ISPRS Journal of Photogrammetry and Remote Sensing 189 (2022) 289–301

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

A second-order attention network for glacial lake segmentation from remotely sensed imagery

Shidong Wang [a], Maria V. Peppa [a], Wen Xiao [b,c,a,*], Sudan B. Maharjan [d], Sharad P. Joshi [d], Jon P. Mills [a]

[a] School of Engineering, Newcastle University, Newcastle upon Tyne NE1 7RU, UK
[b] School of Geography and Information Engineering, China University of Geosciences, Wuhan 430074, China
[c] National Engineering Research Center of Geographic Information System, China University of Geosciences, Wuhan 430074, China
[d] International Centre for Integrated Mountain Development, Kathmandu, GPO Box 3226, Nepal

ARTICLE INFO

Keywords:
Deep learning
GLOF
Satellite imagery
Climate change
Landsat
Self-attention mechanism

ABSTRACT

Climate change is increasing the risk of glacial lake outburst floods (GLOFs) in many of the world's most vulnerable and high mountain regions. Simultaneously, remote sensing technologies now facilitate continuous monitoring of glacial lake evolution around the globe, although accurate and reliable automated glacial lake mapping from satellite data remains challenging. In this study, a Second-order Attention Network (SoAN) is devised for the automated segmentation of lakes from satellite imagery. In particular, a novel Second-order Attention Module (SoAM) is proposed to capture the long-range spatial dependencies and establish channel attention derived from the covariance representations of local features. Furthermore, as the dimensions of the input and output tensors are identical and it simply relies on matrix calculations, the proposed SoAM can be embedded into different positions of a given architecture while maintaining similar reference speed. The designed network is implemented on Landsat-8 imagery and outputs are compared against representative deep learning models, demonstrating improved results with a Dice of 81.02% and a F2 Score of 85.17%.

1. Introduction

Over recent decades, glacier ice melt rate has significantly increased due to global warming (Blunden et al., 2020; Shugar et al., 2020). A number of studies have suggested that glaciers in the Hindu Kush Himalaya (HKH), as in other parts of the world have been shrinking (Bajracharya et al., 2020; Maharjan et al., 2018; Nie et al., 2017). As a direct consequence of ice melt, new or expanded glacial lakes are found in such high mountain regions, increasing the risk of glacial lake outburst floods (GLOFs) and in-turn posing a significant threat to downstream communities and infrastructure. Since 1950s, in excess of 50 GLOFs have already been recorded in the HKH region and there may be more unrecorded or undocumented (Veh et al., 2018). The damage caused by GLOFs is often more catastrophic than hydrometeorological floods as peak discharges can surpass monsoonal river discharge by several orders of magnitude. Studies have shown that HKH GLOFs have the highest death toll worldwide (Veh et al., 2020). Moreover, it has been estimated that the glaciers in Nepal lost almost a quarter of their total area from the 1980s to 2010, and the number of glacial lakes increased by 11% (Maharjan et al., 2018; Bajracharya et al., 2014). The expansions of new and existing glacial lakes has led to higher risks of GLOFs, and it is therefore crucial to document and regularly monitor the spatio-temporal evolution of glacial lake in the HKH region.

Large-scale studies of glaciers and glacial lakes in the HKH started in the 1980s using topographic maps, aerial photographs and field investigations. During the years 1999 and 2005, the International Centre for Integrated Mountain Development (ICIMOD) made an inventory of glaciers and lakes in five HKH countries (China, Nepal, Bhutan, India and Pakistan), using topographic maps and available satellite imagery, including Landsat TM, IRS, SPOT (Maharjan et al., 2018). A total of 8,790 lakes were identified, 203 of which were identified as potentially dangerous(Ives et al., 2010). Further mapping of glacial lakes and assessment of GLOF risks was conducted in Nepal during 2009, using Landsat imagery captured in 2005 and 2006. 1,466 lakes were mapped, 21 of them identified as in a critical state (Mool et al., 2011). A more recent study found 3,624 glacial lakes in the Koshi, Gandaki, and Karnali

* Corresponding author at: School of Geography and Information Engineering, China University of Geosciences, Wuhan 430074, China.
E-mail addresses: Shidong.wang@newcastle.ac.uk (S. Wang), Maria-Valasia.Peppa@newcastle.ac.uk (M.V. Peppa), wen.xiao@newcastle.ac.uk, Wen.Xiao@cug.edu.cn (W. Xiao), sudan.maharjan@icimod.org (S.B. Maharjan), sharad.joshi@icimod.org (S.P. Joshi), jon.mills@newcastle.ac.uk (J.P. Mills).

https://doi.org/10.1016/j.isprsjprs.2022.05.007
Received 1 November 2021; Received in revised form 18 May 2022; Accepted 20 May 2022
Available online 29 May 2022

Table 1
Recent studies that investigate the use of deep learning methods for mapping and monitoring of the spatio-temporal evolution of glaciers and glacial lakes.

Reference	Study site	Image source	Deep learning method
Zhang et al. (2021)	Greenland glaciers	Landsat-8; ALOS-1; TSX; Sentinel-1	DeepLabV3+
Dirscherl et al. (2021)	Supraglacial lakes in Antarctica	Sentinel-1	Modified U-Net
Cheng et al. (2021)	Glacial terminus in Greenland	Landsat-8; Sentinel-1; TerraSAR-X	DeepLabV3 + Xception
Qayyum et al. (2020)	Glacial lakes in the HKH region	PlanetScope	VGG U-Net; Efficient U-Net
Wu et al. (2020)	Glacial lakes in south-eastern Tibet	Landsat-8; Sentinel-1	Modified U-Net
Xie et al. (2020)	Glaciers in Karakoram; Himalayas	Landsat-8; ALOS	GlacierNet: Modified SegNet
Baraka et al. (2020)	Glaciers in the HKH region	Landsat-7; SRTM	U-Net

basins of Nepal, China and India, and identified 47 potentially dangerous lakes in that region. The study utilised 2015 to 2018, Landsat-8 Operational Land Imager (OLI) Top of the Atmosphere (TOA) products (Bajracharya et al., 2020).

It has been demonstrated that remote sensing is an effective technology to map glacial lakes at a regular frequency (Song et al., 2014). Although various optical and radar data can today be used for glacier and lake mapping and monitoring, many studies still rely on Landsat imagery thanks to its wide coverage and long serving history (Bhardwaj et al., 2015; Nie et al., 2017; Veh et al., 2018; Qayyum et al., 2020; Wangchuk and Bolch, 2020). Image band index calculation, together with histogram thresholding constitutes the most commonly-applied techniques implemented with optical data for identifying and delineating water bodies. Among several indices the Normalized Difference Water Index (NDWI; Gao (1996)) and the Normalized Difference Snow Index (NDSI; Salomonson and Appel (2004)), which is also known as the Modified NDWI (Xu, 2006; Chen et al., 2017), are the most established indices used for both glacier and lake mapping (Bhardwaj et al., 2015; Chen et al., 2017). Those indices combine the shortwave infrared (SWIR) with the near infrared (NIR) bands in which water absorbs that part of the electromagnetic spectrum, and the visible green band in which water is highly reflective. The Normalized Difference Vegetation Index (NDVI) has also been jointly used alongside the aforementioned indices as it can highlight the presence of vegetation either on land or in water. Typically, if the index value exceeds a preset threshold, the pixel is considered as water. However, such rule-based methods can be inaccurate at the boundaries of water bodies where mixed spectral information of land, vegetation, water and snow may be present. A single threshold value is therefore usually inadequate to distinguish water or snow from the adjacent land (Chen et al., 2017). To refine the results at the boundaries of various land cover classes, and thereby achieve sub-pixel accuracy, previous studies combined rule-based methods either with other algorithms (e.g. the non-local active contour algorithm in Chen et al. (2017) and the hierarchical rule-based classification in Guo et al. (2021)) or with machine learning methods (Veh et al., 2018; Zhang et al., 2020). Rishikeshan and Ramesh (2018) proposed a morphological operator based approach that outperformed the maximum likelihood classification for water body extraction.

With the evolution of cloud computing platforms (e.g. Google Earth Engine (GEE) (Gorelick et al., 2017; Tamiminia et al., 2020)), that allow the analysis of freely available historic satellite image series, the implementation of machine learning algorithms for water extent mapping (Pekel et al., 2016) and glacial lake detection (Veh et al., 2018) has gained increased attraction in the remote sensing community. For instance, random forest (RF) is one of the most commonly used machine learning classifiers in glacier studies (Wangchuk and Bolch, 2020) and is, embedded into the GEE platform thereby enabling a straight forward implementation. Even though previous studies showed that RF can provide accuracy levels better than 90% in glacial lake segmentation, especially when appropriate band indices are used as auxiliary attributes for training the classifier (Veh et al., 2018; Zhang et al., 2020), mixed spectral noise caused by shadows, clouds and ice still remains a fundamental challenge. As seen in Wu et al. (2020), when optical (e.g. Landsat-8) and radar (e.g.Sentinel-1 Synthetic Aperture Radar (SAR)) images were incorporated into the RF classifier training, RF could better distinguish between shadows and frozen lakes, however the RF approach still provided poor results over regions with low spectral reflection.

More recently, following the advancement of deep learning based image understanding, remote sensing studies have shifted from easy-to-implement classic machine learning methods towards convolutional neural networks (CNNs) (Zhu et al., 2017). CNNs have been demonstrated to show improved results over classical machine learning, especially when mapping large geographical areas with varying geomorphological terrain characteristics (Xie et al., 2020). This shift is also evidenced by the high number of publications using CNNs with Earth Observation data in image segmentation applications since 2012 (Hoeser and Kuenzer, 2020; Hoeser et al., 2020). Li et al. (2022) provided a comprehensive review of water body classification methods from optical remote sensing imagery, and demonstrated the advantages and opportunities of deep learning-based methods over non-deep learning-based methods. In relation to glacier and glacial lake mapping, the U-Net (Ronneberger et al., 2015; Baraka et al., 2020), SegNet (Long et al., 2015) and DeepLabV3+ (Chen et al., 2018) CNN architectures have proven particularly popular. The U-Net and SegNet models, alongside their variants, are encoder-decoder architectures and DeepLabV3 + is an improved variant of a naive-encoder to an encoder-decoder architecture. Even though such architectures were initially designed to cope with traditional computer vision datasets (e.g. small-size images used in medical studies), it has been shown that they can also handle the diverse properties found in Earth Observation imagery (e.g. optical versus SAR, dense and heterogeneous classes etc.) (Hoeser and Kuenzer, 2020).

