapping areas to mitigate classification errors and ensure smooth junctions. Secondly, the binary image is converted into a shapefile format. Finally, small and isolated polygons are removed to obtain the final outline dataset.

2.3. Model construction

In this work, we utilize a U-shaped (Cao et al., 2021) Encoder-Decoder network that incorporates a combination of local and global information (Fig. 2). The encoder is implemented using the Local-Global Transformer (LGT), while the decoder consists of multiple Local-Global CNN Blocks (LGCB). The loss function combines dice and cross-entropy loss. Detailed descriptions of the encoder, decoder, and loss function are as follows.

2.3.1. Encoder: Local-Global Transformer(LGT)

LGT was adopted as the encoder to capture local–global information. Vision Transformer(ViT) enjoy great flexibility in modelling long-range dependencies and serve as the most basic component in the next-generation of visual recognition tasks (Chu et al., 2021). However, ViT faces challenges in terms of computational complexity when dealing with high-resolution images compared to natural language. Swin Transformer is present to improve this constraint by computed self-attention only within each spatially grouped non-overlapped sub-window with the shifted window(Liu et al., 2021). Although this approach significantly reduces complexity, it sacrifices inter-window connections and results in uneven window sizes, limiting the receptive field and computational convenience. Thus, in order to capture both fine-grained and short-distance as well as long-distance and global information, we incorporate the locally-grouped self-attention (LSA) and global sub-sampled attention (GSA) modules from Twins-SVT (Chu et al., 2021) in our framework.

The LGT model is structured hierarchically with four stages, each consisting of a Conditional Position Encoding (CPE) and Transformer block. Initially, image patches are flattened into 1D sequence using patch embedding and then input to the hierarchical stages. The hierarchical stages were defined by different patch size. The definition of patch size is crucial for computation and feature extraction. Unlike ViT's consistent patch size approach that leads to low resolution and a single-scale representation, Pyramid Vision Transformer (PVT) (Wang et al., 2021) addresses this issue by incorporating pyramid construction through stacked layers with hierarchical patch sizes. In this study, we adopt the same architecture, designing four stages with patch sizes of 4 $\times$ 4, 8 $\times$ 8, 16 $\times$ 16, and 32 $\times$ 32. within each stage, we include CPE and Transformer encoder. CPE is provide spatial position information to 1D sequence data. Position information is indispensable for the vision tasks, but self-attention operation in Transformers is permutation-invariant, which cannot leverage the order of the tokens in an input sequence. The previous position encoding contains absolute and relative ways with fixed or dynamics systems. While in this work, we employ CPE from the

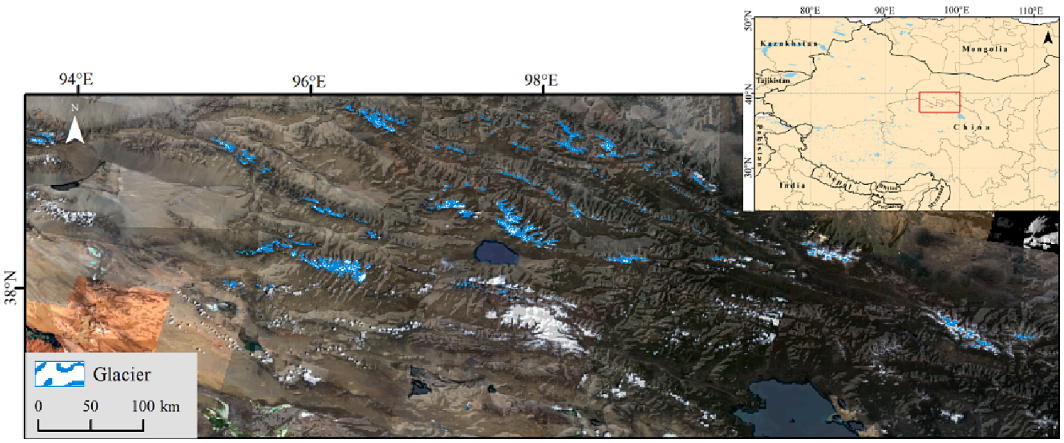

Fig. 1. The glacier outlines were used as ground truth during the model training. The background is Landsat 8 imagery and SRTM DEM.

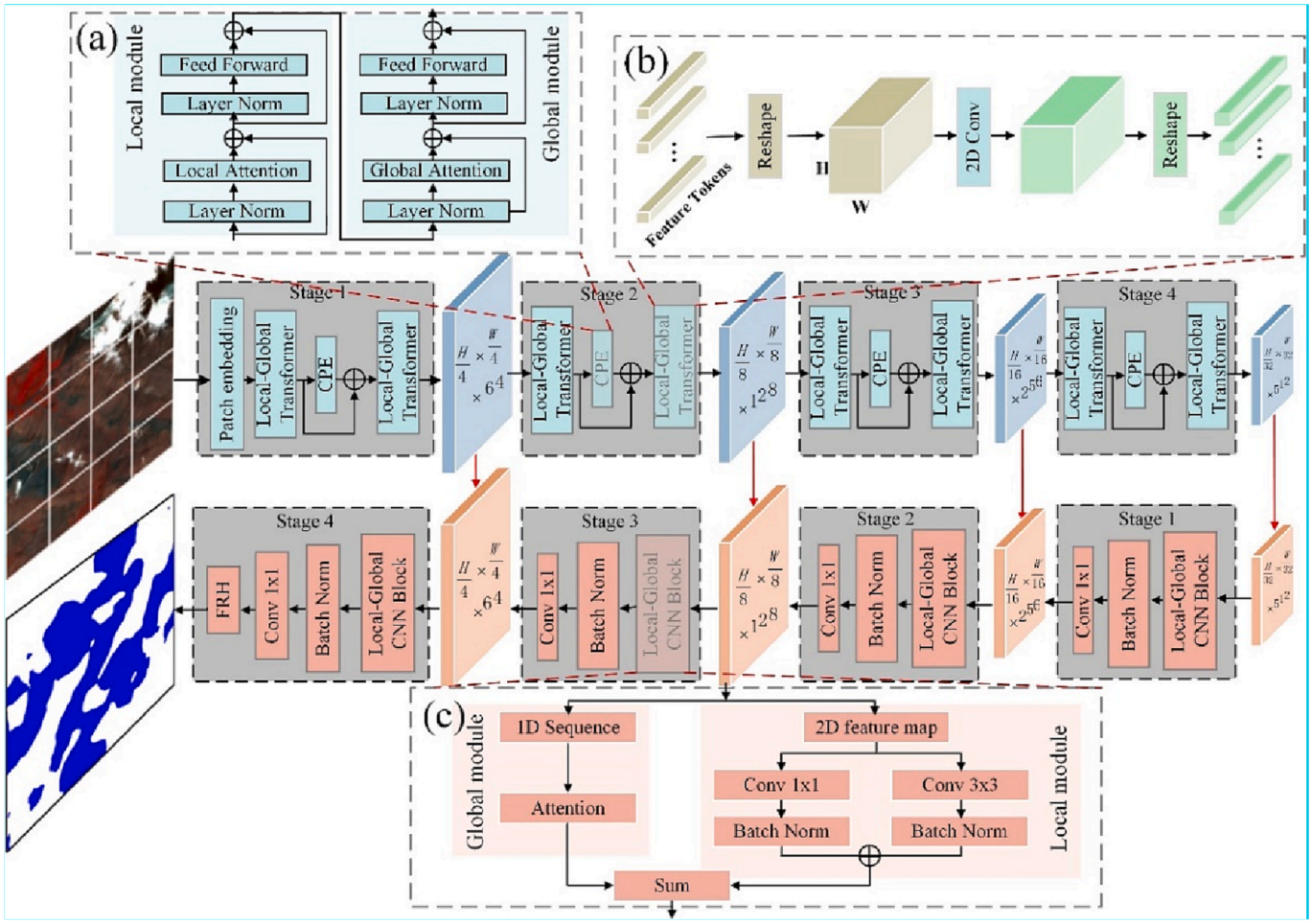

Fig. 2. The overall architecture of the proposed model. The first row of the middle part is the encoder, in which the feature map has been downsampling to size $\frac{H}{32} \times \frac{W}{32} \times 512$, the second row of the middle part is the decoder in which the feature map has been upsampling to the original size and segmented by FRH module. (a) and (b) is the detail of the LGT and CPE respectively. (c) is the detail of the LGCB.

