Ideas from Works Related to Water Body Segmentation from Sentinel S1/S2 Images

Jaya Sreevalsan Nair,

GVCL, IIIT-B

March 11, 2024

Paper-01

A Three-Stage Deep Learning Model for Accurate Retinal Vessel Segmentation (2019)

- Vessels are classified as thin or thick: those less than 3-pixel thickness are “thin”, and otherwise, “thick”.
- 3 stages:
  - Thick segmenter with 1 pooling layer,
  - Thin segmenter with multiple pooling layers (adapted a fully convolutional network FCN), and
  - concatenation using fusion segmenter.

https://ieeexplore.ieee.org/abstract/document/8476171

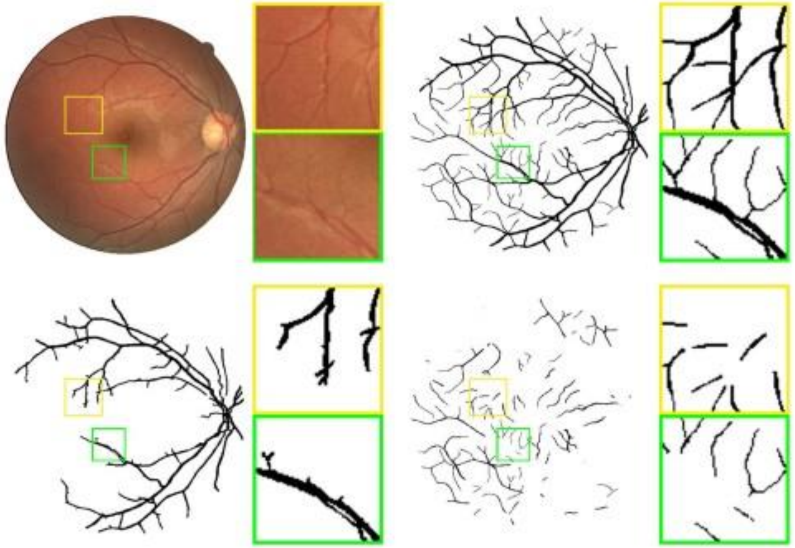


Fig. 1. Analysis of the retinal vessel segmentation problem. Row 1: from left to right, the fundus image and the enlarged patches, and the manual annotation and the annotations of the two fundus image patches. Row 2: from left to right, the manually annotated thick vessels and the annotated thick vessels in the two fundus image patches, and the manually annotated thin vessels and the annotated thin vessels in the two fundus image patches.

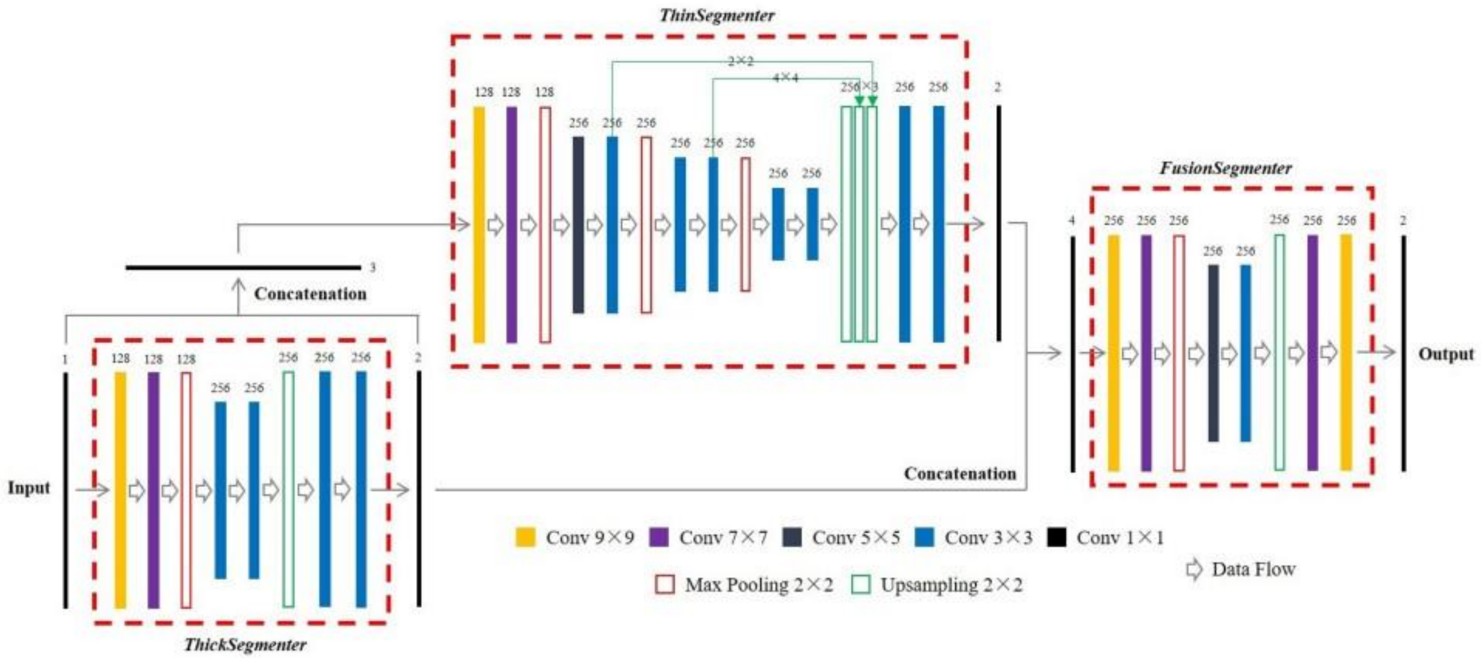


Fig. 2. The overview of the proposed three-stage deep learning framework. The framework consists of three separate models, namely *ThickSegmenter* for thick vessel segmentation, *ThinSegmenter* for thin vessel segmentation and *FusionSegmenter* for vessel fusion respectively.

Paper-02

Automatic Segmentation of River and Land in SAR Images: A Deep Learning Approach (2019)

- For segmentation of surface, river water and land.
  - Lee filter used for reducing the effect of speckle noise is optimal, but results in loss of image information
- Two different implementation of U-Net architecture is studied on SAR images,
  - A U-Net is trained from scratch (Vanilla U-Net)
  - A U-Net model with pretrained weights are used (Transfer U-Net), as learnt by the U-Net model on the ISBI 2015 Cell Tracking dataset
- Experimental results show that the both architectures gave similar performance in terms of F1 score, pixel accuracy and mean IoU.
  - Vanilla U-Net performs slightly better.
  - Transfer U-Net identifies very minute details in the image such as small rivers.

https://uverma.github.io/assets/pdf/AIKE-Vaibhav-2019.pdf

Paper-03

Improved Semantic Segmentation of Water Bodies and Land in SAR Images Using Generative Adversarial Networks (2020)

