SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Efficient Flood Segmentation from Sentinel-1 SAR Imagery Using SegFormer and CNN-Based Deep Learning Models

K. Lakshman Kumar[1*], Ranga Swamy Sirisati[2]

[1*]Department of Computer Science and Engineering, Bharatiya Engineering Science and Technology Innovation University, Gownivaripalli, Gorantla, Andhra Pradesh, India. E-mail: 2022wpcse011@bestiu.edu.in , ORCID id: 0009-0009-2920-8423
[2]Associate Professor, Department of AIML, School of Engineering, Anurag University, Hyderabad.
E-mail: sirisatiranga@gmail.com, ORCID id: 0000-0002-3104-6672

ABSTRACT
Having the good and timely data of flood extent is crucial for disaster management, disaster risk assessment and environmental management. Synthetic Aperture Radar (SAR) imagery is another source of data that can be used to map flooding since it can be acquired under cloudy conditions and does not require daylight. However, the backscatter characteristics of SAR imagery is heterogeneous that makes accurate delineation of flood at pixel level difficult. A light Weight Vision Transformer - based semantic segmentation architecture, SegFo.rmer, is compara.tively evalu.ated in this stu.dy with four CNN-bas.ed mod.els: U-Net, Weig.hted U-Net, .""Deep.LabV3+ and Atten.tion U-Net for pix.el-lev.elflo.oddeline.ation usi.ng Sent.inel-1 SAR data bas.ed on the SEN1Flo.ods11 bench.mark. A unif.iedexperi.mental prot.ocol was used for data preproc.essing & mod.eltrai.ningtest.ing and perfor.manceevalu.ation". Accu.racy, preci.sion, rec.all, F1-sco.re/Dice coeffi.cient, Inters.ection over Uni.on (IoU), rece.iveropera.tingcharact.eristic area und.er the cur.ve (ROC-AUC), qualit.ative segmen.tation anal.ysis, train.able param.eters, and infer.ence time were used to comp.are the evalu.atedmod.els. The results of SegFormer showed the highest overall accuracy, F1/Dice score, IoU, and ROC-AUC. Attention U-Net had maximum recall (70.17%) and UNET had the lowest measured inference time (6.23 ms/image). The number of trainable parameters in SegFormer was as low as 3.71 million, significantly lower than the CNN architectures evaluated. The results show that SegFormer achieves a good compromise between segmentation quality and parameter efficiency for Sentinel-1 SAR flood segmentation, and the comparative study reveals significant trade-offs between flood sensitivity, segmentation overlap, and computational efficiency.

Keywords: Flood segmentation, Sentinel-1 SAR, SegFormer, semantic segmentation, deep learning, remote sensing.

1. Introduction

1.1 Background and motivation

"Floods are one of the most disruptive natural hazards with significant impacts on human". settlements, transportation networks, agricultural land, natural ecosystems, and infrastructure. The knowledge of the spatial extent of inundation is thus relevant for emergency response, damage assessment, evacuations, recovery operations and disaster-risk management. Although conventional field-based surveys can give detailed 'ground truth' information, their use can be limited during active flooding in terms of accessibility, safety, time and geographical area. Satellite-based Earth observation provides an alternative means of obtaining spatially extensive information for flood assessment (Amitrano et al. 2018; Nemni et al. 2020).

Synthetic Aperture Radar (SAR) is highly useful for monitoring floods because microwave observations can be acquired independently of daylight and are substantially less affected by cloud cover than optical observations. These properties are especially important during flood events associated with severe rainfall, when persistent cloud cover can restrict the availability of optical satellite imagery. Sentinel-1 SAR observations have consequently become an important source of information for flood mapping and other environmental monitoring applications(Twele et al. 2016; Liang and Liu 2020; Amitrano et al. 2018; Tay et al. 2020).

"The SEN1Floods11 dataset was introduced to support the development and evaluation of deep-learning approaches for flood mapping using Sentinel-1 imagery"."The benchmark comprises 4,831 georeferenced image chips at 512 × 512 resolution, representing approximately 120,406 km^2 across six continents, 14 biomes, 357 ecoregions, and 11 flood events"."The dataset includes Sentinel-1 observations together with classified surface-water information and provides an important benchmark for machine-learning-based flood mapping. (Bonafilia et al. 2020)".

Despite the advantages of SAR, automated flood delineation remains challenging. SAR images are influenced by speckle noise, surface roughness, vegetation, soil conditions, built-up structures, terrain effects, and variations

in radar backscatter(Li et al. 2019; Liang and Liu 2020; Nemni et al. 2020; Zhao et al. 2021; Saleh et al. 2024). Consequently, flooded surfaces may exhibit heterogeneous responses, while non-flooded surfaces can sometimes generate backscatter patterns similar to those associated with inundated areas. The separation of flooded and background pixels is not reliable because of such characteristics, especially in complex scenes where there are vegetation, urban areas, agriculture fields and fragmented water bodies.

Deep-learning segmentation benefits from the ability of models to learn Deep-learning segmentation benefits from models that can learn hierarchical feature representations directly from image data. Encoder–decoder CNN architectures, including U-Net and its variants, are widely used for semantic segmentation because their skip connections allow high-level semantic information to be combined with spatially detailed features. Such characteristics are useful for flood mapping, where accurate localization of inundation boundaries and retention of smaller flooded regions are important(Ronneberger et al. 2015; Shelhamer et al. 2017; Zhu et al. 2017).

However, conventional CNN architectures primarily obtain contextual information through convolutional operations. Although deeper networks and multi-scale mechanisms can enlarge the effective receptive field, modelling relationships between spatially distant image regions remains an important challenge(Dosovitskiy et al. 2021; Xie et al. 2021; Lv et al. 2023). This issue is particularly relevant to flood segmentation because inundated areas can be spatially extensive, fragmented, irregular, and distributed across heterogeneous land-cover environments.

Attention mechanisms can emphasize informative spatial features (Schlemper et al. 2019), while weighted or class-sensitive loss functions can increase the contribution of under-represented flood pixels during optimization (Sudre et al. 2017; Lin et al. 2017). Nevertheless, improving flood-pixel sensitivity can also increase false-positive predictions(Sudre et al. 2017; Lin et al. 2017). Consequently, overall pixel accuracy alone is insufficient to characterize the quality of a flood segmentation model. Metrics such as precision, recall, F1/Dice and IoU provide additional information about the quality of the predicted flood regions.

Vision Transformers provide another approach to contextual representation by using self-attention to model relationships among spatially separated image regions. SegFormer is a particularly relevant architecture because it combines a hierarchical Transformer encoder that produces multi-scale representations with a lightweight multilayer perceptron decoder. The architecture was designed to provide an efficient alternative to more complex transformer-based segmentation systems (Xie et al. 2021).

The potential of transformer-based approaches for SAR flood mapping has also been demonstrated in recent studies. For example, DAM-Net uses differential attention and transformer-based feature representations for flood detection from bi-temporal SAR imagery and introduced the S1GFloods dataset containing 5,360 image pairs across 46 flood events. The reported results demonstrate the potential of transformer-based contextual modelling for distinguishing flood-related changes from SAR-specific background variations (Saleh et al. 2024). Similarly, DeepSARFlood investigated vision-transformer-based deep ensembles for automated SAR flood inundation mapping and incorporated uncertainty estimation into the flood-mapping framework. The study reported improved IoU through multi-task learning and model ensembles and demonstrated automated processing for large geographical areas (Sharma and Saharia 2025).

Recent benchmark development has also highlighted the importance of considering different geographical and land-cover conditions. UrbanSARFloods, for example, was introduced to address the limited representation of urban flood environments in existing SAR flood datasets. The dataset contains 8,879 512 × 512 chips spanning 18 flood events, 20 land-cover classes, five continents, and approximately 807,500 km^2. Its authors also evaluated several CNN segmentation architectures, demonstrating the importance of dataset diversity and standardized evaluation for SAR flood mapping. (Zhao et al. 2024).

These developments indicate that the research problem is no longer simply whether deep learning can segment floods from SAR imagery. Instead, an important question is which architectural characteristics provide the most useful balance between segmentation quality and computational requirements under a controlled experimental setting. This consideration is particularly relevant when a model is intended for repeated processing of satellite observations during disaster-response applications.

1.2. Research gap

While CNN-based and transformer-based methods have shown great promise for SAR flood mapping, there are a number of issues that remain pertinent.

First, the traditional CNN structures mostly use local feature extraction of the convolution operation. Although U-Net and similar encoder–decoder networks are good at capturing spatial information, they have comparatively limited capacities to directly model long-range contextual relationships. Although attention modules and multi-scale feature extraction do have the ability to enhance the representation of context, they are still embedded in the majority of CNN-based frameworks.

Second, class imbalance is an inherent problem with flood segmentation. The number of background pixels can be significant and much larger than the number of flood pixels, and this can result in a model having a high overall accuracy with comparatively poor flood-region overlap. Weighted-loss methods can help to enhance flood-pixel sensitivity, which may compromise with an increase in the number of false positive predictions. Hence a thorough evaluation should involve several complementary measures to the accuracy.

Third, transformer-based architectures have brought global self-attention in the field of semantic segmentation and have proven to be promising for SAR flood detection. Yet, recent research shows that the performance of the transformer is influenced by the set of data, the flooding state, the input representation and the evaluation protocol. Consequently, the superiority of a transformer architecture should be established experimentally rather than assumed from its architectural design.

Fourth, computational efficiency is an important consideration for practical Earth-observation applications. A flood-mapping model may need to process large quantities of satellite imagery within a limited period. Therefore, model parameters and inference time should be evaluated alongside segmentation quality.

These considerations motivate a controlled comparative investigation of CNN and transformer-based segmentation architectures using the same benchmark and experimental pipeline. In this study, U-Net, Weighted U-Net, DeepLabV3+, Attention U-Net, and SegFormer are evaluated using the SEN1Floods11 Sentinel-1 SAR dataset. The analysis considers segmentation quality, discrimination capability, qualitative prediction behaviour, model complexity, and inference efficiency.