Conditional Position encoding Vision Transformer(CPVT)(Chu et al., 2021b), which handles longer sequences in an inductive manner.. The CPE plugin is inserted in the middle of the Transformer encoder as a default choice (Fig. 2(b)). The Transformer encoder contains LSA and GSA(Fig. 2(a)).Each LSA is followed by a GSA. Residual connection(He et al., 2016) is applied around each of the two sub-layers, followed by layer normalization. The specific details of each block are provided below..

Patch embedding. In the initial stage, the input image $X \in \mathbb{R}^{H \times W \times C}$, where (H, W) represents the original image resolution and C denotes the number of channels, is split into $\frac{HW}{P^2}$ patches. Here, p refer to the patch size (which is set to 4 in the first stage of this study), Subsequently, the flattened patches are processed through projections, resulting in embedded patches of size $\frac{HW}{P^2} \times C1$. in the first stage of this work, $C1$ is set

to 64. Consequently, we obtain the patch embedding, which is then ready for further input into the Transformer operation.

Conditional Position Encoding (CPE). CPE(Fig. 2(b)) reshapes the flattened tokens $X \in \mathbb{R}^{B\times N\times C}$ back to 2D image representation $X^{'} \in \mathbb{R}^{B\times H\times W\times C}$, where B represents the batch size, N corresponds the number of sequence elements, (H, W) denotes the original size of the image, and C indicates the number of channels. Subsequently, the image undergoes a transformation using a function $\mathscr{F}$, resulting in conditional positional encodings $E^{B\times H\times W\times C}$ with the fixed size. The function $\mathscr{F}$ can be implemented as an efficient 2-D convolution with kernel $k(k \geq 3)$ and $\frac{k-1}{2}$ zero paddings. Finally, the 2D image is reshaped into flattened tokens, which serve as the input for the Transformer module.

Locally-grouped Self-Attention (LSA). To efficiently model the data, the Local Self-Attention (LSA) approach divides the 2D images $X \in \mathbb{R}^{H\times W\times C}$ into equally arranged m × n sub-windows(Fig. 2(a)). Within each sub-window, self-attention is calculated among the $\frac{HW}{mn}$ elements to enable communication limited to the sub-window. Self-attention involves mapping a query and a set of key-value pairs to an output, where the query, keys, values (Q, K, V) and output are all vectors. The output is computed as a weighted sum of the values, with the weights determined by a compatibility function of the query with the corresponding key (Vaswani et al., 2017). Specifically, the dot products of Q and K, which have the same dimension d_k, are computed and divided by the scaling factor $\sqrt{d_k}$. A Softmax operation is then applied to obtain the weighted sum by multiplying it with V. The calculations are performed by organizing Q, K, and V into matrices. This computation can be represented by Equation (4).

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

in which the Attention(Q, K, V) is the output metric. A single self-attention is only calculated at one dimension(d_{model}), which is less beneficial than computing self-attention in multi-shaped dimensions. Thus we present Multi-head attention, enabling effective communication among diverse information.. Multi-head attention can be present in equations (5) and (6):

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^O \tag{5}$$

$$head_i = Attention\left(QW_i^Q, kW_i^k, VW_i^V\right) \tag{6}$$

Where the corresponding dimension transformation parameter of (Q, K, V) are the matrices $W_i^Q \in \mathbb{R}^{d_{model}\times d_k}$, $W_i^K \in \mathbb{R}^{d_{model}\times d_k}$, $W_i^V \in \mathbb{R}^{d_{model}\times d_v}$ and $W^O \in \mathbb{R}^{d_{model}\times hd_v}$($h$ is the times of dimension transform). We can express the total cost of the LSA as $O\left(\frac{H^2W^2}{mn}d\right)$, which is highly dependent on the $\frac{H}{m}$ and $\frac{W}{n}$ (defined as k_1 and k_2). Obviously, the computation will more efficient when $k_1 \ll H$ and $k_2 \ll W$, which makes k_1 and k_2 are crucial parameters in this model.

Global Sub-sampled Attention (GSA). To facilitate communication between sub-windows, the global attention layer is incorporated (Fig. 2 (a)). In contrast to directly adding a standard global attention operation after the local layer we employed sub-sampled attention, inspired by the spatial reduction attention(SRA) of PVT, which reduces computation by reducing the spatial scale of two elements K, V of self-attention(Wang et al., 2021). Similarly, GSA employs a single representative to summarize the crucial information for each of the m × n sub-windows. This representative is used to interact with other sub-windows, serving as the K component in self-attention. Consequently, GSA effectively reduces the spatial scale of K in self-attention. The formulation of GSA is given by equations (7) and (8).

$$GSA(Q, K, V) = Concat(head_1, ..., head_n)W^O \tag{7}$$

$$head_j = Attention\left(QW_j^Q, SS(K)W_j^K, VW_j^V\right) \tag{8}$$

Where W_j^Q, W_j^K and W_j^V are still the metric of dimension project(same as the ones in standing Multi-head attention). $SS(K)$ means the Sub-Sampled K, which is related to the number of sub-windows. We reduce the cost of global communication from $O(H^2W^2d)$ to $O(mnHWd) = O\left(\frac{H^2W^2}{k_1k_2}d\right)$.

Finally, we can summarize Transformer Encoder as equation (9):

$$\begin{aligned}
\widehat{z}_{ij}^{\,l} &= \mathrm{LSA}\left(LayerNorm\left(z_{ij}^{l-1}\right)\right) + z_{ij}^{l-1},\\
z_{ij}^{l} &= FFN\left(LayerNorm\left(\widehat{z}_{ij}^{\,l}\right)\right) + \widehat{z}_{ij}^{\,l},\\
\widehat{z}^{\,l+1} &= GSA\left(LayerNorm\left(z^{l}\right)\right) + z^{l},\\
z^{l+1} &= FFN\left(LayerNorm\left(\widehat{z}^{\,l+1}\right)\right) + \widehat{z}^{\,l+1},\\
& i \in \{1, 2, ..., m\}, j \in \{1, 2, ..., n\}
\end{aligned} \tag{9}$$

where $\widehat{z}_{ij}^{\,l}$ and $\widehat{z}^{\,l+1}$ represent the output of LSA in sub-windows (i, j) and the GSA respectively, which refer to the feature maps of the input image. z_{ij}^{l} and z^{l+1} represent the results of Feed-Forward Networks(FFN) (Vaswani et al., 2017).