Studies reported over the last two years (Table 1) have successfully demonstrated that the aforementioned encoder-decoder model variants provide a substantial baseline for automatic segmentation implemented in glacier research. Despite high performance in segmentation results, several limitations exist. These include: a) misclassifications at sharp boundaries between glaciers and icebergs (Zhang et al., 2021), between glacial lake features that are slightly frozen and blue ice or wet snow (Dirscherl et al., 2021), and at glacial lake edges with mixed pixels of ice, clouds, debris or dry/wet snow (Xie et al., 2020), even with very high spatial resolution imagery (e.g. in Qayyum et al. (2020)); b) technical challenges when dealing with training datasets at inconsistent temporal and spatial scales especially when imagery is derived from different sources such as optical and radar (Wu et al., 2020); and c) suboptimal results when mountainous shadows are not entirely masked out (Wu et al., 2020), or in cases of high variability in glacier retreat rates and feature formation of glacial lakes as found in the Nepal Himalayas (Xie et al., 2020).

Moreover, the studies listed in Table 1 all utilised existing encoder-decoder architectures and their contributions mainly focused on the use of multiple satellite data sources and/or very high spatial resolution imagery, as well as the comparative analysis of the adopted model's learning capability with or without a large amount of training data. As far as research on glacial lake mapping is concerned, the development of a neural network itself is still limited. Whilst additional deep learning modules have been combined with the aforementioned encoder-decoder architectures and implemented within Earth Observation (e.g. attention modules for object detection in Yang et al. (2018) and for river segmentation in Xia et al. (2019)), as yet there is no such model development or architecture combination for glacial lake segmentation

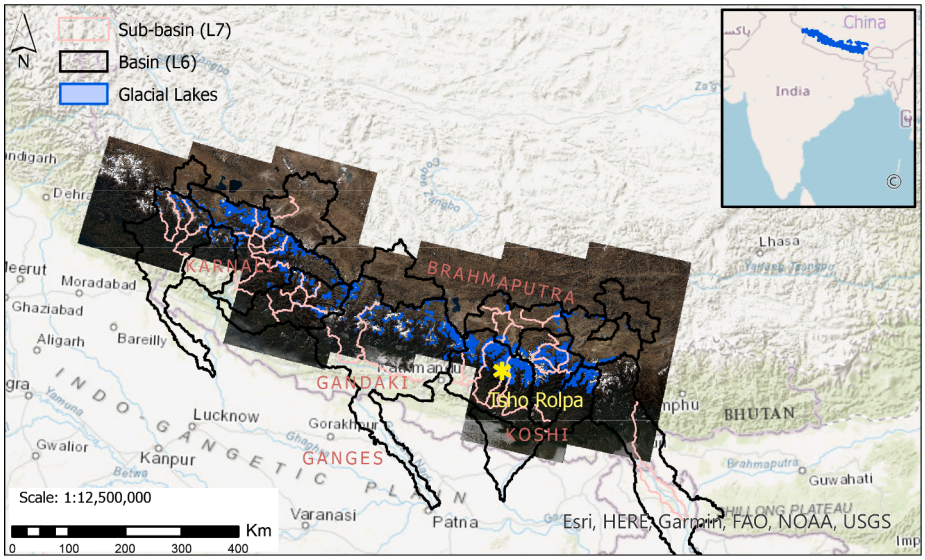

Fig. 1. Overview of the glacial lakes in the Koshi, Gandaki, and Karnali river basins of the HKH region superimposed over a tile mosaic of the Landsat-8 imagery used in the presented experiments. The Tsho Rolpa test site is highlighted in yellow. Basins and sub-basins are level 6 and 7 products respectively of the HydroSHEDS database, retrieved from Lehner (2013), and depicted here for reference.

Table 2
Landsat-8 Surface Reflectance Tier 1 OLI and TIR bands as retrieved by EEDC (2021). Notice that B8 and B9 bands are not listed here since they are not processed to Surface Reflectance by GEE. The last three quality bands were generated in GEE.

Landsat-8 Bands	Centre Wavelength [nm]	Description	Spatial Resolution [m]
B1	443	Coastal aerosol	30
B2	482	Blue	30
B3	561	Green	30
B4	655	Red	30
B5	865	NIR	30
B6	1609	SWIR-1	30
B7	2201	SWIR-2	30
B10	10895	TIR-1	100
B11	12005	TIR-2	100
sr aerosol	–	Aerosol attributes	–
pixel qa	–	Pixel quality attributes	–
radsat qa	–	Quality mask of radiometric saturation	–

applications.

Recently, the HarDNet-MSEG network (Huang et al., 2021) was proposed for polyp segmentation, which uses the encoder part of its predecessor HarDNet and the new decoder part analogous to Cascaded Partial Decoder (i.e., the Receptive Field Block (RFB) Module and the Dense Aggregation, as described in Section 3.2.1. Taking advantages of the low memory traffic of HarDNet and the effectiveness of the Cascaded Partial Decoder, HarDNet-MSEG can outperform existing well-known architectures in terms of segmentation accuracy and inference speed, including U-Net and its variants, and DeepLabV3+. However, the main purpose of introducing a structure similar to the Cascaded Partial Decoder is to use rapid inference for detecting salient objects, and the segmentation accuracy of the model is expected to be further improved. In addition, the effectiveness of deploying HarDNet-MSEG into glacier lake segmentation from satellite imagery is still unknown, as it involves the processing of numerous bands with variant characteristics, which makes the segmentation more challenging.

Recently, non-local neural networks (Wang et al., 2018) have been proposed to leverage the efficiency of self-attention mechanisms to eminently improve the performance of various computer vision tasks. Furthermore, a linear attention mechanism (Li et al., 2020) has been introduced to approximate the dot-product attention map by its first-order Taylor expansion, thereby decreasing computational cost. Inspired by the aforementioned work, a novel attention module is proposed in this study, which adopts the matrix product to capture adequate long-distance spatial dependence. On this basis, additional second-order statistics and appropriate matrix normalisation are introduced to establish the relationships between the feature spatial locations and the channels. The proposed attention module is highly compatible with the adopted HarDNet-MSEG backbone network so that it can be transferred to the glacier lake image segmentation scenario to further improve the precise cutting of boundary areas, small and irregular regions. To the best of our knowledge, this is the first non-local module that implements channel attention derived from matrix covariance representations and can work concurrently with original spatial attention maps.

To achieve this, first, a publicly available dataset extracted semi-automatically from Landsat-8 images in the HKH region is prepared. Then, a second-order attentive architecture is introduced, and two variant models are derived for comparison with state-of-the-art networks. The influences of different combinations of training input and image conditions are also analysed. In addition, the Tsho Rolpa glacial lake, one of the most dangerous glacier lakes in the region (Shrestha and Nakagawa, 2014; Bajracharya et al., 2020), is used as a test case to further compare the segment at ground level using high resolution aerial imagery acquired from an Unmanned Aerial System (UAS).

2. Study area and dataset description

A region in HKH, consisting of 3,624 glacial lakes, is adopted as the study area in the presented experiments (Fig. 1). The glacial lakes were retrieved from ICIMOD (2020) and Bajracharya et al. (2020) in a vector shapefile format. The retrieved dataset comprised the glacial lake boundaries, primarily based on Landsat imagery from 2015 and 2016 plus two tiles from 2017 and 2018, which were all formed on paleo-glacier landforms (Bajracharya et al., 2020; Maharjan et al., 2018). Water pixels located over river or non-glacial lakes were not included in the datasets as they were filtered out via a validation process utilising high resolution Google Earth Imagery and manual cross-validation. Similar processes were applied using high resolution Google Earth Imagery to amend erroneous or missed boundaries that did not fit the aforementioned years' dynamics (ICIMOD, 2020; Bajracharya et al.,

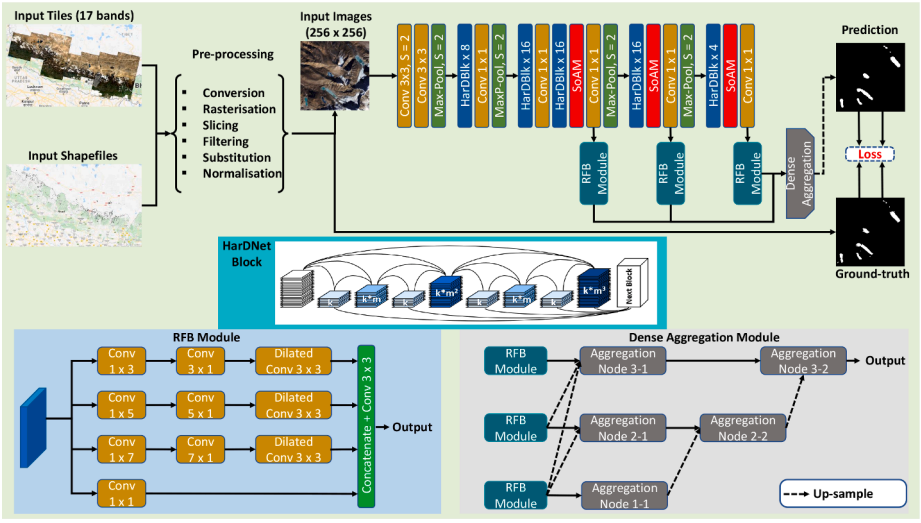

Fig. 2. Overview of the proposed Second-order Attention Network (SoAN). The input tiles and shapefiles are preprocessed, and the features of the resulting images are extracted by an encoder consisting of a series of generic convolution and max pooling operations, as well as HarDNet blocks and Second-order Attention Modules (SoAM) (illustrated in Fig. 3) inserted at specific locations. The decoder is composed of three Receptive Field Block (RFB) modules and a Dense Aggregation Module. Note that HarDNet blocks, RFB modules and Dense Aggregation Module are borrowed from Huang et al. (2021).