- Improve the U-Net methodology by augmenting the dataset of manually annotated images using Generative Adversarial Networks (GANs)
  - Deep Convolutional GANs (DC-GANs) to generate SAR images and corresponding masks
- Proposed Model:
  - The generator is given the latent space (raw SAR and ground truth masks)
  - The discriminator is fed real images and generated images.
  - Outcome: New SAR image patches
- Use raw and generated SAR images in U-Net model

https://www.worldscientific.com/doi/abs/10.1142/S1793351X20400036

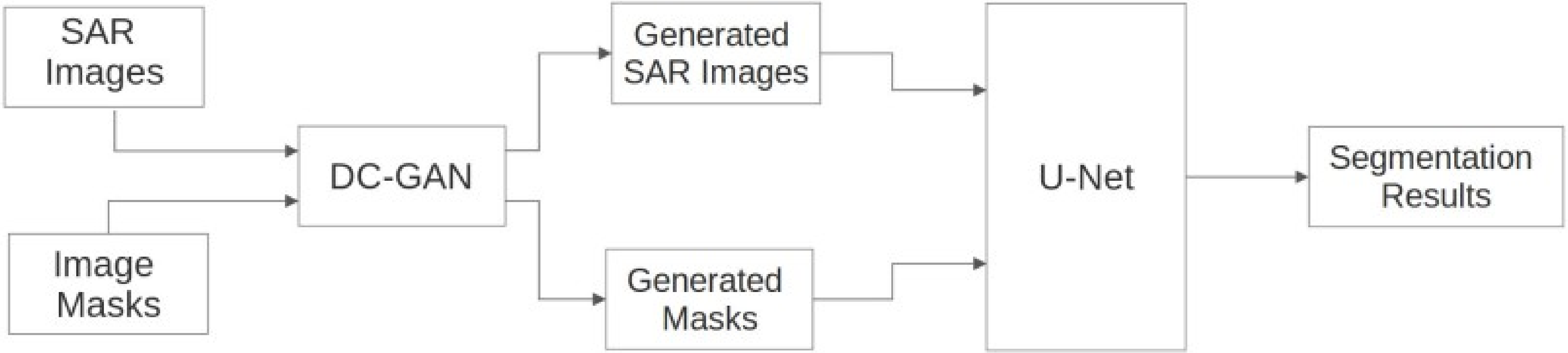


Fig. 2. Using DC-GANs to augment segmentation process of U-Net.

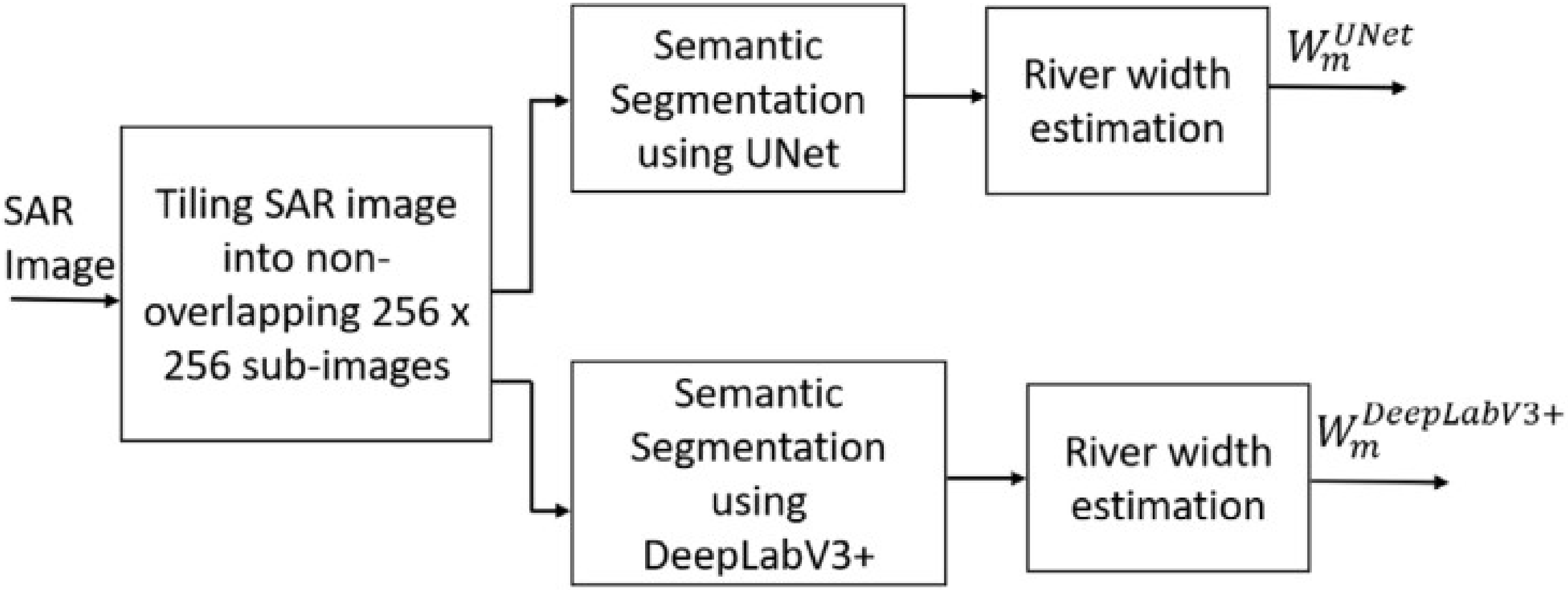
Paper-04

DeepRivWidth: Deep Learning Based Semantic Segmentation Approach for River Identification and Width Measurement (2021)

DATA PREPARATION

- The SAR images were acquired for ~3.5 years (Apr 2017 - Dec 2020)
    - 45 images (one image per month) with a resolution of 1967x3004 pixels were obtained.
    - Each pixel in the SAR images was then hand-labeled manually into two categories — rivers and non-rivers.
- A pair of images (SAR image and labeled image) was padded with zeros at the right and bottom ends to get an image of dimension 2048x3072.
    - Uniform crops of 256x256 were then taken from these images, which resulted in 96 sub-images from a single image.

https://uverma.github.io/assets/pdf/DeepRivWidth-Ujjwal2021.pdf



SAR Image

Tiling SAR image into non-overlapping 256 x 256 sub-images

Semantic Segmentation using UNet

River width estimation

W_m^{UNet}

Semantic Segmentation using DeepLabV3+

River width estimation

$W_m^{DeepLabV3+}$

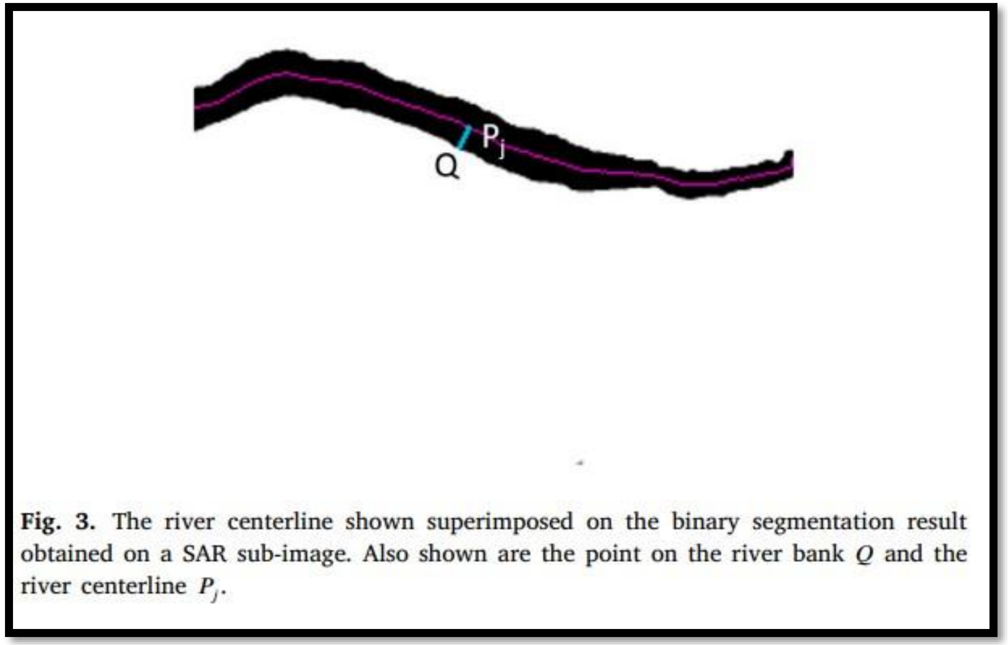


Fig. 3. The river centerline shown superimposed on the binary segmentation result obtained on a SAR sub-image. Also shown are the point on the river bank Q and the river centerline P_j.