2.3.2. Decoder: Local-Global CNN Blocks (LGCB)

LGCB(Wang et al., 2022b) contains a local and global branch for feature extraction and a Feature Refinement Head (FRH) for segmentation. The local branch utilizes a CNN to extract spatial information, while the global branch incorporates an attention layer to capture 1D information. Unlike the standard Transformer that solely relies on CNN for decoding, our design incorporates both CNN and attention layers in the global branch. The following sections provide a detailed introduction to each block.

Local-Global CNN Blocks (LGCB). The local branch of the model employs two parallel convolutional layers with kernel sizes of 3 and 1, followed by batch normalization operations, to extract local contextual information.. The global branch involves dividing the 2D image into sub-windows and performing multi-head attention within each sub-window. To capture the global interaction between sub-windows, a cross-shaped window context interaction module is introduced, which combines the results of pooling layers. Specifically, the cross-shaped window modules utilize horizontal and vertical average pooling layers to calculate the pixel-wise dependencies across windows, representing the horizontal and vertical relationships. The details can refer to Unet-Former (Wang et al., 2022b).

Feature Refinement Head(FRH). To integrate the global–local and spatial-channel information in both the encoder and decoder, we employ the Feature Refinement Head (FRH) as the segmentation head. The FRH consists of spatial-wise and channel-wise representations. In the spatial branch, a depth-wise 3 × 3 convolution is used to generate a spatial-wise attentional map $S \in \mathbb{R}^{h\times w\times 1}$, where h and w represent the spatial resolution of the feature map. The channel branch starts with a global average pool to generate a channel-wise attentional map $C \in \mathbb{R}^{1\times 1\times c}$, where c represent the channel dimension. Subsequently, two 1 × 1 convolutional layers are applied: the first reduces the channel dimension c by a factor of 4, and the second expands it back to the original dimension. Finally, the two attentional maps are fused through a summation operation to facilitate semantic segmentation.

2.3.3. Loss function

We adopt the Dice loss function, which has been extensively utilized in semantic segmentation tasks (Sudre et al., 2017). In such tasks, the learning process often encounters difficulties in distinguishing the foreground region from the background, resulting in incomplete or missing detection of the foreground. The Dice loss addresses this limitation (Milletari et al., 2016). In this work, the dice loss l_d for the two class can be calculated as follow (10):

$$l_d = 1 - \frac{\sum_n^N p_n r_n + \varepsilon}{\sum_{n=1}^N p_n + r_n + \varepsilon} - \frac{\sum_{n=1}^N (1-p_n)(1-r_n) + \varepsilon}{\sum_{n=1}^N 2 - p_n - r_n + \varepsilon} \quad (10)$$

Where r_n and p_n represent the reference and prediction respectively, ϵ is a constant to avoid denominators of zero.

3. Experiment results and discussion

To evaluate the model performance, a series of experiments including the comparison of different models, model modules and test on different band combinations were set. Thus, this section contains the model implementation details, evaluation metrics and experiment results.

3.1. Implementation details and evaluation metrics

The dataset used in this study consisted of a total of 3024 image-label pairs, with each image and label resized to 512 × 512 pixels. All experiments are conducted on a single GeForce RTX 2080 Ti GPU. During the training process, we employed the AdamW optimizer with a learning rate of $1e^{-4}$ and utilized the cosine strategy to adjust the learning rate. The batch size was set to 4.

To evaluate the performance of the model,the overall accuracy(OA), mean F1 score(F1) and mean intersection over union(mIoU) were employed as the evaluation metric:

$$OA = \frac{tn + tp}{tn + tp + fn + fp} \quad (11)$$

$$F1 = 2\frac{PR}{P + R} \quad (12)$$

$$IoU = \frac{tp}{tp + fn + fp} \quad (13)$$

Where tp, fp, tn and fn denote the number of true positives, false positives, true negatives and false negatives respectively, $P = \frac{tp}{tp+fp}$ and $R = \frac{tp}{tp+fn}$.

3.2. Comparison between different models

To compare the capability of CNN, Attention mechanism and Transformer to capture local and global information in glacier extraction, the proposed model was evaluated against U-Net, DeepLab V3+, Attention DeepLab V3+, and Swin Transformer. U-Net has been previously employed for single glacier ice and ocean segmentation (Baumhoer et al., 2019; Mohajerani et al., 2019; Zhang et al., 2019). focusing on identifying the spectral and spatial heterogeneity of ice and ocean at the calving front of Greenland or Antarctic glaciers. In these work, the both low and high level local spatial information. was captured by symmetrical encoder-decoder architecture and CNN of U-Net. In contrast, DeepLab V3 + has been employed for delineating calving fronts of multiple Greenland glaciers using diverse datasets (Cheng et al., 2021; Zhang et al., 2021). DeepLab V3 + incorporates a decoder that combines low and high-level features from the encoder and utilizes Atrous Spatial Pyramid Pooling (ASPP) to handle multi-scale information and expand the receptive field. However, these methods primarily focus on extracting local information at the ice/ocean boundary. Attention DeepLab V3+, on the other hand, is designed to extract information from wide-range mountain glaciers (Chu et al., 2022). Chu et al. introduced the Convolutional Block Attention Module (CBAM) as a parallel connection to the ASPP of DeepLab V3 + to selectively emphasize different regions of importance. Notably, none of the above methods consider global information. Therefore, we introduce Transformer to glacier extraction. Vision Transformer (ViT) captures long-distance dependencies in image features through position embedding and multi-head attention. Swin Transformer further enhances ViT by employing hierarchical feature maps with different patch sizes in different stages and incorporating shifted windows attention within each Swin block, enabling it to handle large-scale images efficiently.

The accuracy and extraction results of the compared models are presented in Table 3 and Fig. 3, respectively. Among the compared models, U-Net exhibits the lowest performance, while DeepLab V3 + demonstrates a significant improvement of 27.4% in OA compared to U-Net. The glacier area extracted by U-Net shows fragmentation, as indicated by the red frame in Fig. 3(b) for the second and third rows. In contrast, DeepLab V3 + produces more contextually coherent results, reflecting the expanded receptive field achieved through atrous convolution. Furthermore, the inclusion of an attention layer further improves the mean accuracy by 0.017 for DeepLab V3 +. Fig. 3(d) illustrates that Attention DeepLab V3 + sporadically detects glacier patches missed by DeepLab V3 +. However, all the results of the above CNN-based models exhibit spatial discontinuities around the marginal regions of the glacier area. While the Transformer model enables a more comprehensive extraction of the glacier area and also further enhanced the mean accuracy of Attention Deeplab V3 + by 0.025, as depicted in the red frame of the first row in Fig. 3(e) and Table 2, indicating its capability to capture long-distance image information. Despite this, all the models still exhibit vague and coarse details of the glacier edge. In contrast, our proposed model achieves both a complete glacier area and clear edge, as illustrated in the red frame of the first row in Fig. 3(f). The evaluation results in Table 2 also confirm the superior performance of our proposed model.

3.3. Ablation study

To evaluate the performance of LSA and GSA, CPE and LGCB, we conduct a series of ablation experiments. The results are shown in Table 3.

Locally-grouped self-attention (LSA) and Global sub-sampled attention (GSA) We evaluate the performance of the different combinations of LSA and GSA in each stage of encoder. The experiments were performed using a U-Net architecture with a ResNet18 backbone. The results, as shown in Table 3, indicate that models employing pure Global or Local attention (LL, LL, LL, LL and GG, GG, GG, GG) performed poorly. This can be attributed to the limited capacity of capturing only local information and having a small receptive field. On the other hand, the interleaved combination of global and local attention (LL, LG, LG, GG) demonstrated relatively higher precision, while the global–local combination in each stage (LG, LG, LG, LG) yielded the best overall performance.