2020). In the presented study, the datasets were used directly as retrieved from ICIMOD (2020) without any further amendment. A list of Landsat imagery used in the presented experiments is shown in Table 5 in the Appendix.

In particular, 25 raw tiles of Landsat-8 Surface Reflectance (SR) covering the studied region, shown in Fig. 1, were collected for the years of 2015, 2016, 2017 and 2018 in order to match the updated glacial lake vectorised inventories. The minimum and maximum lake area included in the collected inventories is 0.003 sq. km. and 5.414 sq. km, respectively. Table 2 lists the characteristics of each Landsat-8 band as described in EEDC (2021). The tiles were extracted via a GEE javascript, as amended from a previous script found in Aryal (2020). All bands shown in Table 2 were resampled at a 30 m spatial resolution. An initial visual inspection ensured that there were no clouds over the glacial lakes for each of the 25 tiles that were used in the presented experiments. All Landsat images except for one have less than 10% cloud coverage, as can be seen in Table 5 of the Appendix. Fig. 10 in the Appendix shows that no clouds are observed over glacier lake boundaries, even for the Landsat image with the percentage of highest cloud coverage.

3. Methodology

3.1. Data pre-processing

Well-implemented data pre-processing is an important step in the machine learning process that can eliminate redundant information or unreliable data, and is beneficial to the deep learning model to aid interpretation of the processed outputs. The preprocessing techniques adopted in our model can be concisely summarised as follows:

- **Conversion:** Vector shapefiles are converted into image masks to serve as the supervision of the model training process.
- **Rasterisation:** Conversion of vector shapefiles into raster image format.
- **Slicing:** Aims to slice the raw tiles and the corresponding mask imagery into patches with a size of 256 × 256 pixels.
- **Filtering:** Sliced patches are filtered according to the predefined area threshold of the glacial lake.
- **Substitution:** All NaN pixel values in the filtered patches are input with 0.
- **Normalisation:** To facilitate model convergence during training, normalisation is computed across all bands. All training samples are first stacked to generate the mean and standard deviation for each dimension. For each dimension of a given input patch, the difference between that dimension and the produced mean and the quotient of the generated standard deviation is then sequentially calculated.

The image patches and corresponding masks obtained through the above manipulations can be randomly selected in pairs and used as training, validation and test data, respectively.

3.2. Second-order Attention Network (SoAN)

The overall architecture of the proposed Second-order Attention Network (SoAN) is illustrated in Fig. 2.

3.2.1. Network backbone

The backbone network adopted for semantic segmentation of the glacier lake is HarDNet-MSEG (Huang et al., 2021). In particular, this approach is superior to many well-known methods in terms of accuracy and inference speed, such as DeepLabv3+ (Chen et al., 2018), U-Net (Ronneberger et al., 2015) and its variants (Huang et al., 2021). As shown in Fig. 2, the backbone network consists of the HarDNet blocks (Chao et al., 2019) (the predecessor of HarDNet-MSEG Huang et al. (2021)), RFB, dense aggregation scheme and several basic units of convolution and pooling layers.

The encoder part of the network is primarily composed of Harmonic Densely Connected blocks (Chao et al., 2019), together with several commonly used convolution and pooling layers. The HarDNet block, as the basic unit of the Harmonic Densely Connected Network (HarDNet) (Chao et al., 2019), reduces the number of shortcuts between different convolution blocks, and instead increases the number of channels in the key layers, thus achieving a balance between computational memory and accuracy. Furthermore, it introduces several 1 × 1 convolution layers as transitions in appropriate locations to ensure that the computational density will not decrease. Unlike the most well-known U-Net (Ronneberger et al., 2015) based segmentation network, the foremost components of the decoder component are designed to decode in the Cascade partial and dense aggregation way. It applies the RFB modules at several deeper layers, namely, a multi-branch network with different kernel size convolutions and dilated convolution layers to effectively enlarge the receptive fields of feature maps in various resolutions. Afterwards, upsampling based on bilinear interpolation is employed to adjust feature maps of different resolutions to an identical scale. Finally, the element-wise matrix multiplication is used to aggregate features that has been adapted to the same scale to form the final representation.

3.2.2. Second-order Attention Module (SoAM)

The design of HarDNet-MSEG (Huang et al., 2021) ensures that the

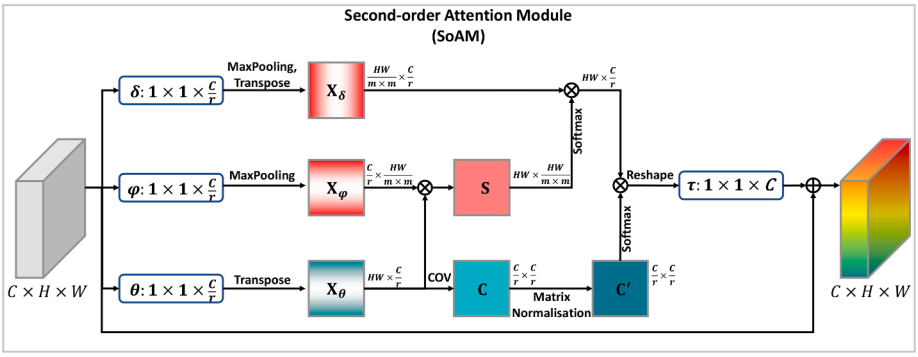

Fig. 3. Overview of the proposed Second-order Attention Module. Where $\otimes$ and $\oplus$ denote the matrix multiplication and the element-wise sum operation. COV is the process of calculating the covariance matrix. C, H and W represent the channel, height and width of the input feature, respectively. r and m indicate the factor that decreases the number of feature channels and the stride of the Maxpooling operation.

network obtains higher segmentation accuracy at a faster inference speed. However, the feature receptive field increased by adjusting the network structure is not adequate to effectively capture the long-range dependence between local features, and it completely neglects the correlations between feature channels and spatial information. Inspired by the recent successful use of attention mechanisms (Vaswani et al., 2017; Wang et al., 2018), a novel Second-order Attention Module (SoAM) is introduced, incorporating second-order statistics into the non-local neural network (Wang et al., 2018). This approach ensures the modeling of long-range dependence of the position in the feature map and effectively captures the correlation between the feature channels and their spatial information.

Details of the proposed SoAM are in Fig. 3. Specifically, given an input feature denoted by $\mathcal{X}_{in} \in C \times H \times W$ (where C is the number of feature channels, H and W represent the spatial height and width of the input tensor), it can be written simply in a matrix form by collapsing the H and W to HW. Then, 1×1 convolution layers are adopted (i.e., $\delta(\cdot)$, $\phi(\cdot)$ and $\theta(\cdot)$), the actual operations of embedded Gaussian functions introduced in Wang et al. (2018), to project the number of channels C to C/r, where r is the reduction factor. The results obtained selectively undergo subsampling and transpose operations to obtain representations of different spatial resolutions, which can be denoted as $\mathbf{X}_{\delta}, \mathbf{X}_{\phi}$ and $\mathbf{X}_{\theta}$in turn and hold the property of $\mathbf{X}_{\phi} = \mathbf{X}_{\delta}^{\top}$.

The representation $\mathbf{X}_{\theta}$generated through the function θ can be used to play two separate roles, namely, the spatial attention and the channel attention. Specifically, the spatial attention map $\mathbf{S}$ can be obtained using:

$$\mathbf{S} = \mathbf{X}_{\theta} \otimes \mathbf{X}_{\phi}, \tag{1}$$

where the shape of $\mathbf{S}$ comes into $HW \times \frac{HW}{m \times m}$. The factor m is the stride of subsampling (i.e., the maxpooling operation). By using this operation, it not only reduces the amount of pairwise calculations by $\frac{1}{m \times m}$, but also makes the computation sparse and incorporates the feature discrepancy at different spatial resolutions.

The channel attention map is $\mathbf{C}$acquired by collecting the second-order statistics of $\mathbf{X}_{\theta}$. Formally, it can be expressed as:

$$\mathbf{C} = \mathbf{X}_{\theta}^{\top} \bar{\mathbf{I}} \mathbf{X}_{\theta}, \tag{2}$$

where $\bar{\mathbf{I}} = \frac{1}{HW}\left(\mathbf{I} - \frac{1}{HW}\mathbf{1}\mathbf{1}^{\top}\right)$with the shape of $\mathbb{R}^{HW \times HW}$, the $\mathbf{I}$ and $\mathbf{1}$ denote the identity matrix and the all-ones matrix, respectively.

Since the produced matrix $\mathbf{C} \in \mathbb{R}^{\frac{C}{r} \times \frac{C}{r}}$ is a symmetric positive semi-definite (SPD) matrix lying in the space of Riemannian manifold denoted by Sym^{+}, the manifold needs to be flattened into Euclidean space so that commonly used metrics can be adopted to measure the distance between different projected elements. The natural choice to preserve the loss of geometric structure during the projection is to compute the logarithm of the covariance matrix $\mathbf{C}$ because it can endow the Riemannian manifold of SPD matrices with a Lie group structure. However, the logarithm metric is often numerically unstable in the practise since it may change the magnitudes of small eigenvalues considerably, and potentially reversing the order of eigenvalue significance (Li et al., 2017). The matrix square root is an alternative approach to approximately measure the geodesic distance of the Riemannian manifold that is more stable due to the allowance of non-negative eigenvalues, and is adopted in this case to improve the numerical stability.