1.3 Objectives

The objectives of this study are:

- "To evaluate SegFormer for semantic flood segmentation using Sentinel-1 SAR imagery".
- "To implement four representative CNN-based segmentation architectures—U-NetWeighted U-Net& DeepLabV3+and Attention U-Net—for comparative analysis"".
- "To evaluate the investigated architectures using accuracy& precisionas well as recall F1-score/DiceIoU and ROC-AUC"".
- "To examine pixel-level prediction behaviour using confusion matrices and qualitative segmentation maps".
- "To compare the computational characteristics of the models using trainable parameter counts and inference time".
- "To identify the relative strengths and trade-offs between CNN-based and transformer-based architectures for SAR flood segmentation".

1.4 Contributions

The main contributions of this study are:

- Controlled CNN–Transformer evaluation: A common experimental framework is used to compare four representative CNN architectures with SegFormer for Sentinel-1 SAR flood segmentation.
- "Multi-metric segmentation assessment: The models are evaluated using accuracy, precision, recall, F1/Dice, IoU, and ROC-AUC rather than relying on a single performance indicator.
- Pixel-level error analysis: Confusion matrices and qualitative segmentation maps are used to examine false-positive, false-negative, and boundary-related prediction behaviour.
- Computational assessment: Trainable parameter counts and measured inference time are reported alongside segmentation performance to examine the performance–efficiency trade-off.
- Empirical assessment of SegFormer: The study evaluates whether a lightweight transformer-based architecture can provide strong segmentation performance while maintaining a substantially smaller parameter footprint than the investigated CNN architectures.

1.5 Organization of the paper

The remainder of this paper is organized as follows. Section 2 reviews CNN-based attention-based and transformer-based approaches for SAR flood segmentation and identifies the research gap addressed by this study.Section 3 describes the SEN1Floods11 dataset preprocessing procedure and evaluated segmentation architectures as well as training configuration also evaluation metrics. Section 4 presents the quantitative and qualitative plus ROC& confusion-matrix and computational results. Section 5 discusses the limitations of the study and directions for future research. Section 6 presents the conclusions followed by the declarations.

2. Related Work

2.1. CNN-based SAR flood segmentation

"Deep-learning-based semantic segmentation has become an important approach for automated flood mapping from remote-sensing imagery". "CNN based encoder–decoder is a good architecture for this task due to the integration of semantic and spatial information". "The influential encoder–decoder paradigm has been developed in U-Net where skip connections are used to pass spatial information from the encoder to the decoder which allows for detailed predictions at the pixel level (Ronneberger et al. 2015; Shelhamer et al. 2017)".

"The SEN1Floods11 benchmark was very useful as a starting point for flood mapping using Sentinel-1 imagery with deep learning approach". "The dataset was created by (Bonafilia et al. 2020; Bai et al". "2021) "to train and test the fully convolutional neural networks for surface-water and flood segmentation"." It has been geographically diverse and will therefore be helpful for assessing the deep-learning techniques for a variety of flood events and environmental conditions".

Further studies have adopted CNN-based constructions that include attention, multi-scale feature extraction, residual connections, and other means of feature representation. Atrous convolutions and multi-scale

contextual representation are the two key ideas in DeepLab-type architectures (Chen et al. 2018), which aim to model features at different spatial scales. In attention-based encoder–decoder networks, extra feature-selection mechanisms are added to highlight relevant spatial regions (Schlemper et al. 2019).

The current popular CNN-based methods are still appealing due to their well-defined training strategy, ease of implementation, and powerful spatial inductive bias. But their contextual representation is still mainly based on convolutional operations. This is especially important in SAR imagery, which can have geographically large flood zones and where spatial relationships of context can be shared between areas distant to each other.

2.2 Attention and weighted-loss approaches

SAR flood segmentation is affected by both complex image characteristics and class imbalance. Flood pixels can occupy a relatively small proportion of an image compared with non-flooded background pixels. Consequently, an unweighted training objective may encourage the model to prioritize the dominant background class.

Weighted-loss strategies address this problem by assigning greater importance to under-represented classes during model optimization. In flood segmentation, this can increase the model's sensitivity to inundated regions. However, greater sensitivity does not necessarily imply better segmentation overlap because additional flood predictions may increase false-positive rates(Sudre et al. 2017; Lin et al. 2017).

Attention mechanisms provide another approach to improving feature representation. By assigning greater importance to informative spatial features, attention modules can help segmentation networks focus on relevant regions and suppress less informative background features. Such approaches can be particularly useful when flooded surfaces exhibit similar backscatter characteristics to other land-cover categories(Schlemper et al. 2019).

Recent SAR flood studies have continued to explore attention and contextual feature extraction. DAM-Net, for example, explicitly addresses SAR-specific difficulties such as similar backscatter responses and speckle noise using differential attention and transformer-based representations. The study reports that modelling broader contextual relationships can help distinguish water-body changes from pseudo-changes in SAR imagery (Saleh et al. 2024). Nevertheless, attention mechanisms integrated into CNNs still retain the local inductive structure of convolutional feature extraction. This provides motivation for examining architectures in which self-attention plays a more fundamental role in contextual representation.

2.3 Transformer-based SAR flood mapping

Vision Transformers introduced self-attention-based image representation as an alternative to conventional convolutional feature extraction(Dosovitskiy et al. 2021). The self-attention mechanism allows relationships between spatially separated image regions to be represented directly, providing a mechanism for modelling broader contextual dependencies.

SegFormer combines this concept with a hierarchical Transformer encoder and a lightweight MLP decoder. Rather than using a complex decoder, SegFormer aggregates multi-scale features generated by its hierarchical encoder. The original architecture was designed to combine efficiency with strong semantic segmentation performance (Xie et al. 2021).

Transformer-based approaches have increasingly been applied to SAR flood detection. DAM-Net uses a differential attention metric-based architecture to process bi-temporal SAR imagery and distinguish flood-related changes from pseudo-changes. Its S1GFloods dataset was designed to provide broader geographic and event diversity for SAR flood detection. (Saleh et al. 2024). DeepSARFlood further demonstrates the potential of Vision Transformer-based methods by combining deep ensembles, multi-task learning, and uncertainty estimation for rapid SAR flood inundation mapping. The framework reported improved IoU and was designed for automated large-area processing, demonstrating the increasing emphasis on both accuracy and operational efficiency in SAR flood mapping. These studies indicate that transformer-based models can provide useful global contextual representation. However, many transformer-based flood-mapping studies employ specialized architectures, bi-temporal change detection, ensembles, or additional components. Consequently, there remains value in evaluating a comparatively lightweight general-purpose semantic segmentation architecture such as SegFormer against established CNN baselines under a common experimental setting.

2.4 Benchmark datasets and evaluation challenges

Dataset characteristics strongly influence the observed performance of flood-segmentation models. The SEN1Floods11 dataset was designed as a global benchmark and includes 11 flood events across diverse geographical and ecological environments(Bonafilia et al. 2020). However, subsequent work has highlighted limitations associated with the representation of specific flood environments. UrbanSARFloods was developed partly because existing large-scale SAR flood datasets had limited representation of urban flooding. The dataset provides Sentinel-1 intensity and interferometric-coherence information and includes 8,879 chips covering 18 flood events, 20 land-cover classes, and five continents (Zhao et al. 2024).

These developments demonstrate that reported model performance must be interpreted in relation to dataset composition and evaluation protocol. Differences in geographic distribution, flood events, spatial resolution, preprocessing, class balance, train/test partitioning, and evaluation metrics can substantially affect the resulting performance.

For this reason, the present study does not attempt to claim that its numerical results are directly superior to every published flood-segmentation result. Instead, the primary comparison is performed within a controlled

experimental framework, where the five investigated architectures are evaluated using the same experimental dataset and evaluation procedure.

Table 1. Representative deep-learning approaches for SAR-based flood mapping

Study	Dataset / Data Source	Method / Architecture	Main Focus
Bonafilia et al. (2020)	SEN1Floods11	FCN/CNN	Benchmark dataset for Sentinel-1 flood mappingZ
Nemni et al. (2020)	SAR flood imagery	Fully Convolutional Neural Network (FCNN)	Rapid flood segmentation
Ghosh et al. (2022)	Sentinel-1	Deep-learning architectures	Automatic flood detection
Choi et al. (2022)	Sentinel-1	U-Net / HRNetV2	Flood segmentation
Li et al. (2023)	Flood datasets	GeoAI / Deep Learning	Flood inundation mapping and generalization
Ghosh et al. (2024)	Sentinel-1 / NASA benchmark dataset	Nested U-Net	Automatic flood detection
Pech-May et al. (2024)	Sentinel-1	U-Net	Flood-area segmentation and visualization
Saleh et al. (2024)	S1GFloods	DAM-Net	Transformer-based flood detection
Zhao et al. (2024)	UrbanSARFloods	CNN-based segmentation	Urban and open-area flood mapping
Sharma and Saharia (2025)	SAR flood imagery	DeepSARFlood – Vision Transformer-based deep ensembles	Automated flood inundation mapping with uncertainty estimation
Present study	**SEN1Floods11 subset**	**U-Net, Weighted U-Net, DeepLabV3+, Attention U-Net, SegFormer**	**Controlled CNN–Transformer comparison and computational-efficiency assessment**

2.5 Research gap

The reviewed literature demonstrates a progression from conventional CNN-based segmentation toward attention-enhanced and transformer-based approaches for SAR flood mapping. Nevertheless, several issues remain important.

First, CNN architectures are still good baselines for semantic segmentation but they cannot model long range contextual relationships directly due to their local operations. Second, weighted-loss and attention mechanisms can enhance the sensitivity of flood detection, but there could be trade-offs between precision and recall in some cases. Third, transformer-based architectures have shown some good results, but their usefulness in practice must be evaluated based on their architecture and not taken for granted.

Another gap is related to combined assessment of segmentation quality and computational properties. Though many studies have focused on accuracy, F1-score or IoU, for operational flood mapping, the size and processing time of the model should also be considered. A high segmentation quality model which requires a lot of computation may be not so useful for repeated or large area processing.