Application and Results

- The river width is measured as the distance between two points on the riverbank along the direction orthogonal to the localized centerline of the river. Two-step computation:
  - Computation of river centerline (medial axis or topological skeleton) and
  - Measuring the distance between two points on the riverbank along the orthogonal direction to the river centerline.
- U-Net gave accurate results compared to DeepLabV3+ for river width estimation
  - Ground truth Is manually determined river width
  - Rivers of Mangalore-Udupi region affected by frequent floods, and is eco-sensitive

https://github.com/ArjunChauhan0910/DeepRivWidth

Paper-05

Dynamic Snake Convolution based on Topological Geometric Constraints for Tubular Structure Segmentation (2023)

- Use the specificity of tubular structures to guide DSCNet to simultaneously enhance perception in three stages: feature extraction, feature fusion, and loss constraint.
  - A dynamic snake convolution to accurately capture the features of tubular structures by adaptively focusing on slender and tortuous local structures.
  - Subsequently, a multi-view feature fusion strategy to complement the attention to features from multiple perspectives during feature fusion, ensuring the retention of important information from different global morphologies.
  - Finally, a continuity constraint loss function, based on persistent homology, is proposed to constrain the topological continuity of the segmentation better.

https://openaccess.thecvf.com/content/ICCV2023/papers/Qi_Dynamic_Snake_Convolution_Based_on_Topological_Geometric_Constraints_for_Tubular_ICCV_2023_paper.pdf

Challenges

- Thin and fragile local structure.
  - Thin structures account for only a small proportion of the overall image with limited pixel composition.
  - Moreover, these structures are susceptible to interference from complex backgrounds, rendering it difficult to precisely discriminate subtle target variations by the model.
  - Consequently, the model may struggle to differentiate these structures, resulting in the fracture of the segmentation.
- Complex and variable global morphology.
  - Figure 1 shows the complex and variable morphology of thin tubular structures, even within the same image.
  - Morphological variations are observed in targets located in different regions, depending on the number of branches, the location of bifurcations, and the path length.
  - The model may tend to overfit features that have already been seen, resulting in weak generalization when the data exhibits unprecedented morphological structures.

https://github.com/YaoleiQi/DSCNet

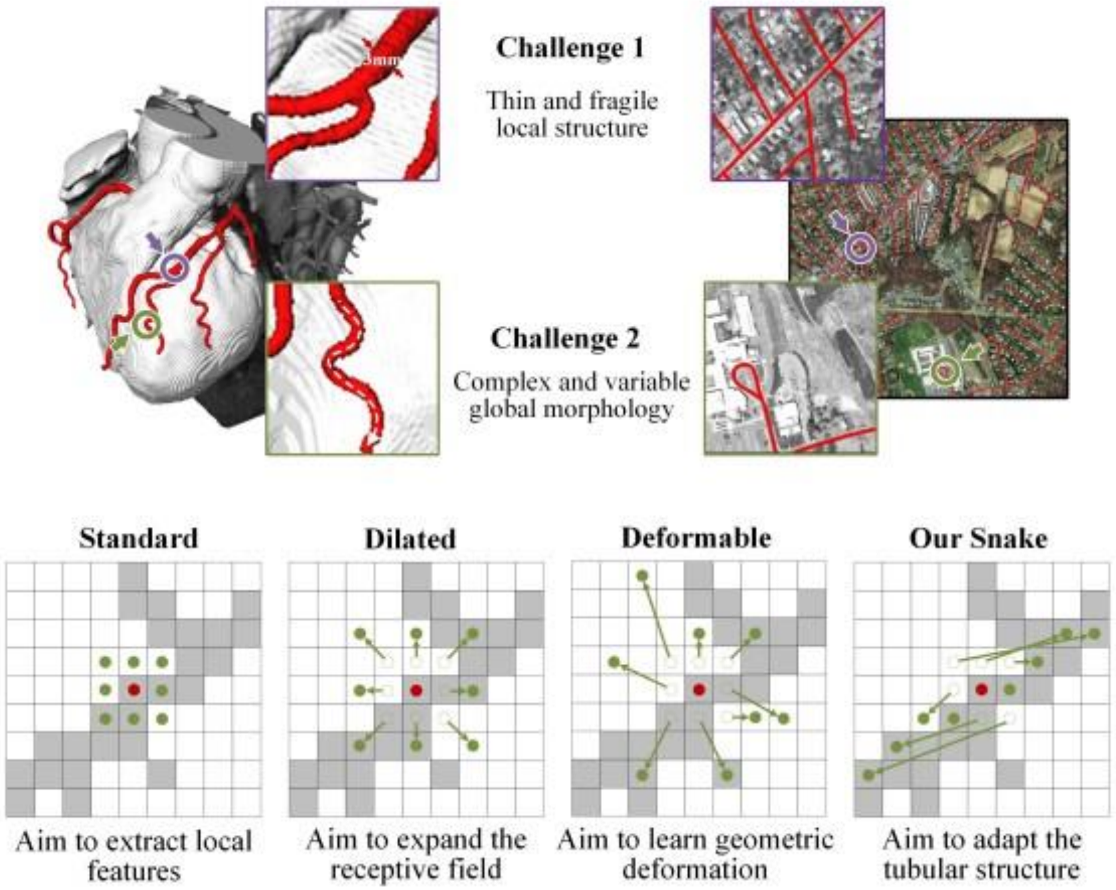

3mm
Challenge 1
Thin and fragile local structure
Challenge 2
Complex and variable global morphology
Standard
Dilated
Deformable
Our Snake
Aim to extract local features
Aim to expand the receptive field
Aim to learn geometric deformation
Aim to adapt the tubular structure

Figure 1. **Challenges.** The above figure shows a 3D heart vascular dataset and a 2D remote road dataset. Both datasets aim to extract tubular structures, but this task faces challenges due to fragile local structures and complex global morphology. **Motivation.** The standard convolutional kernel is intended to extract local features. On this basis, deformable convolutional kernels have been designed to enrich their application and adapt to geometric deformations of different targets. However, due to the aforementioned challenges, it is difficult to focus efficiently on the thin tubular structures.