Given an SPD matrix, there is a unique square root, which can be accurately calculated by EIG or SVD. Specifically, the eigenvalue decomposition of matrix $\mathbf{C}$ can be written as: $\mathbf{C} = \mathbf{U}\mathrm{diag}(v_i)\mathbf{U}^{\top}$, where $\mathbf{U}$ is orthogonal matrix and $\mathrm{diag}(v_i)$ is a diagonal matrix of eigenvalues. Then, the square root of $\mathbf{C}$ can be denoted by $\mathbf{C}^{'} = \mathbf{U}\mathrm{diag}\left(v_i^{\frac{1}{2}}\right)\mathbf{U}^{\top}$, namely $\mathbf{C}^2 = \mathbf{C}^{'}$. However, the computation of both SVD and EIG functions are not well-supported by GPU (Wang et al., 2020; Wang et al., 2021; Li et al., 2017; Li et al., 2018). Moreover, as the covaraince matrix has a significant risk of being degenerated in practise (i.e., one or more of its normalised eigenvalues may be identical), a small ridge term η is added to preserve the singularity of matrix, as described in Wang et al. (2020, 2021). The process can be denoted as:

$$\Sigma = \mathbf{C} + \eta \mathrm{trace}(\mathbf{C})\mathbf{I}, \tag{3}$$

where $\mathrm{trace}(\cdot)$ is the matrix trace operation and $\mathbf{I} \in \mathbb{R}^{\frac{C}{r} \times \frac{C}{r}}$ is an identity matrix. The matrix Σ is a proxy matrix for simplifying the iteration process, whereby the Newton-Schulz Iteration method (Li et al., 2018) is applied to solve the matrix square root and the corresponding coupled iterations can be represented as:

$$\begin{aligned} \mathbf{P}_k &= \frac{1}{2}\mathbf{P}_{k-1}(3\mathbf{I} - \mathbf{Q}_{k-1}\mathbf{P}_{k-1}) \\ \mathbf{Q}_k &= \frac{1}{2}(3\mathbf{I} - \mathbf{Q}_{k-1}\mathbf{P}_{k-1})\mathbf{Q}_{k-1}, \end{aligned} \tag{4}$$

where the iteration starts with setting two proxy matrices $\mathbf{P}_0$ and $\mathbf{Q}_0$ to $\frac{\Sigma}{\mathrm{trace}(\Sigma)}$ and $\mathbf{I}$, respectively. For $\mathrm{k} = 1, \cdots, \mathrm{N}$ is the number of iterations. The adopted iterative method only involves the matrix product, which guarantees to be parallel optimsed on GPU. As the $\mathbf{P}_0 = \frac{\Sigma}{\mathrm{trace}(\Sigma)}$ is initialised to satisfy the convergence condition of the coupled iteration alters the data magnitudes in a non-trivial manner that needs to be considered to counteract such change. Then, the compensation term recommended by Li et al. (2018) is adopted and results in the normalised matrix $\mathbf{C}^{'}$. Formally, it is computed by:

$$\mathbf{C}^{'} = \sqrt{\mathrm{trace}(\Sigma)}\mathbf{P}_N. \tag{5}$$

After the channel attention map $\mathbf{C}^{'}$ has been generated, it can be fused with the obtained spatial feature map through the matrix multiplication

Table 3
Comparison of the proposed Second-order Attention Network (SoAN) with baseline methods (full 17 bands). Note that w/ and w/o denote with/without matrix normalisation.

Methods	IoU	Dice (F1)	Precision	Recall	F2 Score	#Model Size (MB)	Inference Time(sec/per image)
U-Net	64.56	78.21	71.39	86.66	83.05	**37.55**	**0.0050**
HarDNet-MSEG	65.52	78.87	73.83	84.72	82.27	70.40	0.0150
HarDNet-MSEG (Non-local)	65.62	78.98	71.00	**89.15**	84.76	79.18	0.0157
SoAN (w/o Normalisation)	67.23	80.23	**75.97**	85.08	83.06	79.18	0.0165
SoAN (w/ Normalisation)	**68.33**	**81.02**	74.96	88.19	**85.17**	79.18	0.0183

product to present a form of self-attention mechanism. Formally, it can be expressed as:

$$\text{score} = \text{softmax}(\mathbf{S}) \otimes \mathbf{X}_{\delta} \otimes \text{softmax}(\mathbf{C}^{'}), \tag{6}$$

where $\text{softmax}(\cdot)$denotes the Softmax function. The reshaping manipulation is performed on the joint attention score in order to facilitate further integration with the input feature $\mathcal{X}_{in}$ through the residual connection method. Finally, a learnable transformation function $\tau(\cdot)$ is employed to restore the channel dimension of attention maps from $\frac{C}{r}$ to C, followed by a batch normalisation layer. The complete form of the proposed Second-order Attention Module (SoAM) can be defined as:

$$\mathcal{X}_{out} = \mathcal{X}_{in} + \tau(\text{score}). \tag{7}$$

Although the idea of the SoAM is inspired by non-local neural networks (Wang et al., 2018), there exist significant differences between the two methods. Firstly, it increases the variability of the spatial attention map by imposing the subsampling operators on two of the mapping functions (i.e., the maxpooling operations followed by the $\delta(\cdot)$ and $\phi(\cdot)$ functions). Secondly, it realises the channel attention by computing the $\text{softmax}(\cdot)$ of the second-order statistics of the projected feature $\mathbf{X}_{\theta}$. Thirdly, it provides two variants for gathering the second-order features (i.e., the covariance matrix and the normalised form with matrix square root normalisation) that are computationally efficient. In addition, the proposed SoAM can be flexibly embedded into various position of the given network. Inserting SoAM into the shallow layer of the network has the potential to capture more spatial information due to the relatively high spatial resolution of shallow features. However, plugging the SoAM at the shallow layer can slightly impair performance in practice. The reason for this phenomenon may be insufficient discrimination of shallow features. Considering the fact that the information stored in the feature channels is more dominant than the spatial location and the spatial details of the shallow information can also be well preserved by the deep information, the proposed SoAM is plugged into the place corresponding to each RFB module (seen in Fig. 2)) to ensure that the attention module can effectively extract discriminative features.

3.3. Loss function

The loss function defines how the neural network calculates the overall error from the residual of each training batch. This in turn will affect how the loss function effectively adjusts the coefficients when performing backpropagation. Here the objective function adopted is the standard Dice loss (Milletari et al., 2016; Sudre et al., 2017) that is used for measuring the overlaps and is widely used to evaluate the segmentation performance when ground truth is available. Let $\widehat{Y}$ be the predicted probabilistic map for the foreground label over N image elements $\widehat{y}_n$ (the predicted background class probability being 1-$\widehat{Y}$) and Y is the reference foreground segmentation with voxel value y_n. The generalised form of 2-class of Dice loss (Sudre et al., 2017) can be denoted as:

$$\mathcal{L}_{Dice} = 1 - \frac{\sum_{n=1}^{N} \widehat{y}_n y_n}{\sum_{n=1}^{N} \widehat{y}_n + y_n}, \tag{8}$$

4. Experiments

4.1. Implementation details

As mentioned in Section 2, 25 raw Landsat-8 tiles are queried using a GEE javascript. Apart from the raw Landsat-8 bands (Table 2), the NDSI, NDWI, and the NDVI indices, as well as the elevation retrieved from the Shuttle Radar Topography Mission (SRTM90) and slope computed from the SRTM90 elevation are added to the tiles and all resampled to 30 m spatial resolution. All the tiles are prepossessed following the steps introduced in Section 3, which includes the coversion, rasterisation, slicing, filtering (the filter percentage set to $5e^{-3}$), substitution and normalisation. In particular, both tile images and vector shapefile were sliced into patches with 256 × 256 pixels (641 patches in total). The pre-processed image and mask pairs are randomly split into training, validation and test sets in the ratio of 70%, 10% and 20%, respectively. During training, the random flips horizontally and vertically, random rotations of 90 degrees zero or more times are adopted to augment the training samples. Note that all experiments are conducted on a PC with a single GeForce RTX 2080 Ti GPU.

Experiments are conducted with two novel attentive variants (i.e., SoAN with/without matrix square-root normalisation) based on the proposed SoAM and compared with three existing baseline architectures (i.e., U-Net; Ronneberger et al. (2015), HarDNet-MSEG; Huang et al. (2021) and HarDNet-MSEG with non-local blocks). The output of the first convolution layer is adjusted to 18 in order to satisfy the conditions of processing multiple bands of Landsat-8 imagery. The architecture of the adopted U-Net has been utilised by Baraka et al. (2020) for the glacier mapping, which includes 5 downsampling layers followed by 5 upsampling layers with a bottleneck layer in between. The remaining experiment architecture is based on the HardNet-MSEG (Huang et al., 2021), while the fully-deconvolution layers are removed in order to increase the inference speed. As described in Baraka et al. (2020), the Adam optimiser is employed to optimise the networks. The initial learning rate is set to $1e^{-4}$ with a weight decay of $5e^{-4}$. The $\mathcal{L}_1$ regularisation with a factor of $5e^{-4}$ is applied to prevent the model from overfitting. The ReduceLROnPlateau scheduler with a *patience* of 10 and a reduce factor of 0.1 is used to reduce the learning rate if the quality of metrics read by the scheduler does not improve for a *patience* number of epochs. The batch size is set to 32 and the models are trained for 300 epochs for all experiments. The stride m and intermediate channel rate r mentioned in the SoAM are set to 2. The η in Eq. (3) is set to $1e^{-2}$. In addition, the iteration number k in Eq. (4) is set to 5 based on experiments (see Section 4.2.4). Note that the proposed attentive models can be efficiently trained in an end-to-end manner.

The metrics adopted to compare the effectiveness of the proposed models and baseline methods include:

$$\text{Dice} = \text{F1} = \frac{2 * tp}{2 * tp + fp + fn}, \quad \text{IoU} = \frac{tp}{tp + fp + fn}, \quad \text{Recall} = \frac{tp}{tp + fn},$$
$$\text{Precision} = \frac{tp}{tp + fp}, \quad \text{F2} = \frac{5\text{Precision} * \text{Recall}}{4\text{Precision} + \text{Recall}}, \tag{9}$$

where tp, fp, tn and fn denote the number of true positives, false

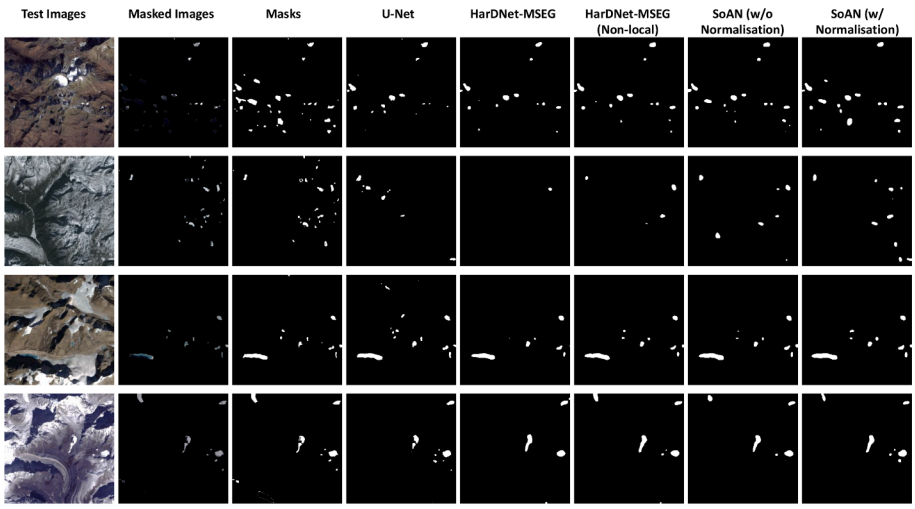

Fig. 4. Visualisation of testing images using different methods.