Thus, in this study, U-Net, Weighted U-Net, DeepLabV3+, Attention U-Net, and SegFormer are compared in a controlled study on the SEN1Floods11 benchmark. The comparison combines metrics based on segmentation, ROC-AUC, confusion-matrix analysis, qualitative prediction analysis, the number of parameters and the inference time. This provides a broader assessment of the relative behaviour of CNN-based and transformer-based segmentation architectures for Sentinel-1 SAR flood mapping.

3. Materials And Methods

3.1 Overall Experimental Framework

The experimental framework was designed to provide a controlled comparison between conventional

convolu.tionalneu.ralnetw.ork (CNN)-bas.edsema.nticsegmen.tationarchite.ctures and a lightw.eightVis.ionTransf.ormerarchit.ecture for flo.oddeline.ation from Sent.inel-1 synth.eticaper.ture rad.ar (SAR) imag.ery. The comp.letework.flowcons.ists of data.setprepar.ation, ima.ge–mask verifi.cation, preproc.essing, data.setpartit.ioning, mod.elimpleme.ntation, mod.eltrai.ning, predi.ction, quanti.tativeevalu.ation, qualit.ativeasses.sment, and computa.tionalanal.ysis.

The publicly available SEN1Floods11 dataset was selected as the benchmark for the experiments. The dataset provides Sentinel-1 SAR imagery together with corresponding flood-related ground-truth labels and was originally developed to support deep-learning-based flood mapping (Bonafilia et al. 2020). From the available dataset, the experimental subset used in this study consisted of 446 image–mask pairs, which were divided into 312 training pairs, 67 validation pairs, and 67 independent testing pairs.

Five semantic segmentation architectures were evaluated: U-Net, Weighted U-Net, DeepLabV3+, Attention U-Net, and SegFormer. The first four represent CNN-based approaches with different feature extraction or feature-weighting mechanisms, whereas SegFormer represents a Vision Transformer-based segmentation architecture. The same prepared dataset and evaluation procedure were used to enable a controlled comparison.

The experimental analysis considered both segmentation quality and computational characteristics. Segmentation quality was evaluated using accuracy, precision, recall, F1-score, Dice coefficient, IoU, ROC-AUC, and pixel-level confusion matrices. Qualitative comparison of predicted flood masks was additionally performed to examine spatial continuity and flood-boundary delineation. Computational characteristics were assessed using the number of trainable parameters and measured inference time per image.

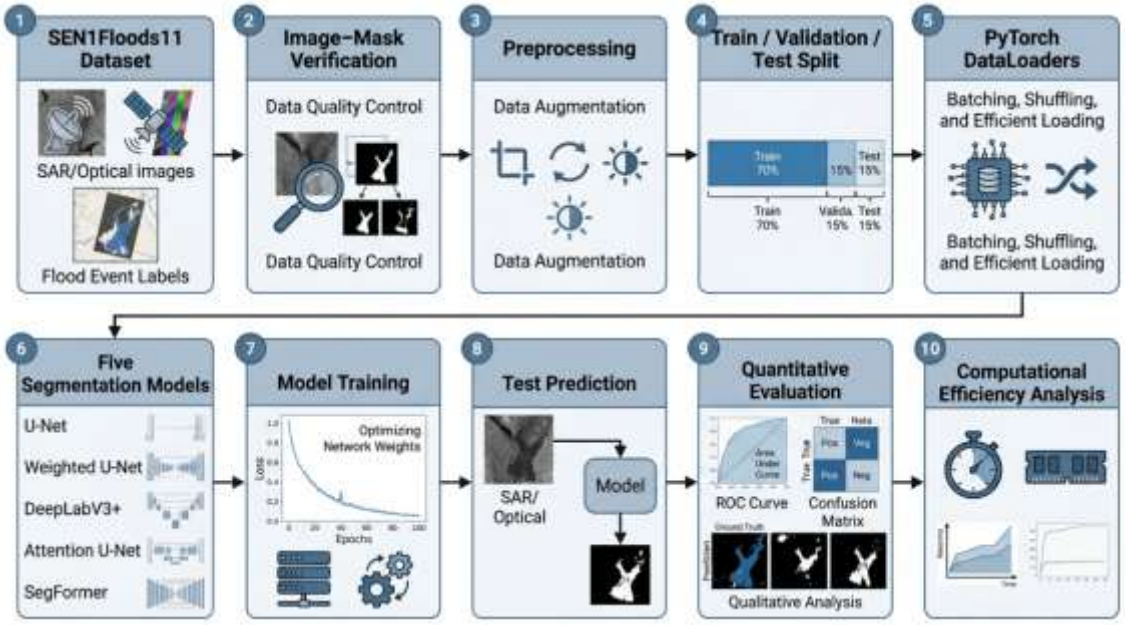

Fig. 1. Overall experimental workflow for Sentinel-1 SAR flood segmentation using CNN-based and SegFormer architectures.

3.2 Dataset

The SEN1Floods11 benchmark dataset was used for the experimental evaluation. It was developed to support machine-learning-based flood and surface-water mapping using Sentinel-1 satellite observations. The original benchmark contains georeferenced image chips representing flood events across diverse geographical and ecological conditions (Bonafilia et al. 2020).

For the present experiments, a subset of 446 image–mask pairs was used. Each sample consisted of a Sentinel-1 SAR image and its corresponding binary flood segmentation mask. The segmentation task was formulated as a two-class pixel-level classification problem, with flood representing the foreground class and background representing the non-flooded class.

The selected samples were divided into training, validation, and testing subsets. The training subset contained 312 image–mask pairs, while the validation and testing subsets contained 67 pairs each. The test subset was kept separate from the training process and was used only for final performance evaluation.

Table 2. Dataset configuration used in the experiments

Parameter	Configuration
Dataset	SEN1Floods11

Satellite/Sensor	Sentinel-1 SAR
Task	Binary semantic segmentation
Classes	Flood, Background
Experimental samples	446 image–mask pairs
Training samples	312
Validation samples	67
Testing samples	67
Input image size	256 × 256 pixels
Deep-learning framework	PyTorch

3.3 Representative Dataset Samples

Representative Sentinel-1 SAR images and their corresponding ground-truth masks were examined before model training to verify the consistency of the segmentation task. The SAR observations exhibit different spatial distributions of inundated regions and varying surrounding land-cover characteristics. The corresponding masks provide pixel-level reference labels used during supervised model training and evaluation.

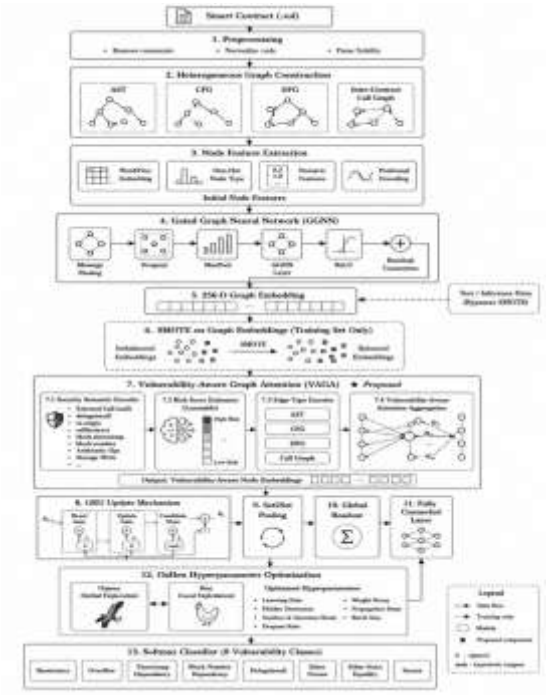

Fig. 2. Representative Sentinel-1 SAR images and corresponding ground-truth flood masks from the experimental dataset.

3.4 Data Preprocessing

The image–mask pairs were preprocessed with a uniform pipeline before training the model. The preprocessing steps were taken to provide the same input data to all evaluated architectures, and to make the image size and its segmentation labels uniform across the experiments.

3.4.1 Image–Mask Verification

The SAR images each had a corresponding segmentation mask, which was attached to the image prior to being added to the experimental dataset. To avoid wrong pairing between the input observations and ground-truth labels, image–mask correspondence was verified.

3.4.2 Image Resizing

The input images were resized to 256 × 256 pixels. To enable batch based computation on the GPUs, they used the same spatial input to all five architectures with the same input dimensions.

3.4.3 Pixel Normalization

The SAR image values were normalized prior to being fed to the neural networks. Normalization helps to scale the numerical input range, and helps to keep model training stable.

3.4.4 Binary Mask Preparation

The ground-truth masks were converted into a binary representation corresponding to the two segmentation classes:

$$Y(x,y) \in \{0,1\} \quad (1)$$

Where

0=Background and1= flood

3.4.5 Tensor Conversion and Data Loading

The images and the masks were then converted to PyTorch tensors after the preprocessing. To support mini-batch processing for training and evaluation, the processed samples were loaded using PyTorchDataLoaders.

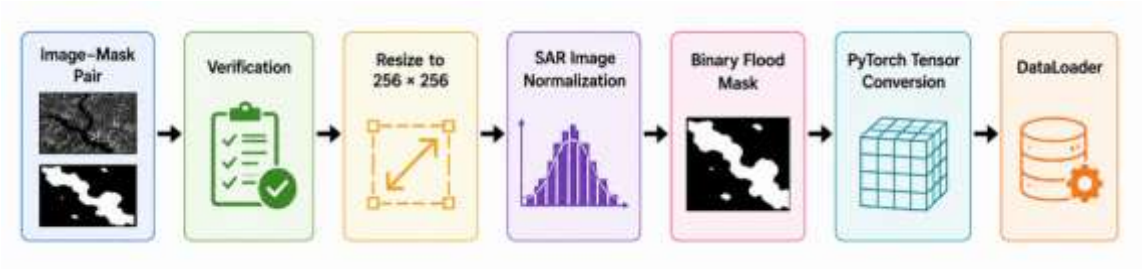

Fig. 3. Preprocessing pipeline used for Sentinel-1 SAR flood segmentation.