Salient Points of DSCNet

- Variety of applications/datasets
  - In 2D, the DRIVE retina dataset and the Massachusetts Roads dataset.
  - In 3D, a dataset called Cardiac CCTA Data.
- Multiple validation metrics
  - Volumetric scores: Mean Dice Coefficient (Dice), Relative-Dice coefficient (RDice), CenterlineDice (clDice), Accuracy (ACC) and AUC are used to evaluate the performance of the results
  - Topology errors: Calculate the topology-based scores including the Betti Errors for Betti numbers β0 and β1. To objectively verify the continuity of the coronary artery segmentation, the overlap until first error (OF) is used to evaluate the completeness of the extracted centerline.
  - Distance errors: Hausdorff Distance (HD) is also widely used to describe the similarity between two sets of points, which is recommended to evaluate the thin tubular structures

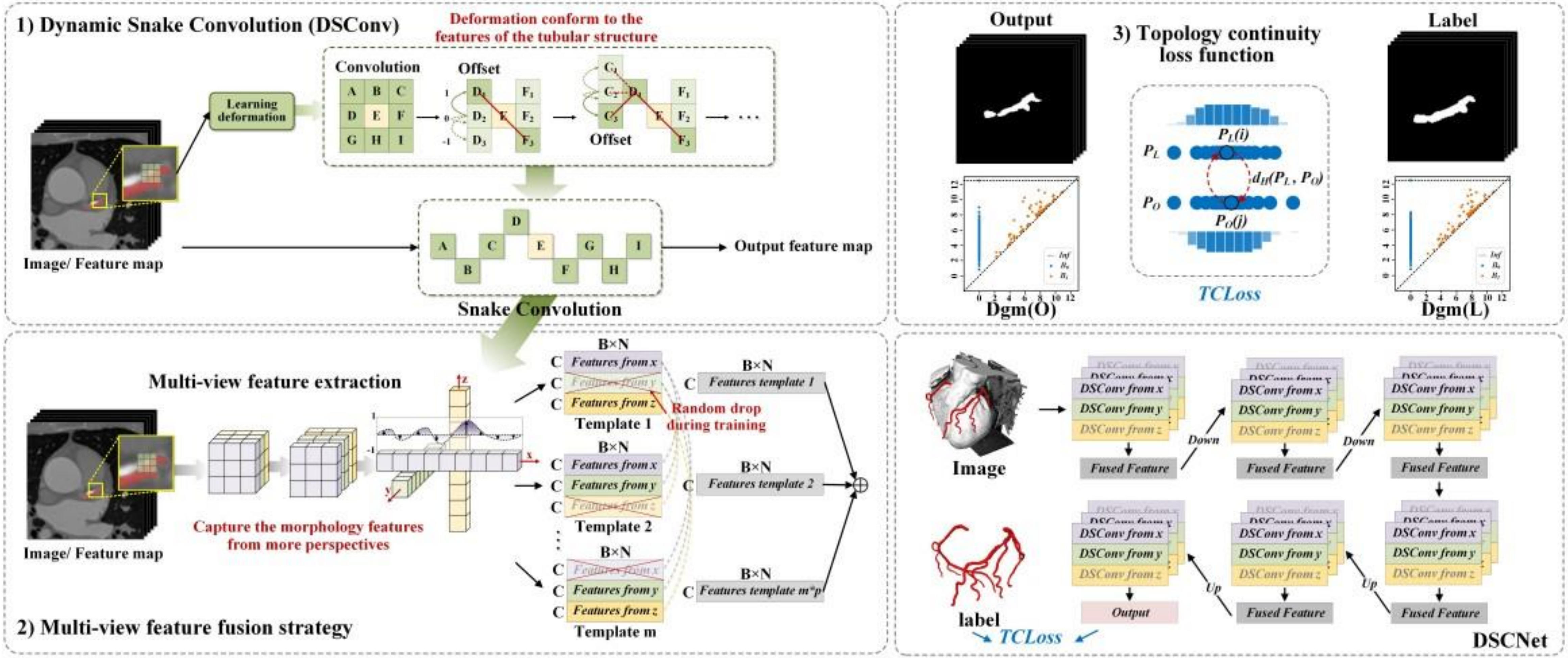
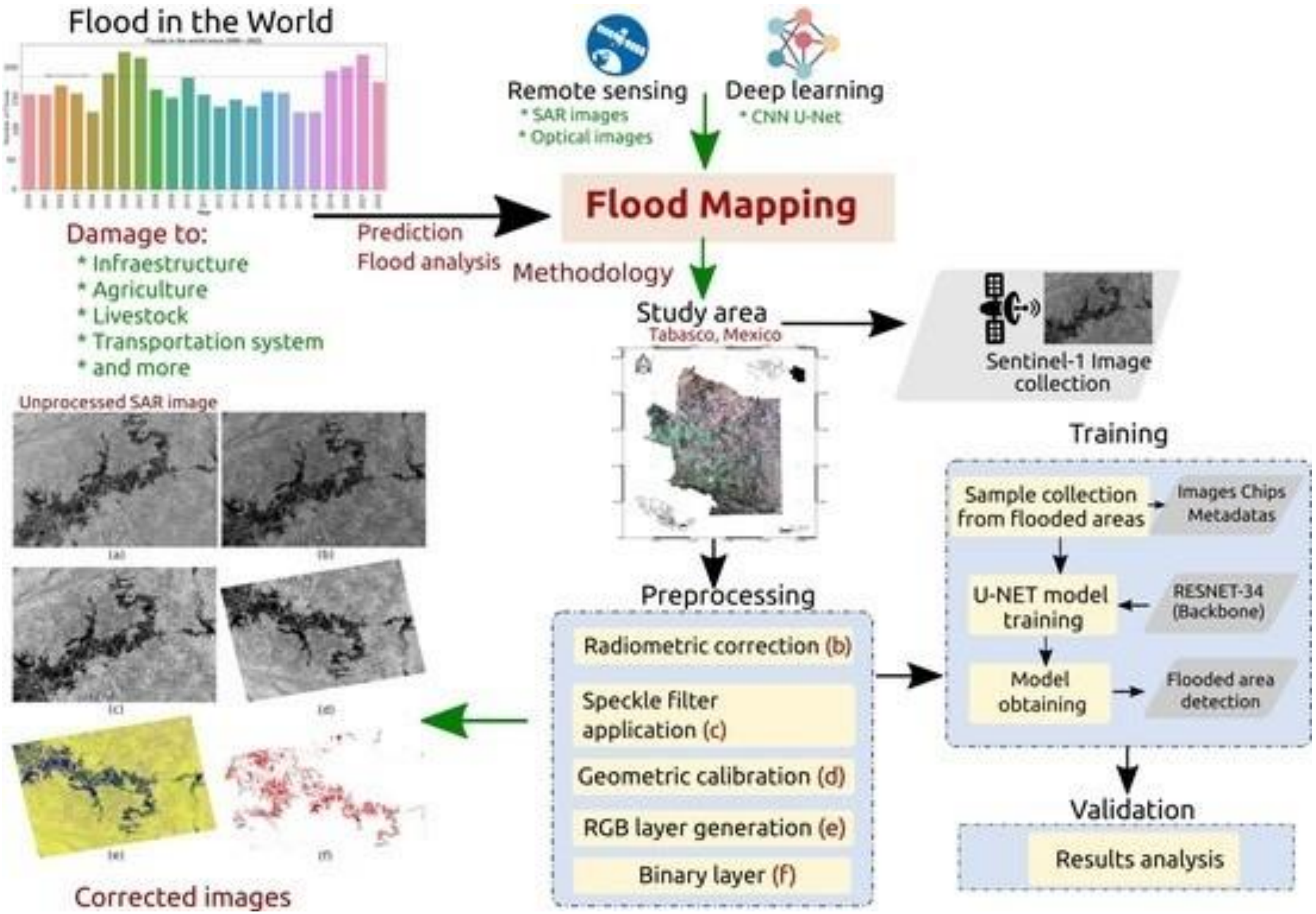


Figure 2. **Methodology.** Schematic overview of our proposed method illustrated on an example of the 3D coronary artery segmentation. Our method has three sections: (1) Dynamic snake convolution (DSConv), which learns the deformation according to the input feature map, adaptively focuses on the slender and tortuous local features under the knowledge of the tubular structure morphology. (2) Multi-view feature fusion strategy, which generates multiple morphological kernel templates based on our DSConv and is used to observe the structural characteristics of the target from multiple perspectives. (3) Loss function, called topological continuity constraint loss function (TCLoss), is based on Persistent Homology to guide the network to focus on the fracture regions with abnormally low pixels/voxels distribution and realize continuity constraint.