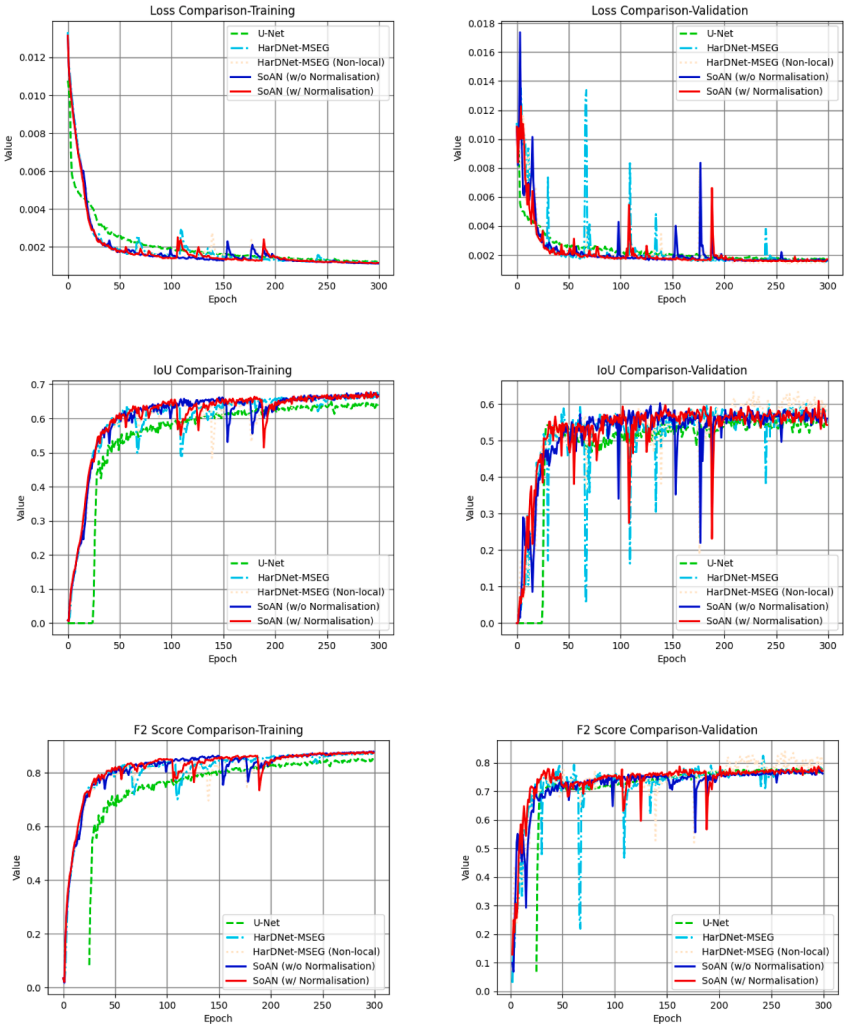

Fig. 5. Comparison of different models in terms of loss, IoU and F2 Score during training and validation.

positives, true negatives and false negatives, respectively. Note that the F1 score is numerically equivalent to the Dice score in the scenario of binary segmentation. In addition to these metrics, the comparisons also include the inference time for a single image and the model size that is computed by: $\frac{\#\text{Parameters}\times 4}{1024\times 1024}$.

4.2. Experimental results

4.2.1. Evaluation on Landsat-8

Comprehensive experiments were conducted on the pre-processed dataset to thoroughly compare the performance of the five deep

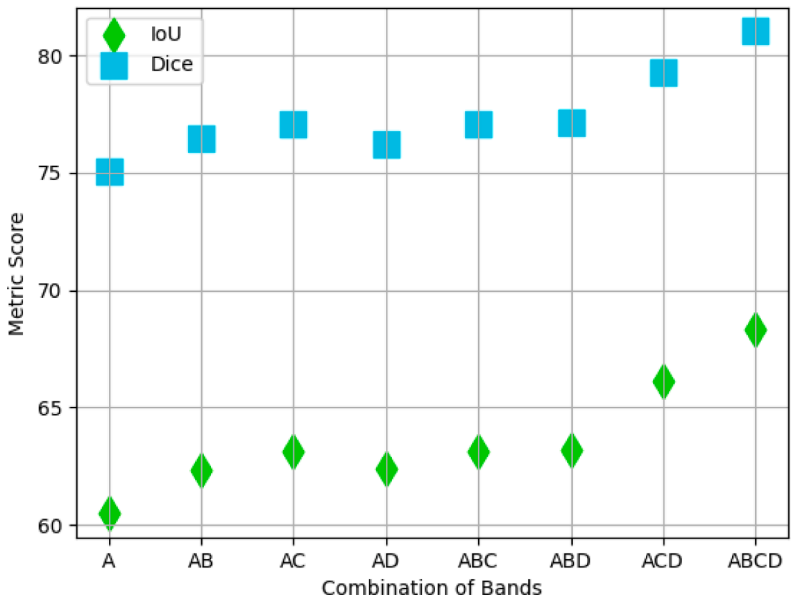

Fig. 6. Experimental results for band selection. All bands are divided into four groups according to their characteristics and are represented by A ('B1'-'B11'), B ('sr_aerosol', 'pixel_qa' and 'radsat_qa'), C ('ndvi', 'ndsi' and 'ndwi') and D ('elevation' and 'slope'), respectively.

Table 4
The performance was obtained by plugging the proposed SoAM into different positions of the adopted backbone.

	Plugging in SoAM after n-th HarDBlock				
	5^{th}	$5^{th}, 4^{th}$	$5^{th}, 4^{th}, 3^{rd}$	$5^{th}, 4^{th}, 3^{rd}, 2^{nd}$	$5^{th}, 4^{th}, 3^{rd}, 2^{nd}, 1^{st}$
IoU	64.79	65.39	68.33	65.51	62.48
Inference Time (sec/per image)	0.0158	0.0168	0.0183	0.0204	0.0214

learning models on the glacier lake segmentation task. As shown in Table 3, the methods listed for comparison include the standard segmentation network U-Net (Ronneberger et al., 2015), the latest HarDNet-MSEG (Huang et al., 2021), the original HarDNet-MSEG with non-local attention block, and two second-order attention networks (SoAN with/without the square-root matrix normalisation). HarDNet-MSEG with non-local blocks first calculates the matrix product of $\mathbf{X}_\theta$ and $\mathbf{X}_\phi$, and then multiplies $\mathbf{X}_g$ to obtain the final attention score. For SoANs, the second-order statistics $\mathbf{C}$ of unsampled local feature $\mathbf{X}_\theta$are introduced on the basis of covariance of $\mathbf{X}_\theta$to effectively capture the correlations between the channels and the feature spatial information. The generated $\mathbf{C}$ or its square-root normalisation form of $\mathbf{C}^{'}$ can be treated as the attention maps of feature channels and integrated into the spatial-based attention map to re-weight the input feature $\mathcal{X}$. In Table 3, the proposed SoANs achieve the best results compared to all listed methods. In particular, the IoU obtained by SoAN with/without matrix normalisation can exceed the U-Net (Ronneberger et al., 2015) and HarDNet-MSEG (Huang et al., 2021) by approximately 3%. By imposing matrix normalisation, F1 (Dice) and F2 scores of the proposed SoAN can reach 81.02% and 85.17% respectively. In comparison of HarDNet-MSEG (Huang et al., 2021) with U-Net (Ronneberger et al., 2015), the relative gains of IoU, Dice and F2 score reflect not only the significance of convolutional layer superposition, but also the profit from the

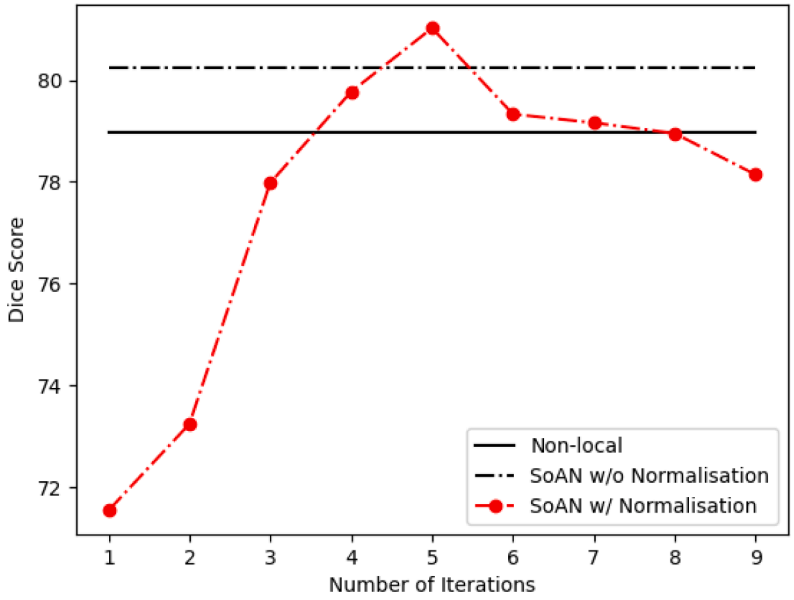

Fig. 7. The effect of the number of Newton-Schulz iterations in SoAM on the performance.