3.5 Segmentation Architectures

Five sema.nticsegmen.tationarchite.ctures were cho.sen to make a compa.risonbetw.een the exis.ting CNN-bas.edarchite.ctures and the transf.ormer-bas.edarchit.ecture. The cho.senmod.elscorre.spond to vari.ousappro.aches for spat.ialfeat.ureextra.ction, spat.ialfeat.urerepresе.ntation and spat.ialfeat.urelocali.zation.

3.5.1 U-Net

U-Net is origi.nallydesi.gned for biome.dical ima.ge segmen.tation, and is an enco.der–deco.der CNN. The enco.dergradu.allyenco.des the hig.herlev.els of seman.tics and the deco.dergradu.allydeco.des the spat.ialresol.ution. Skip connec.tions are used to pass spat.ialinform.ation from enco.dersta.ges to deco.dersta.ges in a simi.larman.ner to pres.erve fine struc.turalinform.ation.

In the curr.entstu.dy, U-Net is used as the base CNN archit.ecture, and the oth.er segmen.tationmod.els are comp.ared to it (Ronneb.erger et al. 2015).

3.5.2 Cla.ss-Weig.hted U-Net

Cla.ss-Weig.hted U-Net is an adapt.ation of the base U-Net archit.ecture, whi.ch invo.lvesadd.ing the cla.sswei.ght to the trai.ningobjec.tive. This appr.oach is partic.ularlyrele.vant to flo.odsegmen.tationbeca.use the num.ber of backg.roundpix.els can substan.tiallyexc.eed the num.ber of flo.odpix.els(Sud.re et al. 2017; Lin et al. 2017).

The weighted loss assigns different contributions to the segmentation classes:

$$L_{weighted} = -\frac{1}{N}\sum_{i=1}^{N} w_{yi}\,[y_i \log(pi) + (1 - y_i) log(1 - p_i)] \qquad (2)$$

where w_{yi}represents the weight associated with the ground-truth class, yiis the target label, and $p_i$is the predicted probability.

3.5.3 DeepLabV3+

DeepLabV3+ employs an encoder–decoder architecture with Atrous Spatial Pyramid Pooling (ASPP). Atrous convolution enables the network to capture contextual information at multiple effective receptive-field sizes without proportionally increasing the number of parameters. The decoder subsequently combines semantic representations with spatial information to generate the final segmentation map(Chen et al. 2018). The multi-scale design of DeepLabV3+ makes it suitable for scenes containing flood regions with different spatial extents.

3.5.4 Attention U-Net with EfficientNet-B3 and SCSE Attention

In this study, the Attention U-Net implementation uses an EfficientNet-B3 encoder with SCSE (Spatial and Channel Squeeze and Excitation) attention in the decoder pathway. The EfficientNet-B3 encoder provides hierarchical feature representations, while the SCSE mechanism enhances informative spatial and channel-wise responses during feature reconstruction (Tan and Le 2019;Roy et al. 2018). This implementation was selected to provide an attention-enhanced CNN counterpart to the baseline U-Net.

3.5.5 SegFormer

SegFormer is a transformer-based semantic segmentation architecture consisting of a hierarchical Transformer encoder and a lightweight MLP decoder (Dosovitskiy et al. 2021; Xie et al. 2021). Unlike conventional CNN architectures, the encoder uses self-attention-based representations to capture contextual relationships between image regions.

The hierarchical encoder produces multi-scale feature representations, while the lightweight decoder combines these features to generate the final segmentation prediction. The relatively compact decoder contributes to the computational efficiency of the architecture.

In this study, SegFormer is evaluated as the transformer-based alternative to the four CNN architectures under the same experimental evaluation framework.

Table 3. Characteristics of the investigated segmentation architectures

Model	Architecture	Main mechanism	Role in comparison
U-Net	CNN encoder–decoder	Skip connections	Baseline CNN
Weighted U-Net	CNN encoder–decoder	Class-weighted optimization	Class-imbalance handling
DeepLabV3+	CNN	ASPP + decoder	Multi-scale contextual representation
Attention U-Net	CNN encoder–decoder with EfficientNet-B3 encoder	SCSE attention	Attention-enhanced feature localization
SegFormer	**Vision Transformer**	**Hierarchical self-attention + MLP decoder**	**Transformer-based comparison**

3.6 Training Strategy and Experimental Configuration

All five architectures were evaluated using the same dataset partition, input resolution, preprocessing pipeline, independent test set, evaluation metrics, and prediction threshold, while model-specific optimization settings were retained according to the experimental configurations described in Table 4. The input image size was fixed at 256 × 256 pixels and mini-batch training was employed (Paszke et al. 2019). The U-Net baseline was trained using Adam (Kingma and Ba 2015) with a learning rate of 1×10^{-4} for 10 epochs, whereas Weighted U-Net and DeepLabV3+ were trained for 20 epochs using Adam with a learning rate of 1×10^{-4}. Attention U-Net was trained for 25 epochs using AdamW (Loshchilov and Hutter 2019) with a learning rate of 1×10^{-4} and weight decay of 1×10^{-5}. SegFormer was trained for 20 epochs using AdamW (Loshchilov and Hutter 2019) with a learning rate of 5×10^{-5} and weight decay of 1×10^{-5}. The batch size was 4 for all models. Weighted configurations used a positive-class weight of 3.0. The combined loss can be expressed as:

$$L_{total}=L_{BCE}+L_{Dice} \quad (3)$$

Where L_{BCE} represents the binary cross-entropy loss and L_{Dice} represents the Dice loss.

For the class-weighted configurations, the binary cross-entropy component was modified using a positive-class weight of 3.0:

$$L_{W\,BCE} = -\frac{1}{N}\sum_{i=1}^{N} w_{yi}[y_i \log(pi) + (1 - y_i)\log(1 - p_i)] \quad (4)$$

where N is the number of pixels, y_i is the ground-truth label, and p_i is the predicted probability for the flood class.

Dice loss can be expressed as:

$$L_{Dice} = 1 - \frac{2\sum_i p_i y_i + \epsilon}{\sum_i p_i + \sum_i y_i + \epsilon} \quad (5)$$

where ϵ is a small constant introduced for numerical stability.

The use of the combined loss provides both pixel-wise classification supervision through BCE and region-overlap supervision through Dice loss(Sudre et al. 2017).

Table 4. Model-specific training configurations used in the experiments

Model	Epochs	Optimizer	Learning Rate	Weight Decay	Loss Configuration
U-Net	10	Adam	1×10^{-4}	0	BCE + Dice
Weighted U-Net	20	Adam	1×10^{-4}	0	Weighted BCE + Dice
DeepLabV3+	20	Adam	1×10^{-4}	0	Weighted BCE + Dice
Attention U-Net	25	AdamW	1×10^{-4}	1×10^{-5}	Weighted BCE + Dice
SegFormer	20	AdamW	5×10^{-5}	1×10^{-5}	Weighted BCE + Dice

Common settings: input size = 256 × 256 pixels, batch size = 4, positive-class weight = 3.0, prediction threshold = 0.5, PyTorch= NVIDIA Tesla T4.

3.7 Performance Evaluation

The evaluation metrics were selected to provide complementary measures of pixel-level classification accuracy and region overlap, following established segmentation-evaluation practice (Taha and Hanbury 2015). The segmentation performance of each model was evaluated on the independent test set. Because flood segmentation is a pixel-level binary classification problem and may be affected by class imbalance, multiple complementary metrics were used.

3.7.1 Accuracy

Accuracy represents the proportion of correctly classified pixels:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives.

3.7.2 Precision

Precision measures the proportion of predicted flood pixels that are actually flood pixels:

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

A higher precision indicates fewer false-positive flood predictions.

3.7.3 Recall

Recall measures the proportion of actual flood pixels correctly detected:

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

Higher recall indicates stronger sensitivity to inundated regions.

3.7.4 F1-score

The F1-score represents the harmonic mean of precision and recall:

$$F1 = \frac{2\times Precision\times Recall}{Precision+Recall} \quad (9)$$

3.7.5 Dice Coefficient

For binary segmentation, the Dice coefficient can be expressed in terms of the confusion-matrix components as:

$$Dice = \frac{2TP}{2TP+FP+FN} \quad (10)$$

Because Dice and F1-score are mathematically equivalent when calculated from the same positive-class predictions, they are reported jointly as F1/Dice rather than as two independent measures.

3.7.6 Intersection over Union

IoU measures the overlap between predicted and ground-truth flood regions:

$$IoU = \frac{TP}{TP+FP+FN} \quad (11)$$

IoUis particularly useful for assessing the spatial overlap between predicted inundation and the reference flood mask.

3.7.7 Receiver Operating Characteristic and AUC

ROC analysis was performed by varying the classification threshold and calculating the corresponding true-positive rate and false-positive rate. The area under the ROC curve (AUC) provides a threshold-independent measure of the model's discrimination capability between flood and background pixels.

3.8 Confusion Matrix Analysis

Pixel-level confusion matrices were generated for each model to examine the distribution of true-positive, true-negative, false-positive, and false-negative predictions. The analysis provides additional insight into the differences between flood detection sensitivity and false-alarm behaviour.

In particular, the confusion matrices allow the higher recall observed for Attention U-Net and Weighted U-Net

to be examined alongside their precision values and therefore provide a more complete interpretation of their segmentation behaviour.

3.9 Computational Efficiency

Computational efficiency was evaluated using two measurements: trainable parameter count and inference time per image.

The number of trainable parameters provides an indication of model size and memory requirements. Inference time measures the elapsed time required to generate a segmentation prediction for an individual input image under the experimental computing environment.

These measures are reported alongside segmentation performance to evaluate whether improved segmentation quality is accompanied by increased computational requirements.

3.10 Experimental Reproducibility

To support reproducibility, the experiments were conducted using the same dataset partition, input resolution, training framework, and evaluation procedure for the five investigated architectures. The computing environment used an NVIDIA Tesla T4 GPU and the PyTorch framework.