Local-Global CNN Block(LGCB). We conducted ablation experiments to evaluate the performance of the encoder using three different decoders: pure CNN, Transformer, and our proposed LGCB. These experiments were conducted using the U-Net architecture with the same encoder as LGT. The results clearly indicate that the CNN decoder achieved the lowest performance. While the Transformer decoder showed a slight improvement in precision, it came at the cost of significantly higher parameters. Therefore, considering both precision and efficiency, LGCB emerged as the best decoder in this experiment.

Conditional position encoder (CPE) The effectiveness of the CPE is proven by comparing the original position encoding (PE) in Vision Transformer (Dosovitskiy et al., 2020). The utilization of CPE instead of

Table 2
Experiment accuracy of different models.

	mIoU	F1	OA
U-Net	0.461	0.612	0.725
DeepLab V3+	0.673	0.816	0.924
Attention DeepLab V3+	0.702	0.822	0.960
Swin Transformer	0.718	0.829	0.962
Our proposed model	0.726	0.843	0.972

Table 3
Ablation study of LSA and GSA, LGCB and CPE. In the LSA and GSA part, L represents LSA and G represent GSA. LG means the specific stage contains one LSA and one GSA. Training time is the time for one batch.

Module	Method	mIoU	F1	OA	Parame-ters(M)	Training time (Mins)
LSA and GSA	(LG, LG, LG, LG) + ResNet18	0.624	0.704	0.921	40.1	6.0
	(LL, LG, LG, GG) + ResNet18	0.612	0.675	0.914	31.2	5.4
	(GG, GG, GG, GG) + ResNet18	0.607	0.642	0.832	25.6	5.8
	(LL, LL, LL, LL) + ResNet18	0.584	0.613	0.784	17.2	5.2
LGCB	LGT + ResNet18	0.624	0.704	0.921	40.1	6.0
	LGT + ResNet50	0.687	0.769	0.935	50.1	6.0
	LGT + LGT	0.730	0.845	0.972	156.2	7.5
	LGT + LGCB (Our proposed)	0.726	0.843	0.972	107.5	6.5
CPE	PE + LGT + LGCB	0.688	0.725	0.963	105.5	6.5

PE leads to an increase of 0.015 in the average accuracy of mIoU, F1, and OA.

3.4. Contribution of different band combinations

To investigate the effect of different band combinations, the input bands are divided into four different groups, namely A ('VV', 'VH' from Sentinel-1GRD), B ('B2′, 'B3′, 'B4′, 'B8′ and 'B11′ from Sentinel-2 SR), C ('ndvi', 'ndsi' and 'ndwi') and D ('elevation' from HMA DEM). In order to examine the effectiveness of various data sources, we excluded the combination of group B and C since group C is derived from the bands in group B. All other permutations of the four groups were tested. Fig. 4 presents the results obtained by evaluating all band combinations on the five models.

The accuracy representation in Fig. 4 is the average of mIoU, F1 and OA of each epoch during the training. Except for the U-Net, the remaining models exhibit consistent accuracy rankings across different band combinations. The accuracy of data combinations in our proposed model fluctuates between0.731 and 0.815. Among the combinations, AD bands perform the worst with an accuracy of 0.763, while the other combinations show a small range of variation (0.026). The contribution of Sentinel-2 bands and image band indexes was assessed by comparing

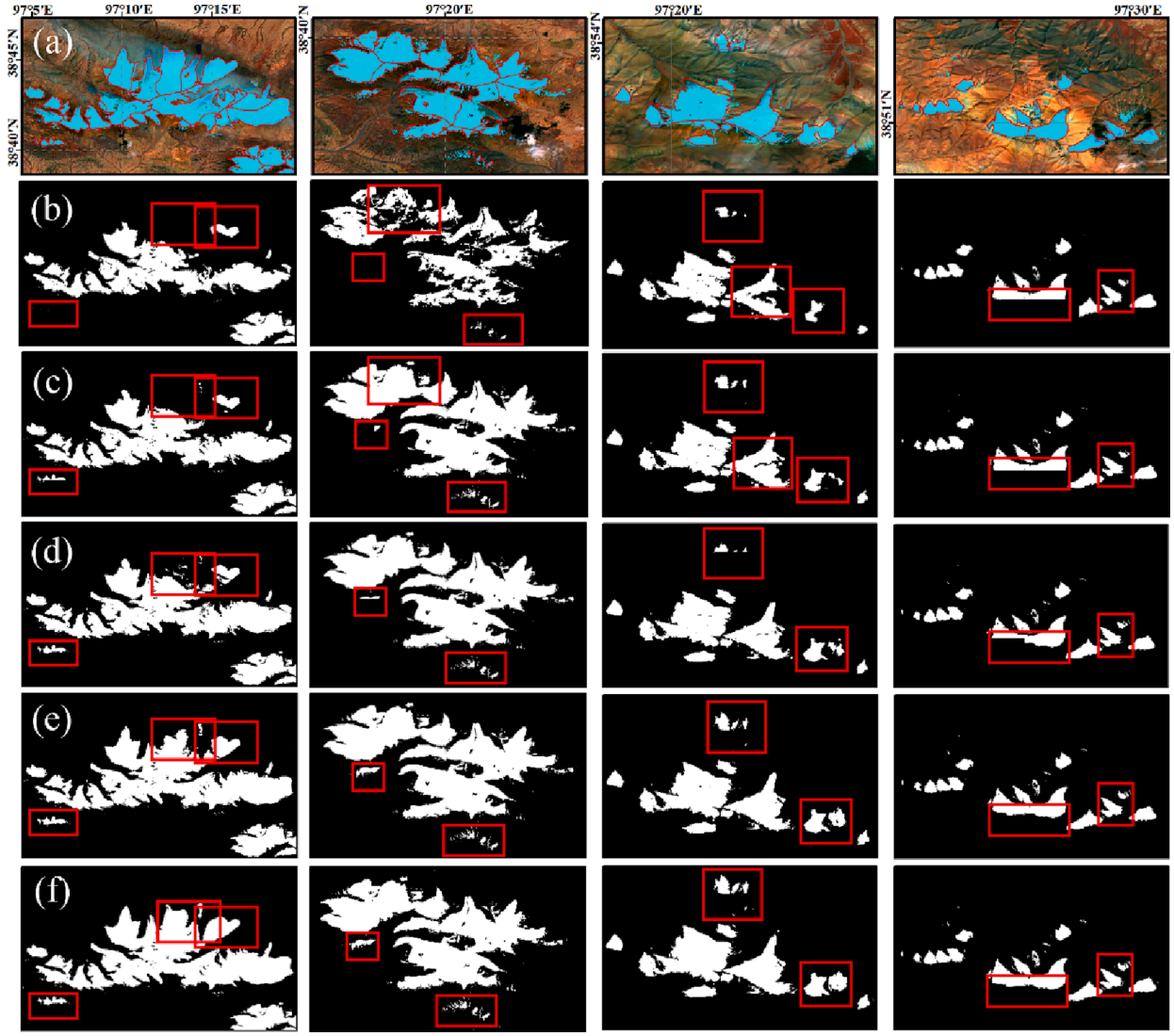

Fig. 3. The visualization results of U-Net, DeepLab V3+, Attention Deeplab V3+, Swin Transformer and our proposed model from (a) to (f) respectively. The four columns represent four regions. The base map of the figure in the first row is the false color composition of SWIN, NIR and green bands from Sentinel-2 imagery, the red line in it is the glacier outline. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