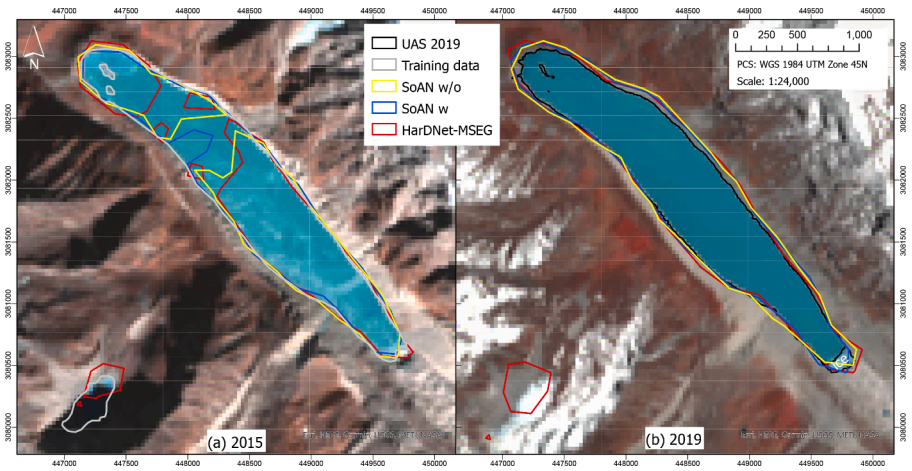

Fig. 8. Visual inspection of the Tsho Rolpa glacial lake boundary as extracted using the three methods HardNet-MSEG, SoAN w and SoAN w/o normalisation, as shown in red, blue and yellow respectively, using (a) the 19/12/2015 training image and (b) the 20/05/2019 Landsat-8 image. The corresponding boundary used for training is shown in grey. The UAS 2019 boundary shown in black is depicted for comparative purposes.

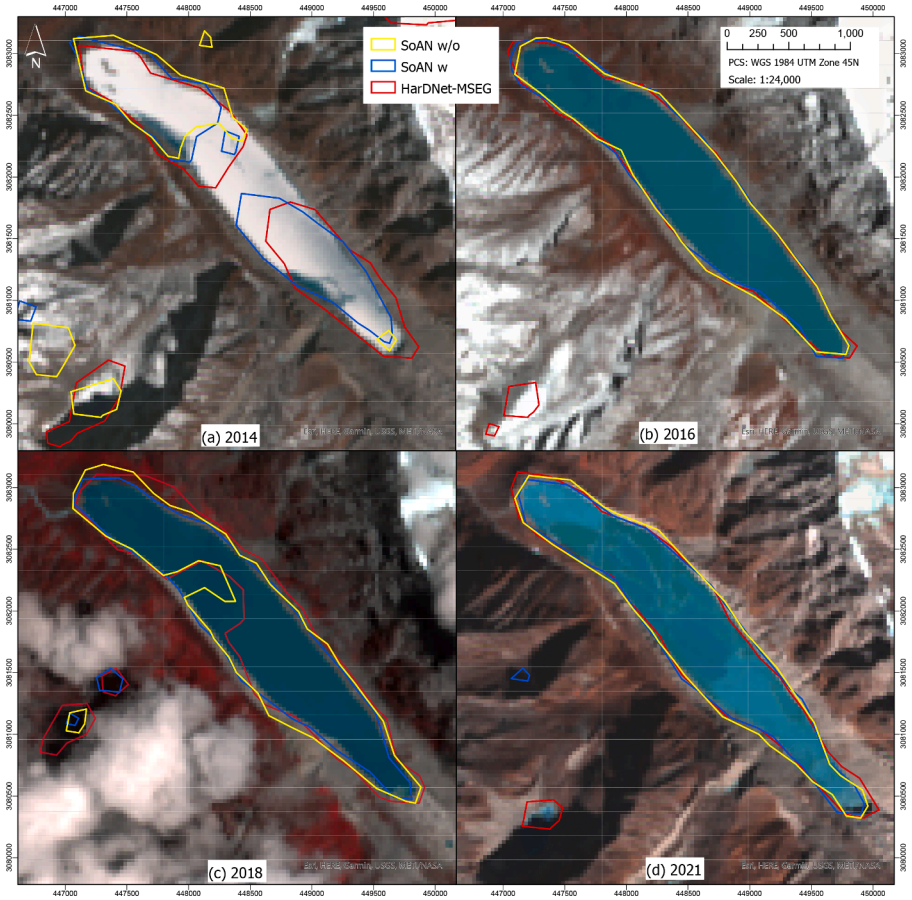

Fig. 9. Visual inspection of the Tsho Rolpa glacial lake boundary as extracted using the three methods HardNet-MSEG, SoAN w and SoAN w/o normalisation, as shown in red, blue and yellow respectively, using (a) the 30/01/2014, (b) the 24/04/2016, (c) the 04/07/2018 and the (d) 01/01/2021 Landsat-8 images.

enlargement of the feature receptive fields thanks to the ingeniously designed network structure. By comparing HarDNet-MSEG and the proposed SoANs, the further improvements of results can be attributed to the effectiveness of the novel SoAM. In terms of the inference time per image and model size, U-Net performs best (i.e., 0.0050 and 37.55 MB) because of its relatively shallow structure. The two proposed SoAN variants can achieve similar inference times to non-local based HardNet-MSEG with the same model size.

In addition to the quantitative results described above, qualitative visualisation of different methods can be seen in Fig. 4. Through the comparison of the patches, it can clearly be seen that, compared with the predictions of U-Net (Ronneberger et al., 2015) and HarDNet-MSEG (Huang et al., 2021), the contours of the glacier lake predicted by the proposed attention-based model are closer to the ground-truth labels. Furthermore, the over-estimated predictions rarely occur after introducing attention mechanisms based on either non-local blocks or second-order statistical attention modules. These results undoubtedly prove the effectiveness of the network for segmenting glacial lakes.

To comprehensively compare the performance of different models, curves of three metrics during training and validation are shown in Fig. 5. For a fair comparison, all hyperparameters of different models are kept constant during the experiment. By comparing the training losses of different models, the models based on the HardNet-MSEG backbone converge faster and lower than the U-Net. Changes in IoU and F2 scores during training also confirm the effectiveness of HarDNet-MSEG models, while the proposed variant of SoAN with square root normalisation performs the best on close examination. Nevertheless, the performance of different models on the validation set fluctuates significantly, which may be due to the relatively small size and randomness of the data.

4.2.2. Test on band combinations

To investigate the effect of using a set of specific bands or different combinations of bands, the input bands are divided into four different groups according to their properties, namely, A ('B1'-'B11'), B ('sr_aerosol', 'pixel_qa' and 'radsat_qa'), C ('ndvi', 'ndsi' and 'ndwi') and D ('elevation' and 'slope'). Then, on the basis of the Landsat-8 band (i.e. group A), additional groups were successively introduced and tested on the model of SoAN with matrix square root normalisation for their gains to the results. As can be seen from Fig. 6, the IoU and Dice scores of introducing group C are higher than adding groups B and D. The ACD group performed the best in a combination of three different groups. The best results are obtained when using all bands, followed by the combination of ACD with slight gaps. In summary, the introduction of Group C has a more pronounced effect on the improvement of segmentation performance and the best results are obtained using all available bands.

4.2.3. Effect of embedding SoAM in different positions

The proposed SoAM, as a standalone module based on attention mechanism, is embedded in three different places of the backbone network (shown in Fig. 2), being compatible with the original HarDNet-MSEG framework. The design of the SoAM guarantees that the dimensions of the input and output tensors are identical, which means this module can be flexibly plugged into different parts of the backbone. Since the main component of the SoAN encoder part is the HarDNet Block, SoAM is considered to be attached after different HarDNet Blocks to maximise its capabilities. As can be seen from Table 4, the best IoU is obtained by appending the SoAM after the 3^{rd}, 4^{th} and 5^{th} HarDNet Blocks, while the lowest IoU appears when plugging SoAM after each HarDNet Block. The reason for the difference may be that the multi-scale features extracted by shallow layers are not effectively transited to the decoder part by the RFB module as after the 3^{rd}, 4^{th} and 5^{th} HardNet Blocks. In addition, as stated in the paper of the HarDNet-MSEG, shallow features are discarded because deep features can also represent the spatial details of shallow information relatively well and contain extensive channel information, thereby reducing computational cost.

4.2.4. Effect of iterations in matrix normalisation

Since the solution of the square root of the matrix is based on an iterative method (Newton-Schulz iterations in Eq. 4), the optimal number of iterations need to be investigated experimentally. Fig. 7 shows the effect of different iterations on the Dice score, and compares three different approaches (i.e., Non-local, with/without square-root normalisations) with the same backbone (i.e., HarDNet-MSEG). Specifically, as the number of iterations increases, the Dice score rises rapidly and then decreases slowly, reaching a peak at the 5th iteration. The proposed SoAN with square root normalisation can outperform the Non-local method after 4 iterations. Besides, the Dice score drops consistently after 5 iterations, indicating that increasing the number of iterations is not helpful for improving the performance. As a result, iteration number k is set to 5 in all experiments to balance efficiency and performance.

4.2.5. Further evaluation on Tsho Rolpa glacial lake

To further evaluate the segmentation results, a glacial lake that has been identified as potentially dangerous, Tsho Rolpa (Longitude: 86.4754, Latitude: 27.8627), is taken as an example. The boundaries of Tsho Rolpa determined using the HardNet-MSEG, SoAN w and SoAN w/o normalisation methods, are extracted and smoothed to remove jagged edges (due to the large ground pixel sizes), using the python environment in ArcGIS Pro. They are then visually compared with the manually labelled boundary shown in Fig. 8(a). It can be seen that the boundaries from the two SoAN models (in blue and yellow colours) match relatively well with the labelled boundary (in grey) but with a single area in the middle miss-detected, whereas the HardNet-MSEG boundary (in red) has many areas of miss-detection. In addition, the surface areas of all four boundaries were calculated in ArcGIS Pro as follows: a) the area of the labelled boundary is 1.606 km^2; the area of the SoAN w/o (in yellow) boundary is 1.041 km^2; the area of the SoAN w (in blue) is 1.536 km^2; and the area of the HardNet-MSEG (in red) boundary is 1.273 km^2. Area calculation showed that SoAN w/o provided a result closer to the labelled reference data. Between the two SoAN models, the SoAN w resulted in one single boundary compared to the SoAN w/o outcome. The image data was acquired in December when the lake was covered by ice, which seemingly caused the error.