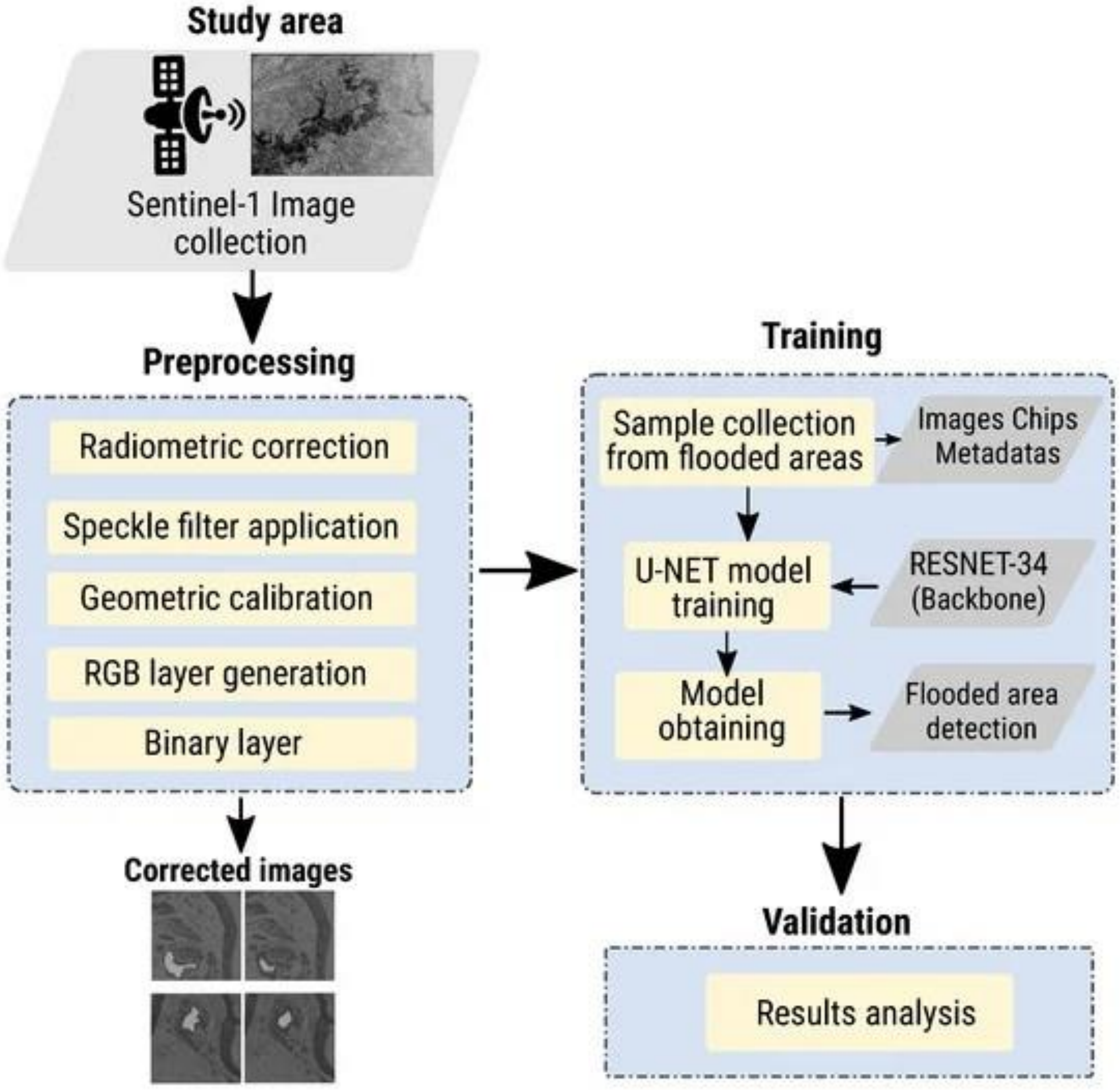
Paper-06

Sentinel-1 SAR Images and Deep Learning for Water Body Mapping (2023)

Workflow for SAR Images



https://www.mdpi.com/2072-4292/15/12/3009



Study area

Sentinel-1 Image collection

Preprocessing

- Radiometric correction
- Speckle filter application
- Geometric calibration
- RGB layer generation
- Binary layer

Corrected images

Training

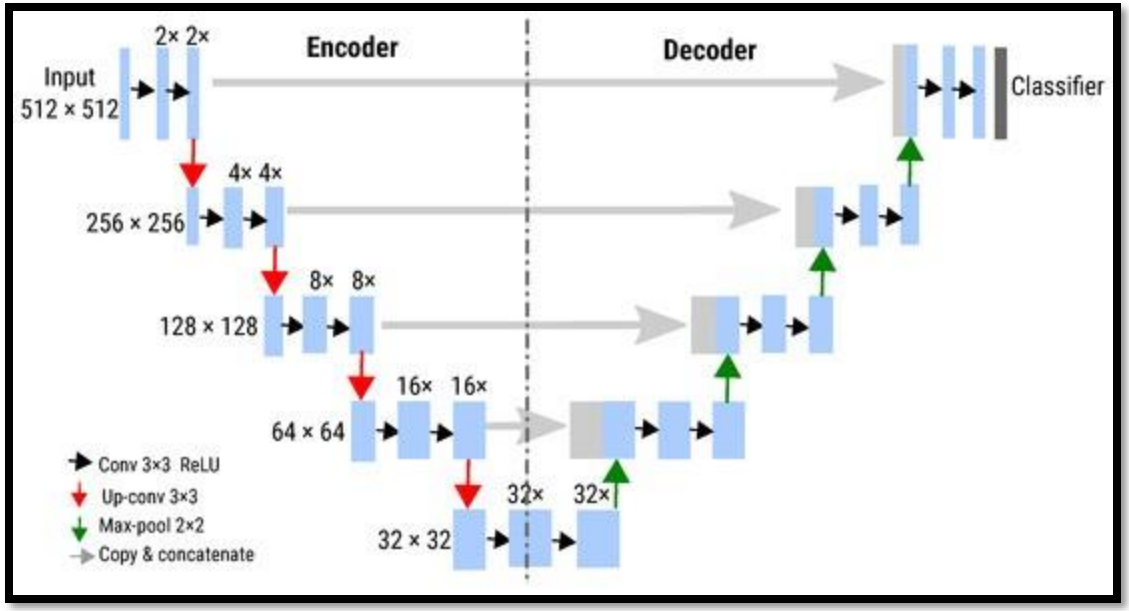
Sample collection from flooded areas → Images Chips Metadatas

U-NET model training ← RESNET-34 (Backbone)

Model obtaining → Flooded area detection

Validation

Results analysis



U-Net structure for SAR image segmentation consists of two paths: encoder and decoder. The encoder is a pre-trained classification network (ResNet) where convolution blocks followed by max-pool downsampling are applied to encode the input image into feature representations at several levels. Each block is a convolution operation and follows a ReLU activation function. The red arrows indicate a 2 × 2 max-pooling layer.

U-Net Results

Training the model with

- 256 samples (chips) and 25 epochs => precision of 82%, recall of 40%, and F1 of 53%.
- 1036 samples and 100 epochs => precision of 94%, recall of 92%, and F1 of 93%.