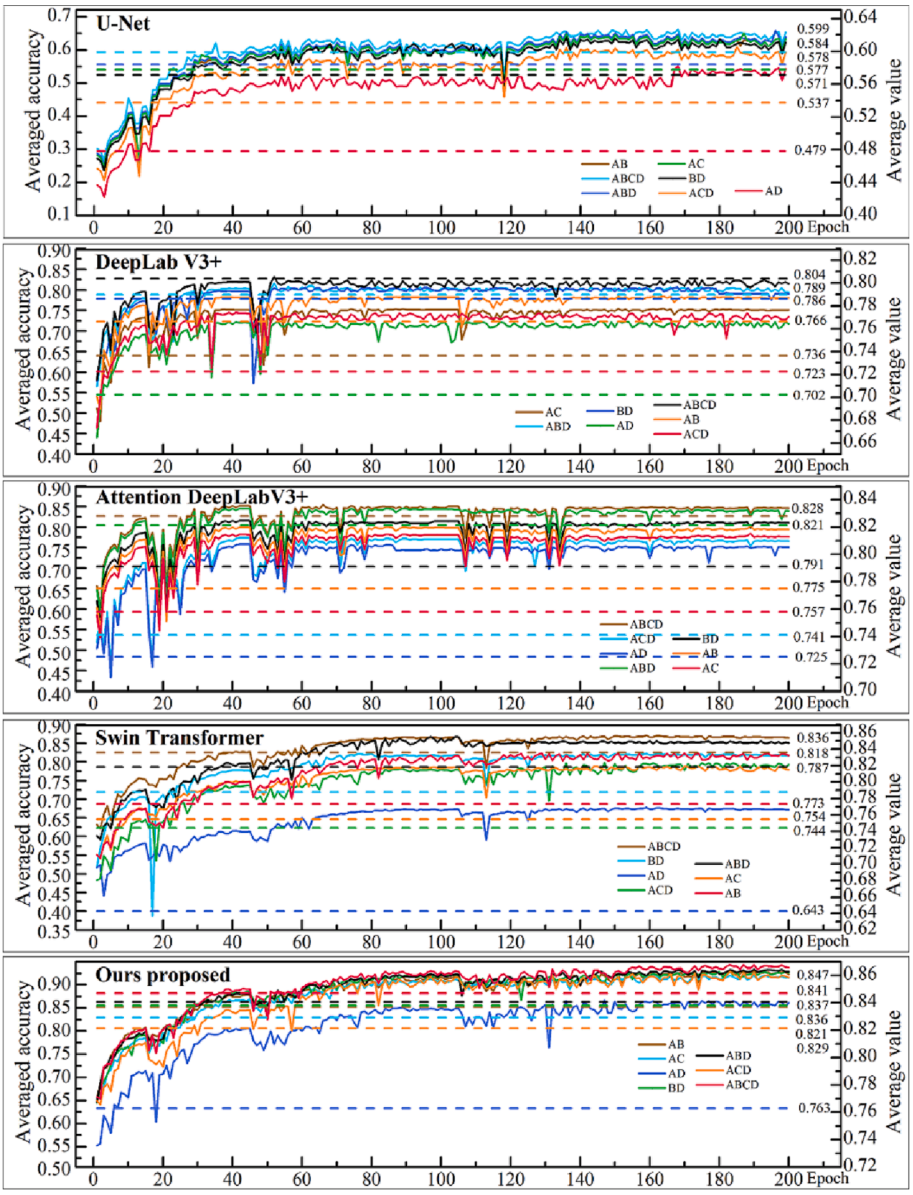

Fig. 4. Experimental results for different band combinations of U-Net, DeepLab V3+, Attention DeepLab V3+, Swin Transformer and our proposed model, where A is 'VV' and 'VH'; B is 'B2′, 'B3′, 'B4′, 'B8′ and 'B11′; C is 'ndvi', 'ndsi' and 'ndwi'; D is 'elevation'. The left vertical axis is the average of mIoU, F1 and OA. The right vertical axis is the average value of accuracy. The average value of each combination is shown following the dotted line.

ABD&ACD and AB&AC. The ABD combination demonstrates a higher accuracy by 0.008 compared to ACD, while AB surpasses AC by 0.015. This suggests that the original Sentinel-2 bands contribute more to glacier identification than the band indices. Furthermore, comparing AC (averaged accuracy = 0.821) and AD (averaged accuracy = 0.763) reveals that image band indexes have a greater contribution than elevation data. Additionally, the experiments confirm the contribution of Sentinel-1 and Sentinel-2 data, the two primary data sources, by comparing BD (averaged accuracy = 0.837) and AD (averaged accuracy = 0.763). The results show a significant accuracy increase of 0.74 with the inclusion of Sentinel-2 imagery.

The visualization results are presented in Fig. 5. It is evident that the results of AD (Fig. 5(b)) exhibit incomplete and fragmented glacier areas, indicating that the combination of Sentinel-1 and elevation data fails to capture the complete extent of glaciers. However, when Sentinel-2 data is added, as seen in Fig. 5(c)-(d), the misclassifications in the red-framed regions of the first, second, and fourth rows gradually diminish. This improvement is observed in the results of AB, BD, ABD, and ABCD, highlighting the significant contribution of Sentinel-2 bands in reducing misclassifications and enhancing the completeness of glacier identification.

3.5. *The model's response to the heterogeneity data.*

In order to extract glaciers in diverse conditions, we integrate Sentinel-1 and Sentinel-2 data, which exhibit distinct heterogeneity. In this section, we assess the capability of our proposed model to handle multi-source heterogeneous data by evaluating its performance on these two datasets under specific and typical glacier conditions. The specific glacier environment includes factors such as snow cover, debris cover, misclassification of water, and image cloud. In the selected region, the glaciers are minimally affected by debris, and the optical image chosen has a cloud cover of less than 5%. Therefore, we examine the circumstances of snow-covered glaciers and lake-terminus glaciers.

As illustrated in the third row of Fig. 6(a)-(b), the results obtained from Sentinel-2 imagery misidentify nearly all snow-covered areas as