As independent comparison, the lake was surveyed by a fixed-wing UAS (eBee; SenseFly (2020)) which captured high resolution images carrying the Sony Cybershot DSC-WX220 digital camera of size 4896 × 3672 pixels on 16 May 2019. The UAS-acquired images were orientated, matched and georeferenced to generate an orthomosaic of the entire lake. More information about the 2019 UAS-acquired datasets can be found in Maharjan et al. (2019). The boundary extracted by the UAS-generated orthomosaic (depicted in black in Fig. 8(b)) is also compared with the boundaries from the three deep learning models using Landsat-8 imagery from the closest dates in May 2019, when no ice is observed on the water surface with the exception of the region close to the glacier terminus. As seen from Fig. 8(b), all three boundaries match well with the high-resolution UAS boundary in general, but they are all slightly larger. In particular, the surface areas of the UAS 2019, SoAN w/o, SoAN w and HarDNet-MSEG boundaries are 1.604 km^2, 1.850 km^2, 1.776 km^2, and 1.808 km^2 respectively. This discrepancy can be attributed to the large 30 m pixel size of the Landsat-8 images compared to UAS orthomosaic with 0.12 m pixel size, and the fact that the majority of the boundary pixels are classified as water. Note at the end of the lake (bottom right in Fig. 8(b)), where the glacier terminus has created some icy debris over the lake, satellite data are able to classify most of the area as part of the lake's water surface.

One key application of a fully automated glacial lake segmentation is to create time-series of lake boundaries to understand the change over time. Four images over the years from 2014 to 2021 are extracted to illustrate the impact of various image conditions. As shown in Fig. 9(a), icy surface, especially when covered by frost (as seen in white) can strongly hinder the segmentation. Whereas, a clear water surface in a warmer season can produce more consistent results, for all the three models (Fig. 9(b)). In addition, satellite image pixel quality also affects the segmentation, regardless of the season as illustrated in Fig. 9(c). The pixel quality is low due to clouds over southwest of the glacial lake, so are the derived band indices. Despite that, the SoAN w model produced the most complete boundary among the three deep learning models. Interestingly, Fig. 9(d) depicts consistent segmentation results during winter time, even with the presence of ice. However, the HarDNet-MSEG boundary (in red) is slightly overestimated in contrast to the two SoAN boundaries. The presented results have demonstrated that lake conditions and image quality can significantly influence the segmentation. It should be noted that, apart from the environmental conditions, the limited amount of training data can also be a potential alternative source of error. Therefore, when training for time-series analysis, it is important to take these factors into consideration and ideally use satellite imagery acquired during optimal (and consistent) environmental conditions.

Moreover, it should be noted that the presented study is based on

Landsat SR products which have well-known issues on very low reflecting surfaces (such as waterbodies and/or shadows) as well as on very high reflecting surfaces such as snow (USGS, 2022), in comparison to TOA products. A further comparative analysis that applies the proposed models on both SR and TOA Landsat imagery from the same acquisition time would aid understanding as to whether erroneous segmentation results are attributable to the limitations of the SR products or the challenging environmental conditions.

Another consideration is the overlaps between adjacent image scenes. In the experimental data, the lateral overlaps (sidelaps) have a time difference of at least six days. This means the two overlapping images are different, due to changes in atmospheric and illumination conditions. The forward overlaps of two consecutive images acquired on the same day have smaller time differences in minutes. In the data, we observed differences in the pixel value of the overlapped region for each image due to different viewing angles and shadows in the mountainous regions surrounding the glacial lakes. In addition, when the sliding window method is applied to the overlapping region between two adjacent images, the resulting patches will contain different spatial information due to the different starting points for processing the overlapping region of two images. In this way, the generated patch samples cover different contents, and each of them is then treated as an independent sample.

5. Conclusion

This study presents a new deep learning network to segment lakes from Landsat-8 imagery. The SoAN includes a novel second-order attention module that incorporates the collection of second-order statistics into the non-local neural network, to effectively capture the correlation between the feature channels and their spatial information. Comparisons with the commonly used U-Net and state-of-the-art HarDNet-MSEG approaches demonstrated the proposed attention module can produce superior results, though can be further improved, mostly in terms of precision. The test on band combinations has shown that indices derived from the image bands can help improve the segmentation, as do elevation and slope. Therefore, it is important to obtain those associated data wherever possible. Further detailed assessment of the results on a particular example lake in the HKH has highlighted significant influence of lake conditions, e.g. image quality and ice cover.

In relation to the technical aspect of the developed model, future work will be focused on automatic lake condition and image quality filtering so that more consistent results can be expected. Regarding the implementation of the proposed model on mapping multi-temporal glacial lake dynamics, it would ideally require segmentation of the lake surface into ice and water. The model could then be trained to incorporate the temporal component under different environmental conditions. In this way it would be feasible to derive a further understanding of the lake conditions over time. Then, ultimately the developed SoAN could support the ongoing glacial lake inventory documentation in the HKH region over multiple years and under different environmental conditions. In addition, as non-glacial lakes might exist in the HKH high mountain region, this could hinder the performance of the developed model. For future considerations and improvements it would be interesting to have a further class of non-glacial lakes provided that accurate labels could be obtained with the aid of expert knowledge. Given the availability of higher resolution satellite data, such as Sentinel-2 and PlanetScope imagery (Qayyum et al., 2020), transferring the developed model to other data sources would also be of interest.

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.

Acknowledgements

The authors would like to acknowledge the valuable comments from the reviewers. This research was supported by the UK Natural Environment Research Council (NERC) SHEAR Catalyst project “WeACT”, Web-Based Natural Dam-Burst Flood Hazard Assessment and Forecasting SysTem (NE/S005919/1); by the NERC project “PYRAMID”, Platform for dYnamic, hyper-resolution, near-real time flood Risk AssessMent Integrating repurposed and novel Data sources (NE/V00378X/1); and by the Water Security and Sustainable Development Hub which is funded by the UK Research and Innovation (UKRI) Global Challenges Research Fund (GCRF) (ES/S008179/1).

Appendix A

Fig. 10
Table 5

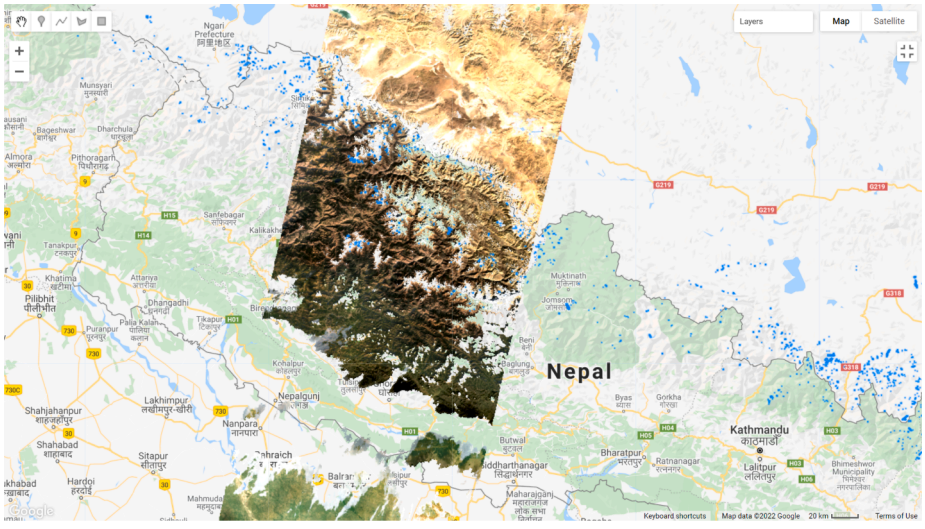

Fig. 10. LC08_143039_20151208 and LC08_143040_20151208 Landsat image tiles with clouded regions masked out. The glacier lake boundaries in blue are superimposed over the tiles. No glacial lakes can be found below the clouded regions. Figure was generated via GEE.

Table 5
25 Landsat-8 image tiles used in the presented experiments as retrieved by ICIMOD.

Landsat image	Cloud Cover [%]
LC08_140041_20151219	4.41
LC08_141040_20151007	1.44
LC08_141040_20151124	5.74
LC08_141040_20151226	0.85
LC08_142040_20151201	1.66
LC08_142040_20151217	1.54
LC08_142040_20160102	1.63
LC08_143039_20151122	4.11
LC08_143039_20151208	2.31
LC08_143040_20151021	4.65
LC08_143040_20151208	14.07
LC08_143040_20151224	2.09
LC08_144039_20150910	4.19
LC08_144039_20151113	4.53
LC08_144039_20151129	3.40
LC08_145039_20151120	3.45
LC08_145039_20161106	2.32
LC08_139040_20161214	0.22
LC08_140040_20160104	0.64
LC08_141040_20180116	0.90
LC08_139040_20151228	0.14
LC08_139041_20151228	2.13
LC08_139041_20171217	2.49
LC08_140040_20151219	0.69
LC08_140041_20151219	4.41
LC08_140041_20161205	3.27

References

Aryal, B., 2020. Query Landsat-7 tiles using GEE. URL: https://github.com/Aryal007/GEE_landsat_7_query_tiles/commits?author=Aryal007.

Bajracharya, S.R., Maharjan, S.B., Shrestha, F., Bajracharya, O.R., Baidya, S., 2014. Glacier Status in Nepal and Decadal Change from 1980 to 2010 Based on Landsat Data. Technical Report International Centre for Integrated Mountain Development and United Nations Development Programme (UNDP).

Bajracharya, S.R., Maharjan, S.B., Shrestha, F., Sherpa, T.C., Wagle, N., Shrestha, A.B., 2020. Inventory of glacial lakes and identification of potentially dangerous glacial lakes in the Koshi, Gandaki, and Karnali River Basins of Nepal, the Tibet Autonomous Region of China. Technical Report International Centre for Integrated Mountain Development and United Nations Development Programme (UNDP).

Baraka, S., Akera, B., Aryal, B., Sherpa, T., Shresta, F., Ortiz, A., Sankaran, K., Ferres, J.L., Matin, M., Bengio, Y., 2020. Machine learning for glacier monitoring in the Hindu Kush Himalaya. arXiv preprint arXiv:2012.05013.