Paper-07

HA-Unet: A Modified Unet Based on Hybrid Attention for Urban Water Extraction in SAR Images (2022)

- Urban water extraction is still a challenging task in automatic interpretation of SAR images.
- The influence of radar shadows and strong scatters in urban areas may lead to misclassification in urban water extraction.
- Local features extracted by CNNs are generally redundant and cannot use global information for water pixel prediction.
- To emphasize the identifiable water characteristics and also exploit the global information of SAR images, a modified U-Net based on hybrid attention mechanism is proposed.

https://www.mdpi.com/2079-9292/11/22/3787

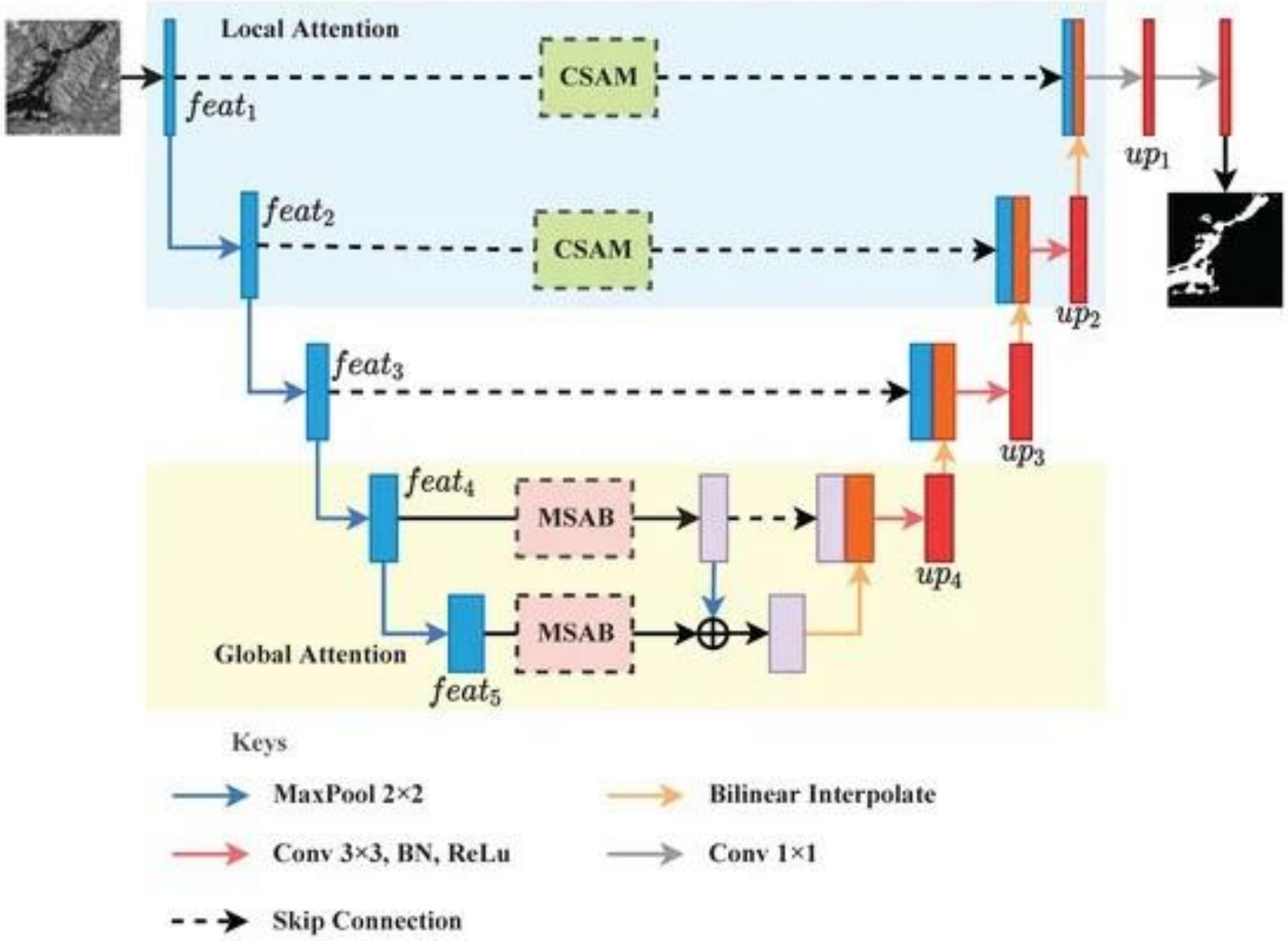
HA-Unet Characteristics

Components for feature extraction ability and the global modeling capability in SAR image segmentation in HA-Unet:

- Channel and Spatial Attention Module (CSAM)
- Multi-head Self-Attention Block (MSAB)

==

- During the feature extraction process, CSAM based on local attention is adopted to enhance the meaningful water features and ignore unnecessary features adaptively in feature maps of two shallow layers.
- In the last two layers of the backbone, MSAB is introduced to capture the global information of SAR images to generate global attention.
- In addition, two global attention maps generated by MSAB are aggregated together to reconstruct the spatial feature relationship of SAR images from high-resolution feature maps.



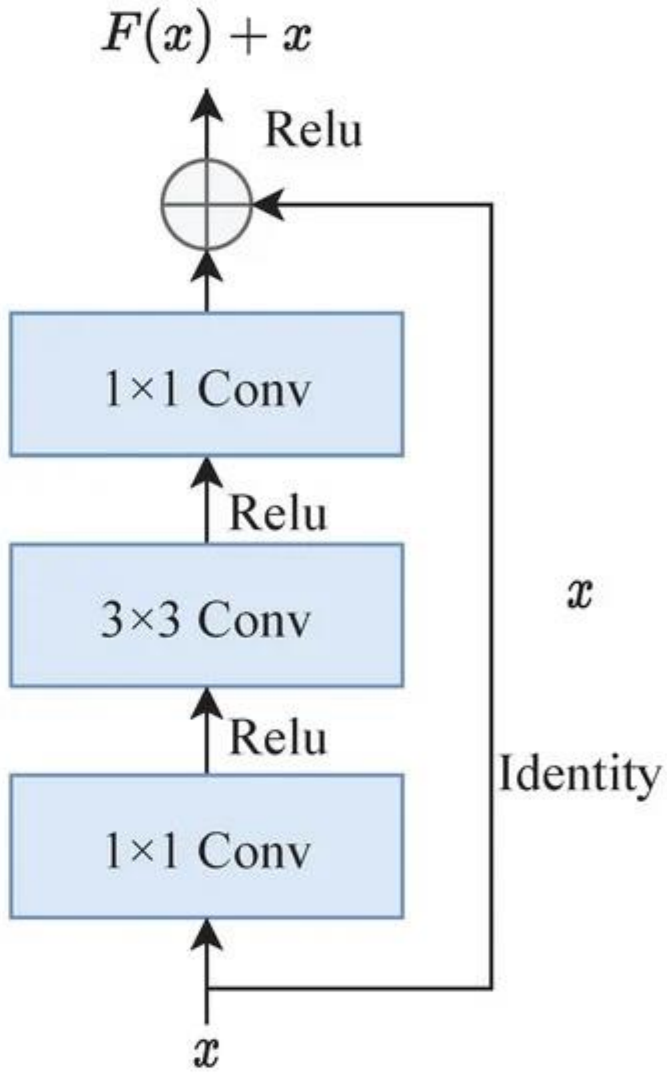
The experimental results on Sentinel-1A SAR images show that the proposed urban water extraction method has a strong ability to extract water bodies in the complex urban areas.

The ablation experiment and visualization results vividly indicate that both CSAM and MSAB contribute significantly to extracting urban water accurately and effectively.

Resnet is widely used in semantic segmentation and target detection, and shows outstanding performance in remote sensing images.