4. Results And Discussion

The experimental results of the five semantic segmentation architectures tested with the test subset SEN1Floods11 are shown in this section. Analysis is done based on quantitative segmentation performance, ROC-AUC, pixel-level confusion matrices, qualitative prediction maps, and computational characteristics. Because flood segmentation is affected by class imbalance, the results are interpreted using multiple complementary metrics rather than accuracy alone.

4.1 Training and Validation Performance

The training and validation histories were examined to assess the convergence behaviour of the five investigated architectures. The models were trained using their respective configurations summarized in Table 4. The resulting validation-loss curves provide an indication of optimization behaviour during the model-specific training periods.

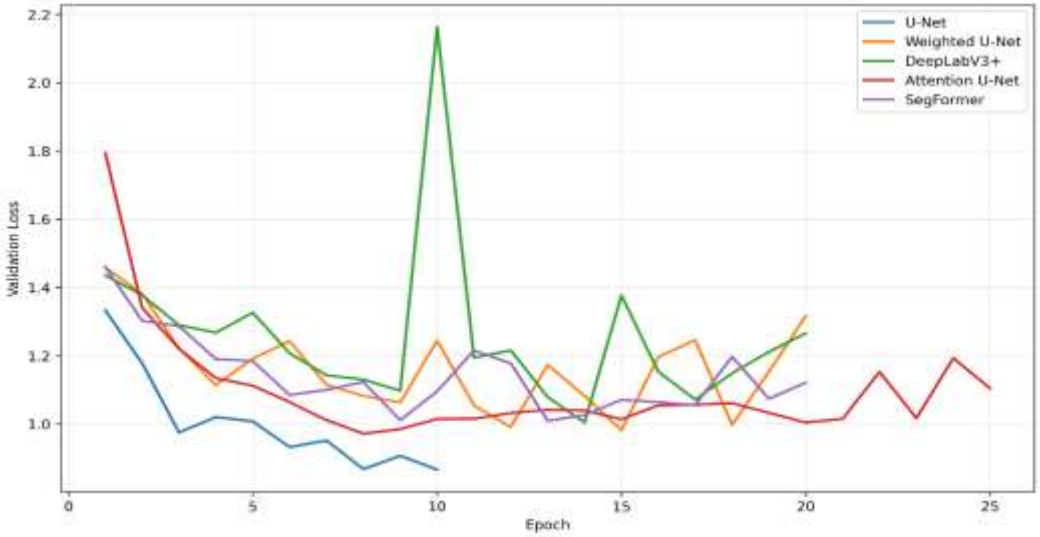

Fig. 4. Validation loss curves of the evaluated semantic segmentation models.

The validation loss curves provide an indication of the convergence behaviour of the investigated models during training. Because convergence behaviour alone does not establish generalization performance, the independent test set was used for the final quantitative comparison

4.2 Quantitative Segmentation Performance

The quantitative performance of the five investigated models is summarized in Table 5. Accuracy, precision, recall, F1/Dice, and IoU were calculated on the independent test subset.

Table 5. Quantitative performance comparison of the evaluated flood-segmentation models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1/Dice (%)	IoU (%)	OC-AUC
U-Net	94.56	67.16	58.78	62.69	45.66	..902
Weighted U-Net	93.56	57.12	68.88	62.45	45.40	..921
DeepLabV3+	91.57	46.84	63.18	53.80	36.80	9.905

Attention U-Net	93.84	58.68	70.17	63.91	46.97	0.914
SegFormer	94.72	65.92	66.43	66.17	49.45	0.951

The results show clear differences among the evaluated architectures. SegFormer achieved the highest accuracy (94.72%), F1/Dice score (66.17%), and IoU (49.45%), indicating the strongest overall overlap between the predicted and reference flood regions among the evaluated models. It also obtained the highest ROC-AUC of 0.951, indicating strong discrimination between flood and background pixels.

U-Net achieved an accuracy of 94.56%, which is close to the SegFormer result. However, the lower F1/Dice score of 62.69% and IoU of 45.66% show that overall pixel accuracy alone is not a criterion to fully assess the quality of flood-region segmentation.

A comparison of the two models revealed that Weighted U-Net outperformed U-Net with a recall of 68.88% compared to 57.49%, demonstrating better sensitivity towards detecting flood pixels. However, its precision decreased to 57.12%. Weighted optimization apparently generated more detections of inundated pixels, and more false positive predictions, as is indicated by this behaviour.

Among the models evaluated, Attention U-net has the highest value of Recall of 70.17%. This means that its attention mechanism helped the model to be able to identify a higher percentage of the reference flood pixels. But with a lower precision of 58.68%, it showed a compromise between flood sensitivity and false-positive suppression when compared to U-Net and SegFormer.

DeepLabV3+ obtained an accuracy of 91.57%, precision of 46.84%, recall of 63.18%, F1/Dice of 53.80%, and IoU of 36.80%. It offers multi-scale feature extraction, meaning that it offers contextual information at various spatial scales, but the performance based on overlap was less than that of the other architectures that were evaluated.

In general, SegFormer was the best in both the accuracy and F1/Dice and IoU, and Attention U-Net was the best in recall and ROC-AUC. These results highlight the varying performance of the models and that a single measure alone cannot be used to evaluate flood segmentation quality.

4.3 Comparative Metric Analysis

To provide a visual comparison of the quantitative results, the principal segmentation metrics are presented in Fig. 5.

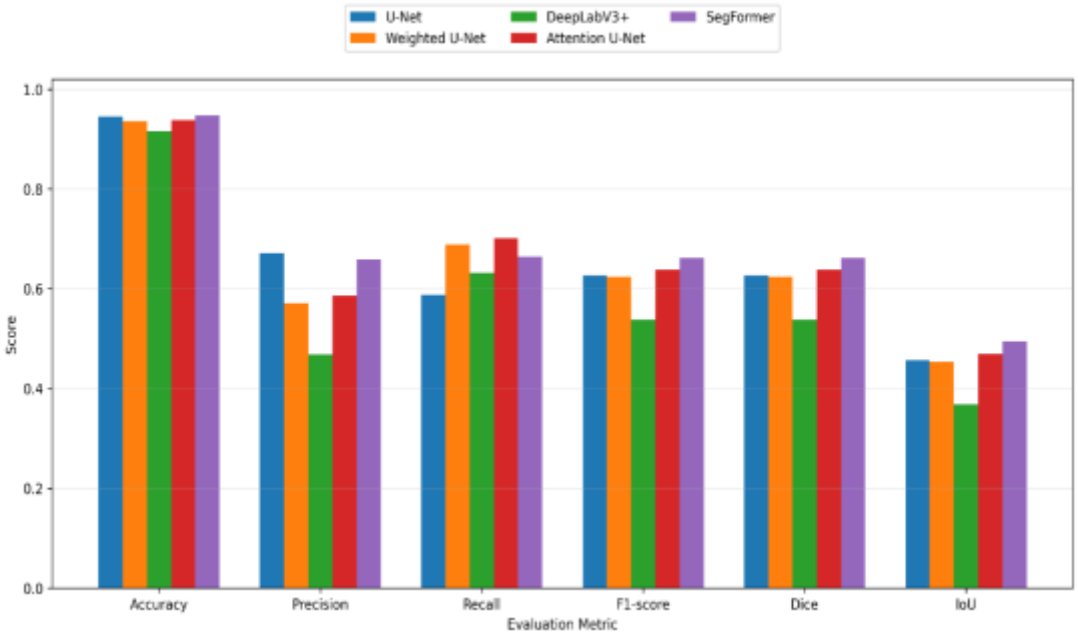

Fig. 5. Comparative performance of the evaluated segmentation models across accuracy, precision, recall, F1/Dice coefficient, and IoU.

Figure 5 shows the best overall performance of SegFormer based on the evaluated segmentation metrics. Specifically, SegFormer achieves the best accuracy, F1/Dice and IoU scores, while U-Net excels in precision and Attention U-Net excels in recall score. It can be seen that SegFormer achieves the best overall performance, especially on overlap-based metrics like Dice and IoU.

The results of the recall indicate a different pattern. The highest recall is obtained by Attention U-Net whereas Weighted U-Net follows. This means that attention and class weighting mechanisms can be used to enhance the identification of flood pixels. But they also have lower precision, suggesting that greater flood sensitivity comes with the trade-off of more false-positive predictions.

Such a compromise is relevant especially in flood mapping. A model with a high number of flood-pixels could result in high recall but also overestimate flood inundation. On the other hand, a high precision model may not capture some real flooded areas. Therefore, F1/Dice and IoU provide useful complementary measures as they consider both false positive and false negative segmentation errors.

4.4 ROC-AUC Analysis

ROC analysis was performed to assess the discrimination capability of each model across different classification thresholds.

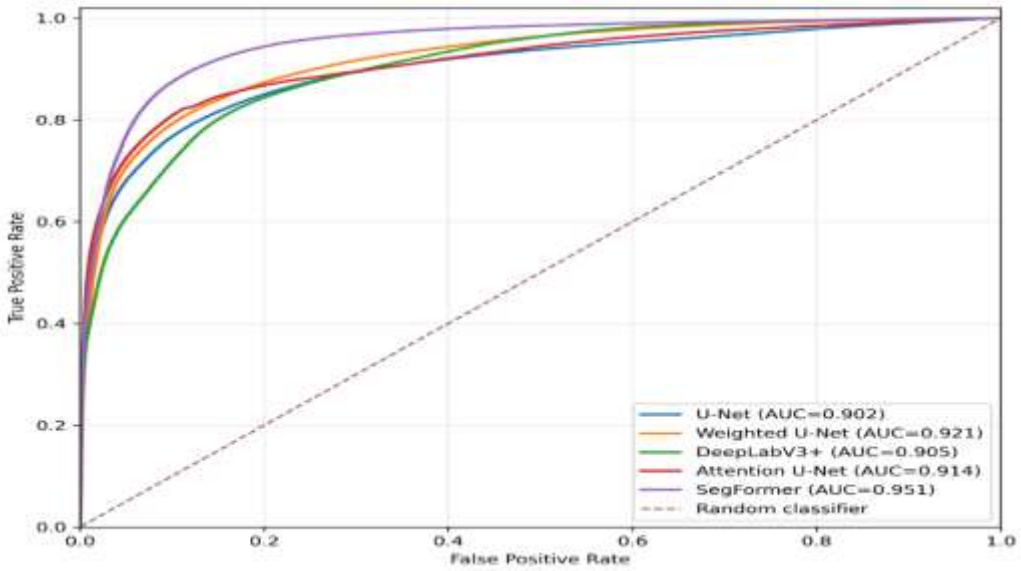

Fig. 6. Receiver operating characteristic (ROC) curves for pixel-level flood classification using the evaluated segmentation models.