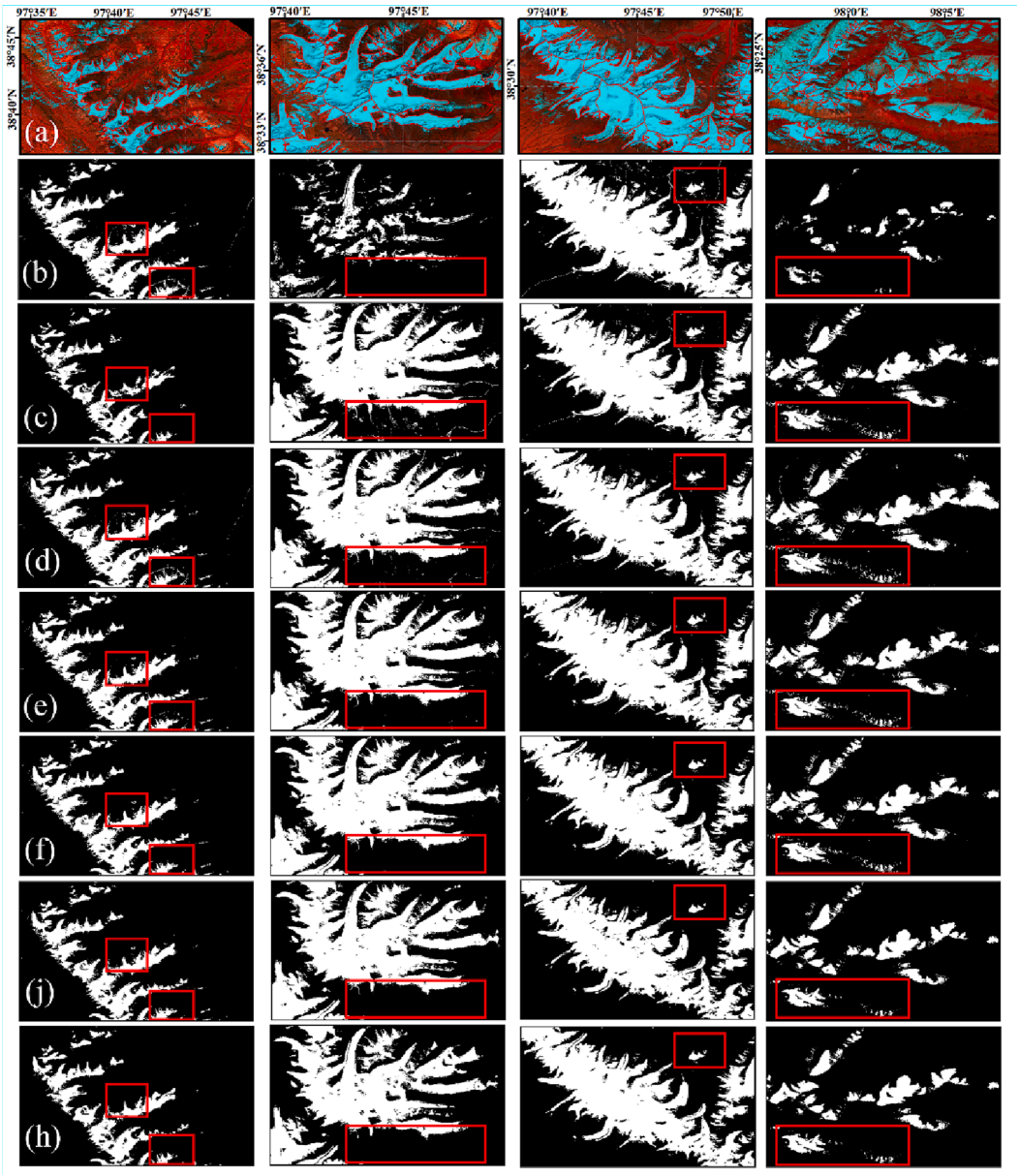

Fig. 5. The visualization results of seven band combinations on our proposed model, which represent AD, ACD, AC, AB, BD, ABD and ABCD from (a) to (h). The base map of the figure in the first row is the false color composition of SWIN, NIR and green bands from Sentinel-2 imagery, the red line in it is the ground truth of the glacier outline. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

glaciers. Similarly, in the experiments using Sentinel-2 data, the lake area is misclassified as a glacier, as shown in the third row of Fig. 6(c)-(d). However, the results obtained from Sentinel-1 data only misjudge a small portion of the snow or lake. This observation highlights the higher potential of Sentinel-1 imagery in accurately extracting glacier outlines under specific conditions.

However, when considering glacier extraction in general environments, as depicted in Fig. 6(e), the results obtained from Sentinel-2 data exhibit clear edges and more accurate delineation of glacier areas. This prompted us to examine the results obtained from both Sentinel-1 and Sentinel-2 data (the first row of Fig. 6). Under specific circumstances, as shown in the first line of Fig. 6(a)-(d), the combined results of the two datasets demonstrate a certain degree of misclassification of snow or lakes, with the misclassified areas falling between the results of Sentinel-1 and Sentinel-2 data. A similar phenomenon is observed under general circumstances, as depicted in the first row of Fig. 6(e), where the results from the two datasets exhibit clear edges and complete areas, but also include misjudged areas extracted from Sentinel-1 data. Thus, the combination of the two datasets does not clearly reflect the specific advantages of each dataset, indicating that our proposed model fails to distinguish the different information obtained from the concatenation of multi-source images. This further reveals that the proposed model is not sufficiently sensitive to heterogeneous features, suggesting that it cannot overcome the impact of complex terrain on glacier extraction. Therefore, in future work, we will consider the adoption of a dual-branch processing approach tailored to specific data.

3.6. Model efficiency

The parameter quantity and training time for a specific module are presented in Table 2. The efficiency and accuracy of the encoder architecture are evaluated using different interleaved methods of LSA and GSA. As shown in Table 2, compared to (LL, LG, LG, GG), our proposed design (LG, LG, LG, LG) achieves a marginal accuracy improvement of

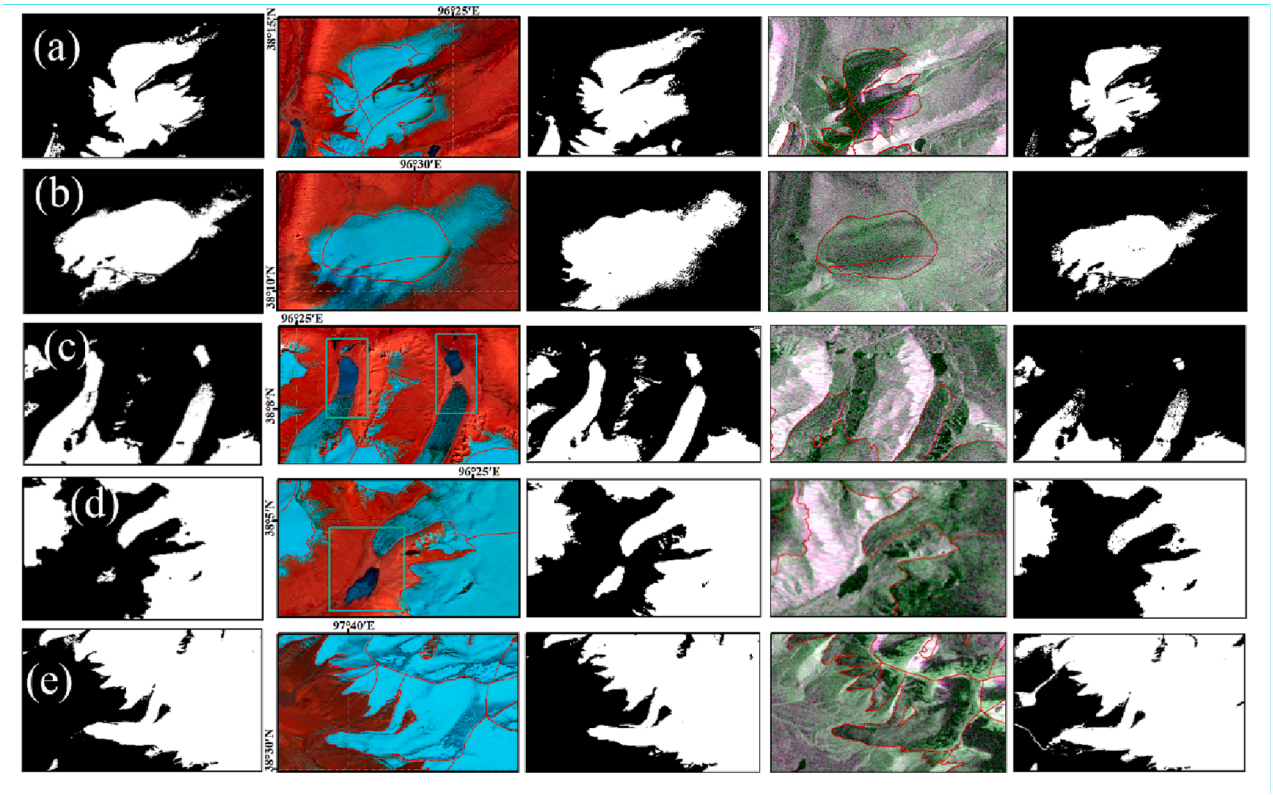

Fig. 6. The extract results from Sentinel-1(the fifth line), Sentinel-2(the third line) and both of them(the first line). The second and the fourth line is the Sentinel-2 image(false color composition of Short Wave Infrared(SWIR), Near Infrared(NIR) and green bands) and Sentinel-1 image(false color composition of VV, VH and VV). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

0.07 at the cost of a significant increase in parameters (8.9 M) and training time. This suggests that while incorporating both high and low-level image information in the fusion of local and global attention contributes less to accuracy, it significantly increases model complexity. Therefore, future research should focus on more efficient strategies for capturing local and global information.