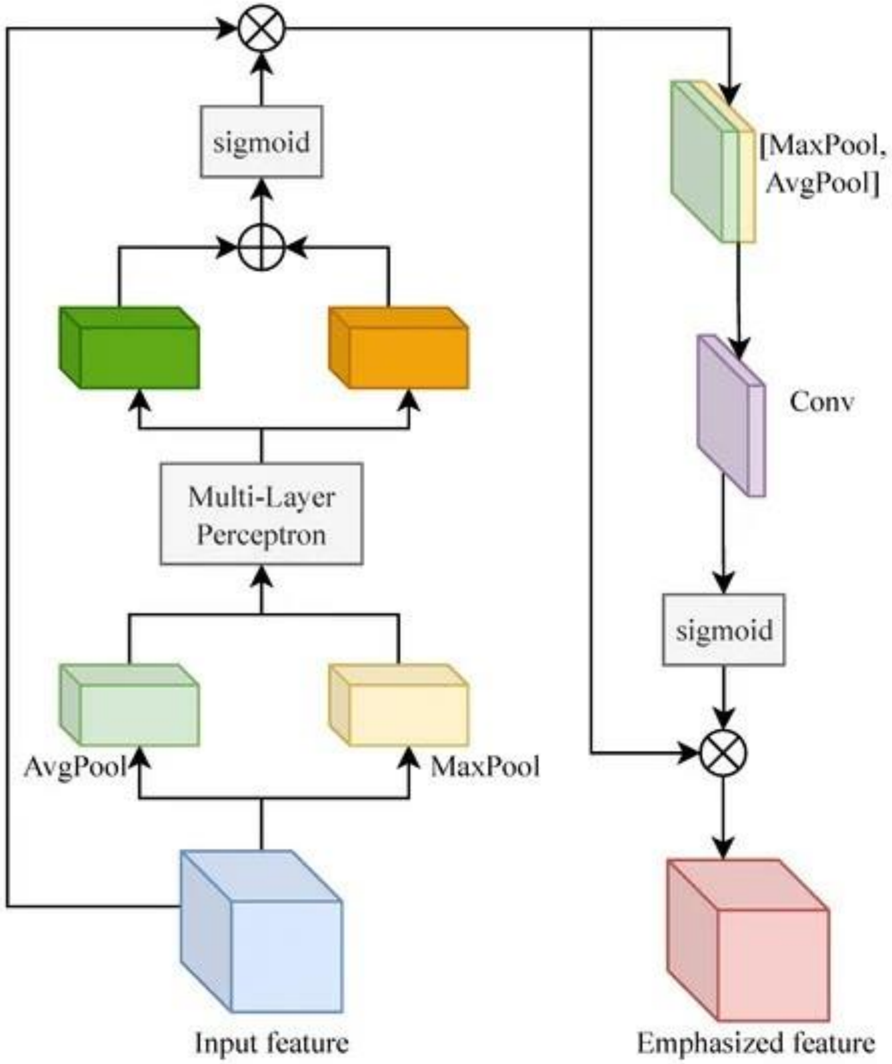
About ResNet

- The key idea of Resnet is the deep residual learning framework
- Instead of directly fitting an underlying mapping $F(x)$, the stacked layers in Resnet fit a residual mapping $F(x)$+x which is performed by a shortcut connection and element-wise addition.
- This framework with a shortcut connection helps Resnet extend the depth of the network to learn richer features without gradient degradation.
- Resnet50 is adopted for water feature extraction in this work based on #parameters and ease of training.



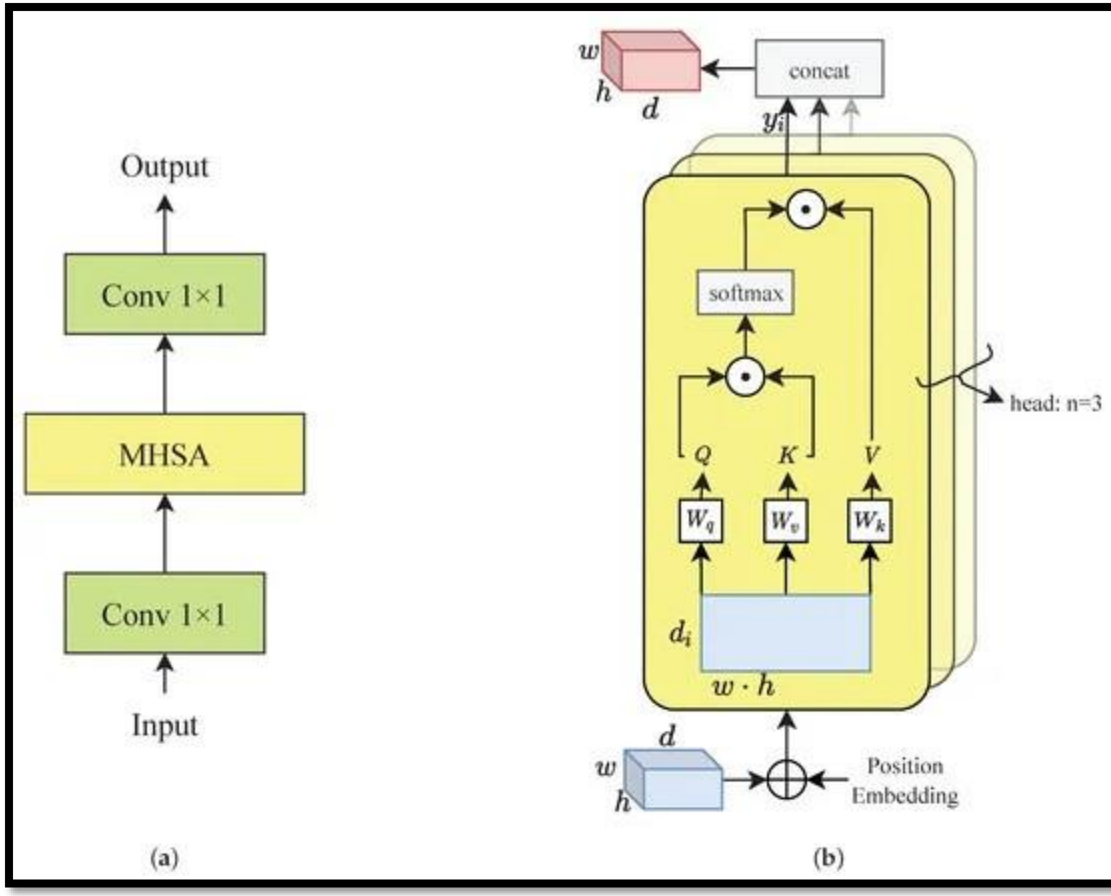
Attention Mechanism: CSAM

- Identifiable and representative feature representations are essential in high accuracy segmentation.
- However, these features extracted by CNNs are redundant, especially at low stage, and they may influence the implicit representation of CNNs.
- Thus, a feature selecting approach is necessary for urban water extraction.
- CSAM based on channel attention and spatial attention is adopted at the early stage of the encoder for adaptive feature enhancement in complex scenes of SAR images.
- Outcome of CSAM: Salient parts of significant properties of water in the feature maps $feat1$ and $feat2$ are focused on water bodies by adaptive enhancement in both channel and spatial dimensions respectively, while the unnecessary ones are suppressed.



Attention Mechanism: MSAB

- Segmentation is a task that requires accurate pixel-level predictions.
- Fine-grained features and long-range dependencies needed to resolve the ambiguities of local pixel prediction.
- In large-scene SAR images, intrinsic correlations among pixels are beneficial to improve classification accuracy, especially for small regional segmentation.
- CNNs have difficulty in capturing the latent contextual correlations of the whole image, since they only process a local neighborhood because of their local receptive field.
- Based on self-attention, MSAB is introduced into the late stages of the encoder to model the long-range dependencies of SAR images.
- In MSAB, the multi-head self-attention (MHSA) layer captures the multiple complex relationships by a concatenation of outputs of n self-attention heads and 1 × 1 convolutions are used to transform the dimensions of output feature maps.