The ROC-AUC results are summarized in Table 5. SegFormer achieved the highest AUC of 0.951, followed by Weighted U-Net (0.921), Attention U-Net (0.914), DeepLabV3+ (0.905), and U-Net (0.902). These results indicate that SegFormer provides the strongest overall discrimination capability among the evaluated models across varying classification thresholds.

The higher AUC obtained by SegFormer indicates stronger discrimination between flood and background pixels across the evaluated threshold range. This observation is consistent with its higher F1/Dice and IoU values.

Weighted U-Net and Attention U-Net also achieved relatively high AUC values. Their performance, combined with their higher recall values, indicates that these architectures can effectively identify flood pixels. However, their lower precision values suggest that some of the additional flood detections correspond to false-positive regions.

The ROC analysis therefore provides additional evidence that SegFormer offers a strong balance between flood detection and background discrimination within the evaluated test set.

4.5 Pixel-Level Confusion Matrix Analysis

Confusion matrices were generated to examine the pixel-level classification behaviour of the five models.

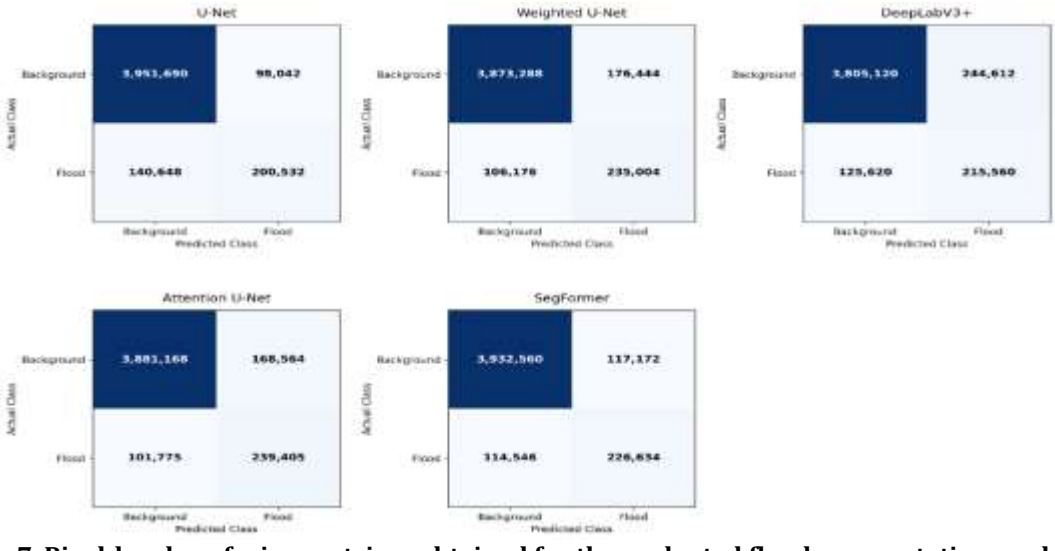

Fig. 7. Pixel-level confusion matrices obtained for the evaluated flood-segmentation models.

The confusion matrices provide detailed information about true-positive, true-negative, false-positive, and false-negative predictions.

The results indicate that the evaluated models correctly classify a large proportion of background pixels, which contributes to their relatively high overall accuracy. This observation also demonstrates why accuracy must be interpreted together with flood-specific metrics such as precision, recall, Dice, and IoU.

Attention U-Net and Weighted U-Net show stronger flood-pixel detection behaviour, consistent with their higher recall values. However, the corresponding increase in false-positive predictions contributes to their lower precision.

SegFormer demonstrates a comparatively balanced prediction pattern, with strong flood-pixel detection while limiting erroneous flood predictions. This behaviour is consistent with its higher F1/Dice and IoU values.

The confusion-matrix analysis therefore supports the quantitative results and provides a more detailed explanation of the precision–recall trade-offs observed among the models.

4.6 Qualitative Segmentation Analysis

Quantitative metrics provide an objective numerical assessment of segmentation performance; however, visual examination of the predicted segmentation maps is important for assessing flood-boundary continuity, fragmented regions, missed inundation, and false-positive predictions.

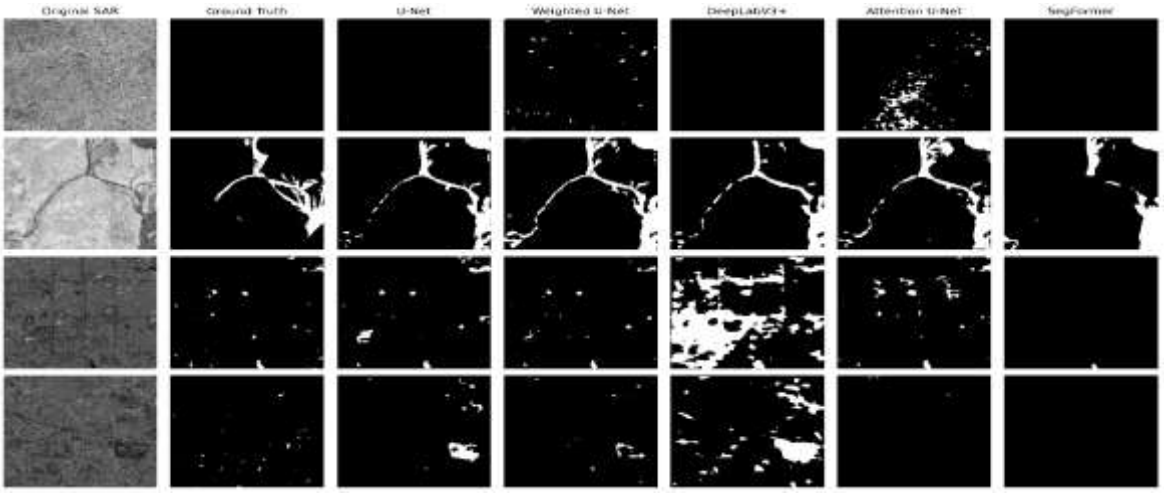

Fig. 8. Qualitative comparison of flood-segmentation predictions generated by the evaluated models for representative Sentinel-1 SAR images.

Figure 8 provides a visual comparison of the segmentation behaviour of the five evaluated architectures. The columns present the input Sentinel-1 SAR image, corresponding ground-truth flood mask, and predictions generated by U-Net, Weighted U-Net, DeepLabV3+, Attention U-Net, and SegFormer, respectively. The ground-truth masks provide the reference flood extent against which the predicted segmentation maps can be visually assessed.

The U-Net predictions identify the major inundated regions but may exhibit local discontinuities or miss smaller flood structures in complex scenes. Weighted U-Net improves sensitivity to flood regions, although some additional false-positive areas may occur. DeepLabV3+ produces comparatively smooth segmentation maps but may fail to recover some narrow or fragmented inundated regions. Attention U-Net demonstrates improved flood-region detection and boundary localization in several representative cases, consistent with its relatively high recall.

SegFormer produces comparatively coherent predictions in several representative cases and shows good spatial correspondence with the reference masks for scenes containing larger or structurally continuous inundated regions. However, some fragmented or low-extent flood regions are not completely recovered. Its hierarchical Transformer encoder provides multi-scale contextual representations that may contribute to its strong overlap-based performance, although the qualitative examples also demonstrate that its behaviour varies across scene characteristics.

4.7 Computational Efficiency

Segmentation accuracy is only one consideration when selecting a model for large-scale Earth-observation applications. Model size and inference time are also relevant, particularly when large numbers of satellite image tiles need to be processed.

The computational characteristics of the five models are presented in Table 6.

Table 6. Computational complexity and inference performance

Model	Trainable parameters (million)	Inference time (ms/image)

U-Net	24.43	6.23
Weighted U-Net	24.43	7.58
DeepLabV3+	22.43	372.27
Attention U-Net	13.22	14.34
SegFormer	**3.71**	**11.57**

From the computational perspective, SegFormer achieved the lowest parameter count among the evaluated models, with only 3.71 million trainable parameters. However, parameter count alone does not directly determine inference speed. U-Net achieved the lowest measured inference time of 6.23 ms/image, whereas SegFormer required 11.57 ms/image. Therefore, SegFormer can be considered parameter-efficient rather than universally computationally faster. In contrast, DeepLabV3+ exhibited the highest inference time at 372.27 ms/image. These results demonstrate that model size and inference speed represent distinct aspects of computational efficiency and should be evaluated separately when considering deployment suitability.

Therefore, the results demonstrate an important performance–efficiency trade-off. SegFormer provides the smallest parameter footprint and the strongest segmentation performance, whereas U-Net provides the fastest measured inference among the evaluated models.

4.8 Performance–Efficiency Trade-off

The combination of segmentation and computational results provides a more complete picture of model behaviour.

SegFormer achieved:

- **94.72% accuracy**
- **66.17% F1/Dice**
- **49.45% IoU**
- **0.951 ROC-AUC**
- **3.71 million parameters**

These results indicate that SegFormer provides the strongest overall segmentation performance among the evaluated architectures while also having the smallest parameter count.

However, U-Net had the lowest F1/Dice and IoU scores while having the fastest measured inference time of 6.23 ms/image.

For U-Net, the highest Recall (70.17%) was obtained, it may be considered to be used when maximizing flood pixels detection is more critical that minimizing False Positive predictions.

Weighted U-Net exhibited a similar sensitivity-oriented trend with a recall of 68.88% and lower precision compared to SegFormer.

Overall, DeepLabV3+ had the lowest segmentation performance of five architectures, while also being significantly slower in terms of measured inference time.