4. Conclusion

In this study we have presented an automated method of glacier extraction based on the Transformer model. The model involves a decoder with Local-global Transformer and decoder with dual branches of CNN and Attention, in which the idea of combining the global–local information is both involved. The Sentinel-1, Sentinel-2, HMA and SRTM DEM data sets were used. The model was trained by the glaciers area of the Qilian Mountains extracted semi-automatically based on GF data. We demonstrated the best accuracy of 0.972 among all other models and the necessity of each module in our proposed model through the experiments on different models and ablation studies. The method we propose can enhance the precision of automatically extracting glacier, while also demonstrating the effectiveness of combining data from various acquisition methods. Our experiments showed that optical imagery plays a significant role in glacier extraction, and SAR data can differentiate snow-covered glaciers and glacier-terminating lakes. Nevertheless, our proposed model is not sensitive for the features extracted from heterogeneous data, thus the dual-branch processing approach tailored to specific data is considered in the future. Moreover, the Transformer-based model makes the local–global information extraction is time-consuming, the further improvement should focus on the efficient interaction of local–global information.

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

This work was support by NationalKey R&D Program of China (2022YFB3903400)

References

Ba, J., Mnih, V., Kavukcuoglu, K., 2014. Multiple object recognition with visual attention. arXiv preprint arXiv.

Baumhoer, C.A., Dietz, A.J., Kneisel, C., Kuenzer, C., 2019. Automated extraction of antarctic glacier and ice shelf fronts from sentinel-1 imagery using deep learning. Remote Sens. 11, 2529.

Bhattacharya, A., Bolch, T., Mukherjee, K., King, O., Menounos, B., Kapitsa, V., Neckel, N., Yang, W., Yao, T., 2021. High Mountain Asian glacier response to climate revealed by multi-temporal satellite observations since the 1960s. Nat. Commun. 12, 1–13.

Blaschke, T., 2010. Object based image analysis for remote sensing. ISPRS J. Photogramm. Remote Sens 65 (1), 2–16.

Bolch, T., Menounos, B., Wheate, R., 2010. Landsat-based inventory of glaciers in western Canada, 1985–2005. Remote Sens. Environ. 114, 127–137.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M., 2021. Swin-unet: Unet-like pure transformer for medical image segmentation. arXiv preprint arXiv: .05537.

Catania, G., Stearns, L., Sutherland, D., Fried, M., Bartholomaus, T., Morlighem, M., Shroyer, E., Nash, 2018. Geometric controls on tidewater glacier retreat in central western Greenland. J. Geophys. Res.: Earth Surf. 123, 2024-2038.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation, Proceedings of the Eur. Conf. Comput. Vis. (ECCV), pp. 801-818.

Cheng, D., Hayes, W., Larour, E., Mohajerani, Y., Wood, M., Velicogna, I., Rignot, E., 2021. Calving Front Machine (CALFIN): glacial termini dataset and automated deep learning extraction method for Greenland, 1972–2019. Cryosphere 15, 1663–1675.