Bhardwaj, A., Singh, M.K., Joshi, P., Singh, S., Sam, L., Gupta, R., Kumar, R., et al., 2015. A lake detection algorithm (LDA) using Landsat 8 data: a comparative approach in glacial environment. Int. J. Appl. Earth Obs. Geoinf. 38, 150–163.

Blunden, J., Arndt, D.S., Bissolli, P., Diamond, H.J., Druckenmiller, M.L., Dunn, R.J.H., Ganter, C., Gobron, N., Lumpkin, R., Richter-Menge, J.A., Li, T., Mekonnen, A., Sánchez-Lugo, A., Scambos, T.A., Schreck, C.J., Stammerjohn, S., Stanitski, D.M., Willett, K.M., Andersen, A., Rosen, R., 2020. State of the climate in 2019. Bull. Am. Meteorol. Soc. 101, S1–S8. https://doi.org/10.1175/2020BAMSSTATEOFTHECLIMATE_INTRO.1.

Chao, P., Kao, C.-Y., Ruan, Y.-S., Huang, C.-H., Lin, Y.-L., 2019. Hardnet: A low memory traffic network. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3552–3561.

Chen, F., Zhang, M., Tian, B., Li, Z., 2017. Extraction of glacial lake outlines in tibet plateau using landsat 8 imagery and google earth engine. IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens. 10, 4002–4009. https://doi.org/10.1109/JSTARS.2017.2705718.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (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. https://doi.org/10.5194/tc-15-1663-2021.

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, 1–27. https://doi.org/10.3390/rs13020197.

EEDC, 2021. Description of USGS Landsat 8 Surface Reflectance Tier 1 - Earth Engine Data Catalog (EEDC). URL: https://developers.google.com/earth-engine/datasets/catalog/LANDSAT_LC08_C01_T1_SR.

Gao, B.-C., 1996. NDWI - a normalized difference water index for remote sensing of vegetation liquid water from space. Remote Sens. Environ. 58, 257–266. https://doi.org/10.1016/S0034-4257(96)00067-3.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sens. Environ. 202, 18–27. https://doi.org/10.1016/j.rse.2017.06.031.

Guo, S., Du, P., Xia, J., Tang, P., Wang, X., Meng, Y., Wang, H., 2021. Spatiotemporal changes of glacier and seasonal snow fluctuations over the namcha barwa–gyala peri massif using object-based classification from landsat time series. ISPRS J. Photogramm. Remote Sens. 177, 21–37.

Hoeser, T., Bachofer, F., Kuenzer, C., 2020. Object detection and image segmentation with deep learning on earth observation data: A Review-Part II: Applications. Remote Sens. 12 https://doi.org/10.3390/rs12183053.

Hoeser, T., Kuenzer, C., 2020. Object detection and image segmentation with deep learning on earth observation data: A Review-Part I: Evolution and recent trends. Remote Sens. 12 https://doi.org/10.3390/rs12101667.

Huang, C.-H., Wu, H.-Y., Lin, Y.-L., 2021. HarDNet-MSEG: A simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 fps. arXiv:2101.07172.

ICIMOD, 2020. Glacial lakes in the Koshi, Gandaki, and Karnali river basins of Nepal, the Tibet Autonomous Region of China, and India. ICIMOD. URL: https://doi.org/10.26066/RDS.1971946.

Ives, J.D., Shrestha, R.B., Mool, P.K., et al., 2010. Formation of glacial lakes in the Hindu Kush-Himalayas and GLOF risk assessment. Technical Report ICIMOD.

Lehner, B.G.G., 2013. Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems. Hydrol. Process. 27, 2171–2186.

Li, P., Xie, J., Wang, Q., Gao, Z., 2018. Towards faster training of global covariance pooling networks by iterative matrix square root normalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 947–955.

Li, P., Xie, J., Wang, Q., Zuo, W., 2017. Is second-order information helpful for large-scale visual recognition?. In: International Conference on Computer Vision (ICCV).

Li, R., Su, J., Duan, C., Zheng, S., 2020. Linear attention mechanism: An efficient attention for semantic segmentation. arXiv preprint arXiv:2007.14902.

Li, Y., Dang, B., Zhang, Y., Du, Z., 2022. Water body classification from high-resolution optical remote sensing imagery: Achievements and perspectives. ISPRS J. Photogramm. Remote Sens. 187, 306–327.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Maharjan, S., Joshi, S., Peppa, M., Xiao, W., Liang, Q., 2021. Digital elevation models and bathymetry data of tsho rolpa glacier lake, Nepal, 2019. doi:10.5285/8e483692-3b65-41d2-a7fd-5a3cd589a71c.

Maharjan, S.B., Mool, P., Lizong, W., Xiao, G., Shrestha, F., Shrestha, R., Khanal, N., Bajracharya, S., Joshi, S., Shai, S., et al., 2018. The Status of Glacial Lakes in the Hindu Kush Himalaya-ICIMOD Research Report 2018/1. Technical Report International Centre for Integrated Mountain Development (ICIMOD).

Milletari, F., Navab, N., Ahmadi, S.-A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). IEEE, pp. 565–571.

Mool, P.K., Maskey, P.R., Koirala, A., Joshi, S.P., Wu, L., Shrestha, A.B., Eriksson, M., Gurung, B., Pokharel, B., Khanal, N.R., Panthi, S., Adhikari, T., Kayastha, R.B., Ghimire, P., Thapa, R., Shrestha, B., Shrestha, S., Shrestha, R.B., 2011. Glacial lakes and glacial lake outburst floods in Nepal. ICIMOD report. doi:978 92 9115 193 6.

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.

Pekel, J.-F., Cottam, A., Gorelick, N., Belward, A.S., 2016. High-resolution mapping of global surface water and its long-term changes. Nature 540, 418–422. https://doi.org/10.1038/nature20584.

Qayyum, N., Ghuffar, S., Ahmad, H.M., Yousaf, A., Shahid, I., 2020. Glacial lakes mapping using multi satellite PlanetScope imagery and deep learning. ISPRS Int. J. Geo-Inform. 9, 560.

Rishikeshan, C., Ramesh, H., 2018. An automated mathematical morphology driven algorithm for water body extraction from remotely sensed images. ISPRS J. Photogramm. Remote Sens. 146, 11–21.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer, pp. 234–241.

Salomonson, V.V., Appel, I., 2004. Estimating fractional snow cover from MODIS using the normalized difference snow index. Remote Sens. Environ. 89, 351–360. https://doi.org/10.1016/j.rse.2003.10.016.

SenseFly, 2020. SenseFly Parrot Group UAV manufacturer - switzerland. URL: https://www.sensefly.com/.

Shrestha, B.B., Nakagawa, H., 2014. Assessment of potential outburst floods from the Tsho Rolpa glacial lake in Nepal. Nat. Hazards 71, 913–936.

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. Nat. Clim. Change 10, 939–945. https://doi.org/10.1038/s41558-020-0855-4.

Song, C., Huang, B., Ke, L., Richards, K.S., 2014. Remote sensing of alpine lake water environment changes on the tibetan plateau and surroundings: A review. ISPRS J. Photogramm. Remote Sens. 92, 26–37.

Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Cardoso, M.J., 2017. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep learning in medical image analysis and multimodal learning for clinical decision support. Springer, pp. 240–248.

Tamiminia, H., Salehi, B., Mahdianpari, M., Quackenbush, L., Adeli, S., Brisco, B., 2020. Google earth engine for geo-big data applications: A meta-analysis and systematic review. ISPRS J. Photogramm. Remote Sens. 164, 152–170.

USGS, 2022. Why are negative values observed over water in some Landsat Surface Reflectance products. URL: https://www.usgs.gov/faqs/why-are-negative-values-observed-over-water-some-landsat-surface-reflectance-products.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. arXiv preprint arXiv:1706.03762.

Veh, G., Korup, O., Roessner, S., Walz, A., 2018. Detecting Himalayan glacial lake outburst floods from Landsat time series. Remote Sens. Environ. 207, 84–97.

Veh, G., Korup, O., Walz, A., 2020. Hazard from Himalayan glacier lake outburst floods. Proc. Nat. Acad. Sci. 117, 907–912.

Wang, S., Guan, Y., Shao, L., 2020. Multi-granularity canonical appearance pooling for remote sensing scene classification. IEEE Trans. Image Process. 29, 5396–5407.

Wang, S., Ren, Y., Parr, G., Guan, Y., Shao, L., 2021. Invariant deep compressible covariance pooling for aerial scene categorization. IEEE Trans. Geosci. Remote Sens. 59, 6549–6561. https://doi.org/10.1109/TGRS.2020.3026221.

Wang, X., Girshick, R., Gupta, A., He, K., 2018. Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7794–7803.

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.

Wu, R., Liu, G., Zhang, R., Wang, X., Li, Y., Zhang, B., Cai, J., Xiang, W., 2020. A deep learning method for mapping glacial lakes from the combined use of synthetic-aperture radar and optical satellite images. Remote Sens. 12, 4020.

Xia, M., Qian, J., Zhang, X., Liu, J., Xu, Y., 2019. River segmentation based on separable attention residual network. J. Appl. Remote Sens. 14, 1–15. https://doi.org/10.1117/1.JRS.14.032602.

Xie, Z., Haritashya, U.K., Asari, V.K., Young, B.W., Bishop, M.P., Kargel, J.S., 2020. GlacierNet: A deep-learning approach for debris-covered glacier mapping. IEEE Access 8, 83495–83510.

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. https://doi.org/10.1080/01431160600589179.

Yang, H., Wu, P., Yao, X., Wu, Y., Wang, B., Xu, Y., 2018. Building extraction in very high resolution imagery by Dense-Attention networks. Remote Sens. 10 https://doi.org/10.3390/rs10111768.

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 https://doi.org/10.1016/j.rse.2020.112265. Cited By 0.

Zhang, M.-M., Zhao, H., Chen, F., Zeng, J.-Y., 2020. Evaluation of effective spectral features for glacial lake mapping by using Landsat-8 OLI imagery. J. Mount. Sci. 17, 2707–2723. https://doi.org/10.1007/s11629-020-6255-4.

Zhu, X.X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing: A comprehensive review and list of resources. IEEE Geosci. Remote Sens. Mag. 5, 8–36.