These results suggest that the choice of the model depends on the type of operation that the modeler is aiming for. In this case, if the main requirement is to achieve good overall segmentation quality while maintaining a compact model, SegFormer could be a good choice. Attention U-Net can be beneficial if maximum sensitivity of flood-pixels is desired. For inference time that is dominated by minimal measured requirements, an alternative is U-Net.

4.9 Comparison with Existing Studies

Care should be taken when comparing the results presented here directly with results from other publications as differences in data sets, spatial splits, pre-processing, input channels, definition of labels, and evaluation procedures can significantly affect reported results.

The original SEN1Floods11 work established an important benchmark for deep-learning-based flood mapping from Sentinel-1 data (Bonafilia et al. 2020). Subsequent studies have introduced specialized attention-based and transformer-based methods, including DAM-Net and DeepSARFlood (Saleh et al. 2024; Sharma and Saharia 2025), demonstrating the increasing use of contextual and transformer-based representations for SAR flood mapping.

Table 7. Comparison of the present study with representative SAR flood-segmentation studies

Study	Dataset	Architecture	Evaluation Focus	Directly Comparable?
Bonafilia et al. (2020)	SEN1Floods11	FCN/CNN	Flood mapping benchmark	No
Nemni et al. (2020)	SAR flood imagery	FCNN	Rapid segmentation	Limited
Choi et al. (2022)	Sentinel-1	U-Net/HRNetV2	Flood segmentation	Limited

Ghosh et al. (2024)	Sentinel-1/NASA benchmark	Nested U-Net	Flood detection	Limited
Saleh et al. (2024)	S1GFloods	DAM-Net	Transformer flood detection	Limited
Zhao et al. (2024)	UrbanSARFloods	CNN	Urban/open-area flood mapping	Limited
Sharma &Saharia (2025)	SAR flood imagery	DeepSARFlood	ViT ensemble + uncertainty	Limited
Present study	**SEN1Floods11 subset**	**5 architectures**	**Segmentation + efficiency**	**Primary controlled comparison**

The primary distinction of the present study is therefore not a claim of achieving the highest numerical result ever reported for SAR flood segmentation. Instead, the contribution lies in the controlled comparison of five segmentation architectures under a common experimental framework, together with simultaneous consideration of segmentation performance and computational characteristics. This distinction is important because published flood-segmentation results obtained using different datasets and protocols should not be treated as directly interchangeable.

4.10 Discussion

The experimental results provide several important observations regarding the use of CNN and transformer-based architectures for Sentinel-1 SAR flood segmentation.

First, SegFormer achieved the strongest overall performance across the principal overlap-based metrics. Its F1/Dice score of 66.17% and IoU of 49.45% were higher than those of the four CNN-based architectures. These results suggest that the hierarchical contextual representation provided by the SegFormer encoder is useful for distinguishing inundated regions from heterogeneous SAR backgrounds.

Second, the results demonstrate that accuracy alone is insufficient for assessing flood segmentation quality. U-Net achieved 94.56% accuracy, which was only slightly below the 94.72% achieved by SegFormer. However, SegFormer produced higher F1/Dice and IoU values. This difference illustrates the effect of background-pixel dominance in a binary flood-segmentation problem.

Third, Attention U-Net achieved the highest recall at 70.17%. This indicates that attention-based feature selection can help identify a larger proportion of actual flood pixels. However, its precision of 58.68% was lower than the precision achieved by U-Net and SegFormer. The result therefore demonstrates the inherent trade-off between flood sensitivity and false-positive suppression.

Fourth, Weighted U-Net also improved recall relative to the baseline U-Net. Its recall increased to 68.88%, compared with 58.78% for U-Net. However, precision decreased from 67.16% to 57.12%. This observation is consistent with the intended effect of emphasizing the flood class during optimization: increased attention to the minority class can improve sensitivity while potentially increasing false-positive predictions.

Finally, SegFormer achieved the best ROC-AUC score of 0.951. The ROC analysis shows that the model was able to keep discriminating between flood and background pixels at various thresholds in the test set examined. This is in addition to the F1/Dice and IoU results and is an extra proof of its balanced classification behaviour.

In terms of computation, SegFormer had the least number of trainable parameters, 3.71M. This is significantly smaller than the CNN architectures explored in this study. The inference measurements, however, show that the number of parameters is not necessarily the determining factor for the shortest inference time. The measured inference time of U-Net was the lowest (6.23 ms/image) and that of SegFormer was 11.57 ms/image. This separation is significant on the occasion of deployment assessment.

The model with the highest inference time of the models was DeepLabV3+ with 372.27ms/image. It shows that multi-scale contextual processing can have significant computational costs depending on the implementation and hardware configuration.

In general, the experimental results indicate that SegFormer has the most comprehensive segmentation performance and is the most parameter-efficient, while the CNN models still have their merits in terms of flood-pixel sensitivity or inference speed. The results thus agree with the concept of architecture-specific evaluation, which is not supposed to use a single model that would be best for all operational needs.

4.11 Summary of Findings

The principal findings of the experimental evaluation are summarized below:

1. SegFormer achieved the highest overall accuracy (94.72%), F1/Dice (66.17%), IoU (49.45%), and ROC-AUC (0.951).
2. Attention U-Net achieved the highest recall (70.17%), demonstrating strong flood-pixel sensitivity.
3. Weighted U-Net achieved 68.88% recall, indicating improved flood detection compared with baseline U-Net, although its precision was lower.
4. U-Net achieved the lowest measured inference time (6.23 ms/image) among the evaluated models.

5. SegFormer had the smallest model size, with only 3.71 million trainable parameters.
6. DeepLabV3+ showed substantially higher inference time, despite its established multi-scale contextual representation.
7. The results demonstrate that segmentation quality, flood sensitivity, model size, and inference speed do not necessarily improve simultaneously, making multi-dimensional evaluation important for SAR flood-mapping applications.

5. LIMITATIONS AND FUTURE RESEARCH

Although the comparative experiments demonstrate the effectiveness of the investigated segmentation architectures, several limitations should be considered when interpreting the results. These limitations also provide directions for extending the present work toward more generalizable and operational flood-mapping systems.

5.1 Dataset and Generalization Limitations

The experimental evaluation was conducted using a subset of the SEN1Floods11 benchmark dataset. Although SEN1Floods11 provides geographically diverse Sentinel-1 observations, the present experiments used 446 image–mask pairs, consisting of 312 training, 67 validation, and 67 testing pairs.

Therefore, the reported results should be interpreted within the scope of this experimental dataset and should not automatically be generalized to all Sentinel-1 flood events or geographical regions.

Future experiments should evaluate the trained models on additional independent SAR flood datasets and geographically distinct flood events. Such cross-dataset evaluation would provide stronger evidence regarding the ability of SegFormer and the CNN baselines to generalize beyond the training distribution.

5.2 SAR Data Representation

In the present study, the focus is on Sentinel-1 SAR imagery. The major benefit of SAR for flood monitoring is the possibility of receiving observations when optical imagery may not be available. However, vegetation, soil moisture, surface roughness, urban structures, terrain and speckle-related variations can impact SAR backscatter.

Future studies might involve more detailed sar representations such as further polarization information, multi-temporal observations and further sar derived features. This information might also enhance the ability to discriminate flooded areas from areas that have similar radar responses.

5.3 Limited Model Selection

Five segmentation architectures were considered in this study: U-Net, Weighted U-Net, DeepLabV3+, Attention U-Net, and SegFormer. Although these models make for an interesting comparison between conventional CNNs and a lightweight transformer architecture, they don't necessarily cover the entire spectrum of modern approaches to semantic segmentation.

In the future, architectures like Swin Transformer, Mask2Former variants or hybrid CNN–Transformer models (Liu et al. 2021; Cheng et al. 2022) could be investigated. These comparisons would identify if the benefits of SegFormer are consistent, when compared to more recent transformer-based segmentation architectures.

5.4 Computational Evaluation

The number of trainable parameters and inference time under the experimental computing environment were taken into account in the computational analysis. While these measurements offer valuable insights into model complexity and processing speed, inference time may vary based on hardware, software implementation, batch size, input dimensions, and the method of measurement.

So, the reported inference times should be viewed as the times observed in the experimental setup and not as universal inference times.

Further studies would be beneficial to assess the performance of the computations on various hardware such as CPUs, GPUs, and edge devices. To get a more complete evaluation of the feasibility of deployment, memory consumption, energy consumption, throughput, and latency may also be examined.

5.5 Lack of Multimodal Information

As of now, the framework is based on Sentinel-1 SAR observations as the main image source. Though SAR is proving very useful in flood monitoring, other data, e.g. optical and multispectral data, can be used to complement information about vegetation, surface characteristics, and land-cover conditions.

The future research could then be focused on Sentinel-1 with Sentinel-2 data fusion. In heterogeneous environments, a multimodal approach could be used, with optical data complementing SAR data to help delineate flood boundaries.

5.6 Multi-Temporal Flood Monitoring

The current experimental design assesses each image–mask pair. Floods are, however, a dynamic phenomenon and change over time.

Observations obtained prior to, during and after the flood events may be integrated into a multi-temporal framework. Temporal information may be useful to enhance the separation of permanent water from newly flooded areas.

Next, temporal SAR sequences and change-aware segmentation models for dynamic flood monitoring will be investigated.

5.7 Operational Deployment and Explainability

Even though the model performance has been assessed both quantitatively and qualitatively in the present study, explainability methods were not explicitly included in the experiments.

Future research could explore explainable AI approaches for generating visualizations of the attention, feature attributions, and gradient-based interpretation methods to understand which spatial features affect flood predictions.

Furthermore, the small parameter size of SegFormer motivates exploring the use of edge or near-edge computing platforms. This deployment may enable the timely processing of satellite observations, during disaster-response operations.

6. Conclusion

A controlled comparison of five semantic segmentation architectures for flood mapping using Sentinel-1 SAR data was presented in this study. The models studied were U-Net, Weighted U-Net, DeepLabV3+, Attention U-Net, SegFormer. All models have been tested in a common experimental framework, which is based on the benchmark SEN1Floods11, common image preprocessing, model-specific training parameters, and an independent test set.