Chu, X., Tian, Z., Zhang, B., Wang, X., Wei, X., Xia, H., Shen, C., 2021b. Conditional positional encodings for vision transformers. arXiv preprint arXiv:.10882.
Chu, X., Tian, Z., Wang, Y., Zhang, B., Ren, H., Wei, X., Xia, H., Shen, C., 2021. Twins: Revisiting the design of spatial attention in vision transformers. Proc. Adv. Neural Inf. Process. Syst. 34, 9355–9366.
Chu, X., Yao, X., Duan, H., Chen, C., Li, J., Pang, W., 2022. Glacier extraction based on high-spatial-resolution remote-sensing images using a deep-learning approach with attention mechanism. Cryosphere 16, 4273–4289.
Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:.11929.
Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., 2012. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. Remote Sens. Environ. 120, 25–36.
Frey, H., Paul, F., Strozzi, T., 2012. Compilation of a glacier inventory for the western Himalayas from satellite data: methods, challenges, and results. Remote Sens. Environ. 124, 832–843.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 770–778.
Holzer, N., Vijay, S., Yao, T., Xu, B., Buchroithner, M., Bolch, T., 2015. Four decades of glacier variations at Muztagh Ata (eastern Pamir): a multi-sensor study including Hexagon KH-9 and Pléiades data. Cryosphere 9, 2071–2088.
Huss, M., Hock, R., 2018. Global-scale hydrological response to future glacier mass loss. Nat. Clim. Change 8, 135–140.
Immerzeel, W.W., Van Beek, L.P., Bierkens, M.F., 2010. Climate change will affect the Asian water towers. Science 328, 1382–1385.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.
Lemke, P., Ren, J., Alley, R.B., Allison, I., Carrasco, J., Flato, G., Fujii, Y., Kaser, G., Mote, P., Thomas, R.H., 2007. Observations: changes in snow, ice and frozen ground. Climate Change 2007, IPCC Chapter 4.
[dataset] Li, J. (2022). The glacier inventory of Qilian Mountain Area (v2.0, 2020). National Tibetan Plateau/Third Pole Environment Data Center, https://doi.org/10.11888/Cryos.tpdc.272461. https://cstr.cn/18406.11.Cryos.tpdc.272461.
Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows, Proc. IEEE/CVF Int. Conf. on Comput, Vis., pp. 10012-10022.
Malenovský, Z., Rott, H., Cihlar, J., Schaepman, M.E., García-Santos, G., Fernandes, R., Berger, M.J.R.S.o.e., 2012. Sentinels for science: Potential of Sentinel-1,-2, and-3 missions for scientific observations of ocean, cryosphere, and land. Remote Sens. Environ. 120, 91-101.
Milillo, P., Rignot, E., Rizzoli, P., Scheuchl, B., Mouginot, J., Bueso-Bello, J.L., Prats-Iraola, P., Dini, L., 2022. Rapid glacier retreat rates observed in West Antarctica. Nature Geosci. 15, 48–53.
Milletari, F., Navab, N., Ahmadi, S.-A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation, 2016 fourth international conference on 3D vision (3DV). IEEE 565–571.
Mohajerani, Y., Wood, M., Velicogna, I., Rignot, E., 2019. Detection of glacier calving margins with convolutional neural networks: A case study. Remote Sens. 11, 74.
Paul, F., Barrand, N.E., Baumann, S., Berthier, E., Bolch, T., Casey, K., Frey, H., Joshi, S., Konovalov, V., Le Bris, R., 2013. On the accuracy of glacier outlines derived from remote-sensing data. Ann. Glaciol. 54, 171–182.
Paul, F., Bolch, T., Kääb, A., Nagler, T., Nuth, C., Scharrer, K., Shepherd, A., Strozzi, T., Ticconi, F., Bhambri, R., 2015. The glaciers climate change initiative: Methods for creating glacier area, elevation change and velocity products. Remote Sens. Environ. 162, 408–426.
Paul, F., Rastner, P., Azzoni, R.S., Diolaiuti, G., Fugazza, D., Le Bris, R., Nemec, J., Rabatel, A., Ramusovic, M., Schwaizer, G., 2020. Glacier shrinkage in the Alps continues unabated as revealed by a new glacier inventory from Sentinel-2. Earth Syst. Sci. Data 12, 1805–1821.
Peng, Y., Li, Z., Xu, C., Zhang, H., Han, W., 2021. Surface velocity analysis of surge region of karayaylak glacier from 2014 to 2020 in the pamir plateau. Remote Sens. 13, 774.
Pritchard, H.D., 2019. Asia's shrinking glaciers protect large populations from drought stress. Nature 569, 649–654.
Racoviteanu, A.E., Arnaud, Y., Williams, M.W., Manley, W.F., 2015. Spatial patterns in glacier characteristics and area changes from 1962 to 2006 in the Kanchenjunga-Sikkim area, eastern Himalaya. Cryosphere 9, 505–523.
Robson, B.A., Nuth, C., Dahl, S.O., Hölbling, D., Strozzi, T., Nielsen, P.R., 2015. Automated classification of debris-covered glaciers combining optical, SAR and topographic data in an object-based environment. Remote Sens. Environ. 170, 372–387.
Rosen, P.A., Hensley, S., Joughin, I.R., Li, F.K., Madsen, S.N., Rodriguez, E., Goldstein, R. M., 2000. Synthetic aperture radar interferometry. Proc. IEEE 88, 333–382.
Salomonson, V.V., Appel, I., 2004. Estimating fractional snow cover from MODIS using the normalized difference snow index. Remote Sens. Environ. 89, 351–360.
Shean, D.E., Alexandrov, O., Moratto, Z.M., Smith, B.E., Joughin, I.R., Porter, C., Morin, P., 2016. An automated, open-source pipeline for mass production of digital elevation models (DEMs) from very-high-resolution commercial stereo satellite imagery. ISPRS J. Photogramm. Remote Sens. 116, 101–117.
[dataset]Shean, D. (2017). High Mountain Asia 8-meter DEM Mosaics Derived from Optical Imagery, Version 1. Boulder, Colorado USA. NASA National Snow and Ice Data Center Distributed Active Archive Center. https://doi.org/10.5067/KXOVQ9L172S2.
Shi, Y., Liu, S., 2000. Estimation on the response of glaciers in China to the global warming in the 21st century. Chi. Sci. Bull. 45, 668–672.
Strudel, R., Garcia, R., Laptev, I., Schmid, C., 2021. Segmenter: Transformer for semantic segmentation. Proc. IEEE/CVF Int. Conf. Comput. Vis. 7262–7272.
Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Jorge Cardoso, M., 2017. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations, Deep learning in medical image analysis and multimodal learning for clinical decision support. Springer 240–248.
Sun, W., Li, Q., Li, B., 2019. Does geographic distance have a significant impact on enterprise financing costs? J. Geogr. Sci. 29, 1965–1980.
Sun, M., Liu, S., Yao, X., Guo, W., Xu, J., 2018. Glacier changes in the Qilian Mountains in the past half-century: Based on the revised First and Second Chinese Glacier Inventory. J. Geogr. Sci. 28, 206–220.
Torres, R., Snoeij, P., Geudtner, D., Bibby, D., Davidson, M., Attema, E., Potin, P., Rommen, B., Floury, N., Brown, M., 2012. GMES Sentinel-1 mission. Remote Sens. Environ. 120, 9–24.
Toutin, T., Cheng, P., 2002. QuickBird–a milestone for high resolution mapping. Earth Observation Magazine 11, 14–18.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Proc. Adv. Neural Inf. Process, Syst, p. 30.
Wang, H., Chen, X., Zhang, T., Xu, Z., Li, J., 2022a. CCTNet: Coupled CNN and Transformer Network for Crop Segmentation of Remote Sensing Images. Remote Sens. 14, 1956.
Wang, X., Girshick, R., Gupta, A., He, K., 2018. Non-local neural networks. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 7794–7803.
Wang, L., Li, R., Zhang, C., Fang, S., Duan, C., Meng, X., Atkinson, P.M., 2022b. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. ISPRS J. Photogramm. Remote Sens. 190, 196–214.
Wang, S., Peppa, M.V., Xiao, W., Maharjan, S.B., Joshi, S.P., Mills, J.P., 2022c. A second-order attention network for glacial lake segmentation from remotely sensed imagery. ISPRS J. Photogramm. Remote Sens. 189, 289–301.
Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L., 2021. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. IEEE Int. Conf. Comput. Vis. 568–578.
Woo, S., Park, J., Lee, J.-Y., Kweon, I.S., 2018. Cbam: Convolutional block attention module. Eur. Conf. Comput. Vis. (ECCV) 3–19.
Xiao, Y., Su, X., Yuan, Q., Liu, D., Shen, H., Zhang, L., 2021. Satellite video superresolution via multiscale deformable convolution alignment and temporal grouping projection. IEEE Trans. Geosci. Remote Sens. 60, 1–19.
Xiao, Y., Yuan, Q., He, J., Zhang, Q., Sun, J., Su, X., Wu, J., Zhang, L., 2022. Space-time super-resolution for satellite video: A joint framework based on multi-scale spatialtemporal transformer. Int. J. Appl. Earth Obs. Geoinf. 108, 102731.
Xu, H., 2006. Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. Int. J. Remote Sens. 27, 3025–3033.
Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., Xu, H., Tan, W., Yang, Q., Wang, J., 2020. Deep learning in environmental remote sensing: Achievements and challenges. Remote Sens. Environ. 241, 111716.
Zhang, G., Chen, W., Li, G., Yang, W., Yi, S., Luo, W., 2020. Lake water and glacier mass gains in the northwestern Tibetan Plateau observed from multi-sensor remote sensing data: Implication of an enhanced hydrological cycle. Remote Sens. Environ. 237, 111554.
Zhang, C., Jiang, W., Zhang, Y., Wang, W., Zhao, Q., Wang, C., 2022. Transformer and CNN Hybrid Deep Neural Network for Semantic Segmentation of Very-High-Resolution Remote Sensing Imagery. IEEE Trans. Geosci. Remote Sens. 60, 1–20.
Zhang, E., Liu, L., Huang, L., 2019. Automatically delineating the calving front of Jakobshavn Isbræ from multitemporal TerraSAR-X images: a deep learning approach. Cryosphere 13, 1729–1741.
Zhang, E., Liu, L., Huang, L., Ng, K.S., 2021. An automated, generalized, deep-learning-based method for delineating the calving fronts of Greenland glaciers from multi-sensor remote sensing imagery. Remote Sens. Environ. 254, 112265.
Zhao, X., Wang, X., Wei, J., Jiang, Z., Zhang, Y., Liu, S., 2020. Spatiotemporal variability of glacier changes and their controlling factors in the Kanchenjunga region, Himalaya based on multi-source remote sensing data from 1975 to 2015. Sci. Total Environ. 745, 140995.
Zhou, C., Zheng, L., 2017. Mapping radar glacier zones and dry snow line in the Antarctic Peninsula using Sentinel-1 images. Remote Sens. 9, 1171.