The experimental results show that there are different strengths in the architectures based on the criterion used for evaluation. Overall, SegFormer outperformed the other models in segmentation, with the highest accuracy of 94.72%, F1/Dice score of 66.17%, IoU of 49.45%, and ROC-AUC of 0.951. Attention U-Net showed the best recall at 70.17%, meaning that it was better able to detect a higher percentage of flood pixels. U-Net attained the quickest measured inference time at 6.23 ms/image, and SegFormer resulted in the smallest model size with 3.71 million trainable parameters.

The results also have shown that accuracy is not enough to assess flood segmentation. Precision, recall, F1/dice and IoU provide valuable insights into the trade-offs between flood sensitivity and false-positive suppression. Thus, the quantitative, qualitative, ROC, confusion-matrix, and computational analyses are combined to give a more thorough evaluation of the architectures investigated.

The results indicate that SegFormer offers a good trade-off between segmentation performance and model size for the tested Sentinel-1 SAR flood-segmentation task. However, interpretation should be done in the context of the experimental data set and computing environment. Further and wider validation is needed with other geographical regions and other SAR flood datasets to conclude more strongly that there is operational generalization.

Future research will focus on the cross-dataset validation, multi-temporal SAR observations, data fusion between Sentinel-1/Sentinel-2, hybrid CNN–Transformer architectures, explainable AI methods, and edge-oriented deployments. These extensions could potentially further strengthen the robustness, interpretability and practical applicability of deep learning based flood mapping systems.

DECLARATIONS:

Funding:
The auth.orsrece.ived no finan.cialsupp.ort for the rese.arch, autho.rship, or public.ation of this arti.cle.

Compe.tinginter.ests:
The auth.orsdecl.are that they have no compe.tingfinan.cial or non-finan.cialinter.ests that cou.ld have influ.enced the work repo.rted in this arti.cle.

Data availability:
The SEN1Flo.ods11 data.set used in this stu.dy is publ.iclyavail.able from the data.setrepos.itorydescr.ibed by Bonaf.ilia et al. (2020). The experi. Mental sub. set, data.setpartit.ioninginform.ation, trai.ned-mod.elconfigu.ration, and evalu.ationproce.dures used in this stu.dy are avail.able from the corresp.ondingaut.hor on reaso.nablerequ.est.

Author contributions:
K. Lakshman Kumar: Conceptualization, methodology, software, data curation, formal analysis, visualization, and writingoriginal draft. RangaswamySirisati: supervision, validation, review and editing. Both authors read and approved the final manuscript.

References

1. Amitrano D, Di Martino G, Iodice A, Riccio D, Ruello G (2018) Unsupervised rapid flood mapping using Sentinel-1 GRD SAR images. IEEE Trans Geosci Remote Sens 56(6):3290–3299. https://doi.org/10.1109/TGRS.2018.2797536
2. Bai Y, Wu W, Yang Z, Yu J, Zhao B, Liu X, Yang H, Mas E, Koshimura S (2021) Enhancement of detecting

permanent water and temporary water in flood disasters by fusing Sentinel-1 and Sentinel-2 imagery using deep learning algorithms: demonstration of Sen1Floods11 benchmark datasets. Remote Sens 13(11):2220. https://doi.org/10.3390/rs13112220
3. Bonafilia D, Tellman B, Anderson T, Issenberg E (2020) Sen1Floods11: a georeferenced dataset to train and test deep learning flood algorithms for Sentinel-1. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp 835–845. https://doi.org/10.1109/CVPRW50498.2020.00113
4. Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation. In: European Conference on Computer Vision, pp 801–818. https://doi.org/10.1007/978-3-030-01234-2_49
5. Cheng B, Misra I, Schwing AG, Kirillov A, Girdhar R (2022) Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 1290–1299.
6. Choi S, Park G, Kang J, Kim G, Youn Y, Lee YW (2022) Flood detection from Sentinel-1 SAR images using U-Net and HRNetV2 models. The Geographical Journal of Korea 56(4):409–419. https://doi.org/10.22905/kaopqj.2022.56.4.8
7. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N (2021) An image is worth 16×16 words: transformers for image recognition at scale. In: International Conference on Learning Representations.
8. Ghosh B, Garg S, Motagh M (2022) Automatic flood detection from Sentinel-1 data using deep learning architectures. ISPRS Ann Photogramm Remote Sens Spatial Inf Sci V-3-2022:201–208. https://doi.org/10.5194/isprs-annals-V-3-2022-201-2022
9. Ghosh B, Garg S, Motagh M, Martinis S (2024) Automatic flood detection from Sentinel-1 data using a nested UNet model and a NASA benchmark dataset. PFG J Photogramm Remote Sens Geoinf Sci 92:1–18. https://doi.org/10.1007/s41064-024-00275-1
10. Kingma DP, Ba J (2015) Adam: a method for stochastic optimization. In: International Conference on Learning Representations.
11. Li W, Lee H, Wang S, Hsu CY, Arundel ST (2023) Assessment of a new GeoAI foundation model for flood inundation mapping. In: Proceedings of the 6th ACM SIGSPATIAL International Workshop on Advances in Geographic Information Systems, pp 102–109. https://doi.org/10.1145/3615886.3627747
12. Li Y, Martinis S, Wieland M (2019) Urban flood mapping with an active self-learning convolutional neural network based on TerraSAR-X intensity and interferometric coherence. ISPRS J Photogramm Remote Sens 152:178–191. https://doi.org/10.1016/j.isprsjprs.2019.04.014
13. Liang J, Liu D (2020) A local thresholding approach to flood water delineation using Sentinel-1 SAR imagery. ISPRS J Photogramm Remote Sens 159:53–62. https://doi.org/10.1016/j.isprsjprs.2019.10.017
14. Lin TY, Goyal P, Girshick R, He K, Dollár P (2017) Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp 2980–2988. https://doi.org/10.1109/ICCV.2017.324
15. 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. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 10012–10022. https://doi.org/10.1109/ICCV48922.2021.00986
16. Loshchilov I, Hutter F (2019) Decoupled weight decay regularization. In: International Conference on Learning Representations (ICLR).
17. Lv J, Shen Q, Lv M, Li Y, Shi L, Zhang P (2023) Deep learning-based semantic segmentation of remote sensing images: a review. Front Ecol Evol 11:1201125. https://doi.org/10.3389/fevo.2023.1201125
18. Nemni E, Bullock J, Belabbes S, Bromley L (2020) Fully convolutional neural network for rapid flood segmentation in synthetic aperture radar imagery. Remote Sens 12(16):2532. https://doi.org/10.3390/rs12162532
19. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Kopf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S (2019) PyTorch: an imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems 32.
20. Pech-May F, Aquino-Santos R, Álvarez-Cárdenas O, Lozoya Arandia J, Rios-Toledo G (2024) Segmentation and visualization of flooded areas through Sentinel-1 images and U-Net. IEEE J Sel Top Appl Earth Obs Remote Sens 17:8996–9008. https://doi.org/10.1109/JSTARS.2024.3387452
21. Ronneberger O, Fischer P, Brox T (2015) U-Net: convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, pp 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
22. Roy, A. G., Navab, N., &Wachinger, C. (2018). Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks. Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, 421–429. Springer. https://doi.org/10.1007/978-3-030-00928-1_48.

23. Saleh T, Weng X, Holail S, Hao C, Xia GS (2024) DAM-Net: flood detection from SAR imagery using differential attention metric-based vision transformers. ISPRS J Photogramm Remote Sens 212:440–453. https://doi.org/10.1016/j.isprsjprs.2024.05.018
24. Schlemper J, Oktay O, Schaap M, Heinrich M, Kainz B, Glocker B, Rueckert D (2019) Attention gated networks: learning to leverage salient regions in medical images. Med Image Anal 53:197–207.
25. Sharma NK, Saharia M (2025) DeepSARFlood: rapid and automated SAR-based flood inundation mapping using vision transformer-based deep ensembles with uncertainty estimates. Sci Remote Sens 11:100203. https://doi.org/10.1016/j.srs.2025.100203
26. Shelhamer E, Long J, Darrell T (2017) Fully convolutional networks for semantic segmentation. IEEE Trans Pattern Anal Mach Intell 39(4):640–651. https://doi.org/10.1109/TPAMI.2016.2572683
27. Sudre CH, Li W, Vercauteren T, Ourselin S, Cardoso MJ (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, pp 240–248. https://doi.org/10.1007/978-3-319-67558-9_28
28. Taha AA, Hanbury A (2015) Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging 15:29. https://doi.org/10.1186/s12880-015-0068-x
29. Tan M, Le Q (2019) EfficientNet: rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning, PMLR 97:6105–6114.
30. Tay CWJ, Yun SH, Chin ST, Bhardwaj A, Jung J, Hill EM (2020) Rapid flood and damage mapping using synthetic aperture radar in response to Typhoon Hagibis, Japan. Sci Data 7:100. https://doi.org/10.1038/s41597-020-0443-5
31. Twele A, Cao W, Plank S, Martinis S (2016) Sentinel-1-based flood mapping: a fully automated processing chain. Int J Remote Sens 37(13):2990–3004. https://doi.org/10.1080/01431161.2016.1192304
32. Xie E, Wang W, Yu Z, Anandkumar A, Alvarez JM, Luo P (2021) SegFormer: simple and efficient design for semantic segmentation with transformers. Adv Neural Inf Process Syst 34.
33. Zhao J, Pelich R, Hostache R, Matgen P, Cao S, Wagner W, Chini M (2021) Deriving exclusion maps from C-band SAR time-series in support of floodwater mapping. Remote Sens Environ 265:112668. https://doi.org/10.1016/j.rse.2021.112668
34. Zhao J, Xiong Z, Zhu XX (2024) UrbanSARFloods: Sentinel-1 SLC-based benchmark dataset for urban and open-area flood mapping. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp 419–429. https://doi.org/10.1109/CVPRW63382.2024.00047
35. Zhu XX, Tuia D, Mou L, Xia GS, 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(4):8–36. https://doi.org/10.1109/MGRS.2017.2762307