Implementation of MSAB

- The input feature x of MHSA layer is appended with positional embedding.
- With the parallel execution of n heads, the MHSA layer learns the richer non-local context.
- To address the high computational complexity in MHSA, only three MHSA layers with four heads are used to construct MSAB in this work.
- In the late stage of the encoder, global attention maps are extracted from $feat4$ and $feat5$ with MSAB, respectively, and global attention maps of different stages are aggregated together to capture long-range dependencies from high-resolution feature maps of SAR images.

Loss Function & Comparative Study

- Considering the imbalanced categorical distribution in the training set for urban water extraction, the loss function based on cross-entropy loss and dice loss is introduced, referred to as L=Lce+Ld.
- The urban water extraction results generated by DeeplabV3+, original U-Net, and the proposed HA-Unet.
  - HA-Unet outperforms the other two models.
- HA-Unet with the hybrid attention has fewer omission errors and commission errors even in complex scenes.
- In comparing the two attention modules, either of the two modules can improve the performance of original Unet gradually in urban water extraction.
- HA-Unet can better understand the characteristics of water boundaries, locations, and shapes owing to the identifiable features from CSAM.
- Limitation: MSAB is resource intensive and needs several parameters.

Paper-08

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 (2021)

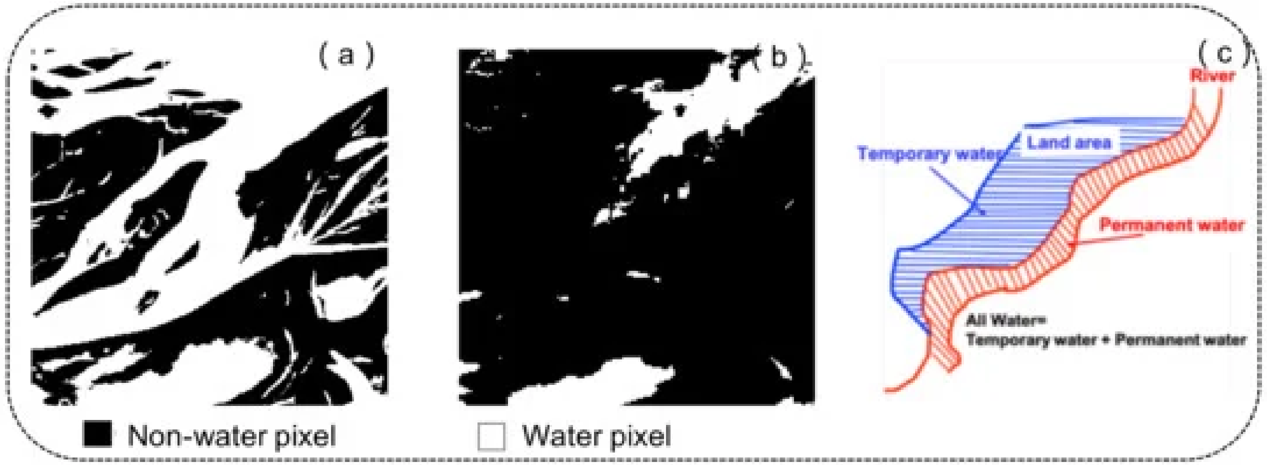
- Using change detection method from multi-temporal remote sensing imageries identifies temporary/permanent water bodies but estimating the water type in flood disaster events from only post-flood remote sensing images is challenging.
- New deep learning algorithms and a multi-source data fusion driven flood inundation mapping approach by leveraging a large-scale publicly available Sen1Flood11 dataset consisting of roughly 4831 labelled Sentinel-1 SAR and Sentinel-2 optical imagery gathered from flood events worldwide in recent years.
- An automatic segmentation method for surface water, permanent water, and temporary water identification, and all tasks share the same CNN architecture.

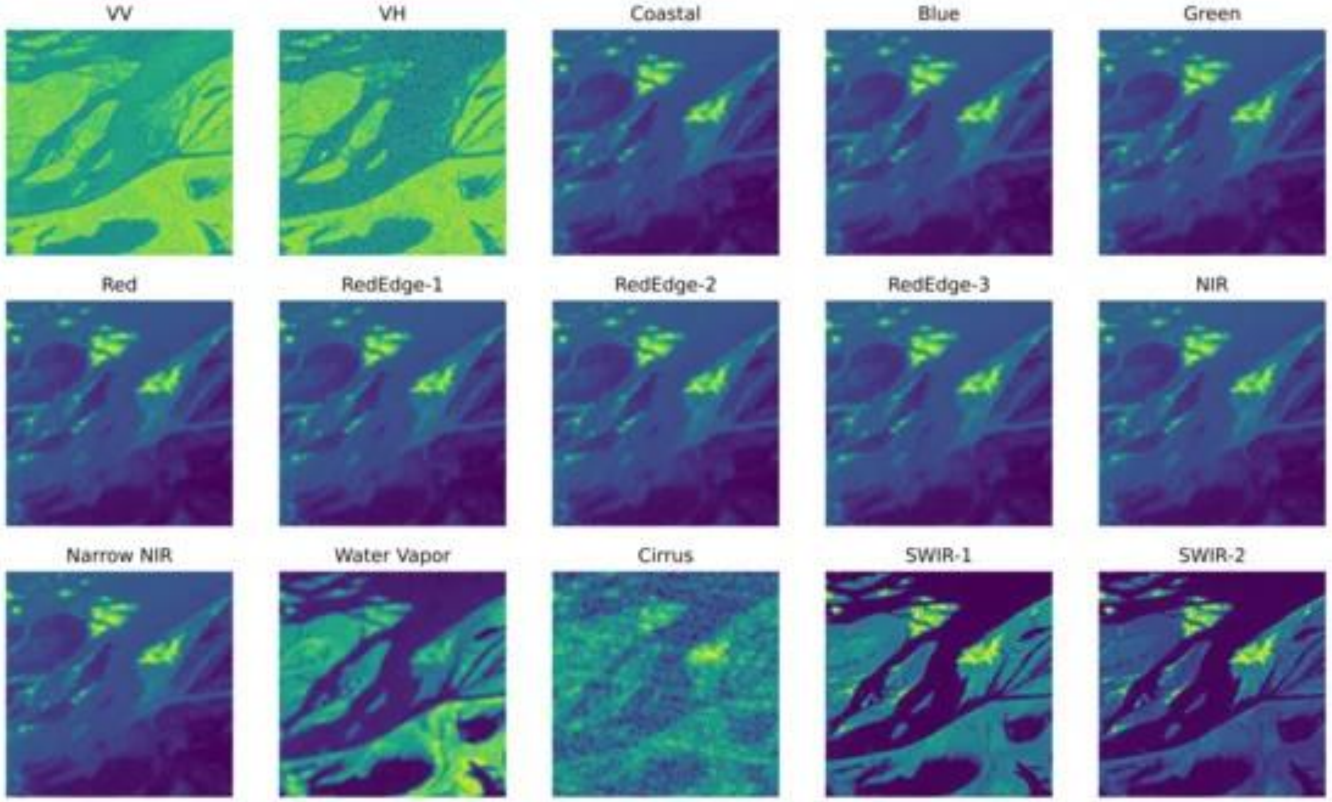
https://www.mdpi.com/2072-4292/13/11/2220

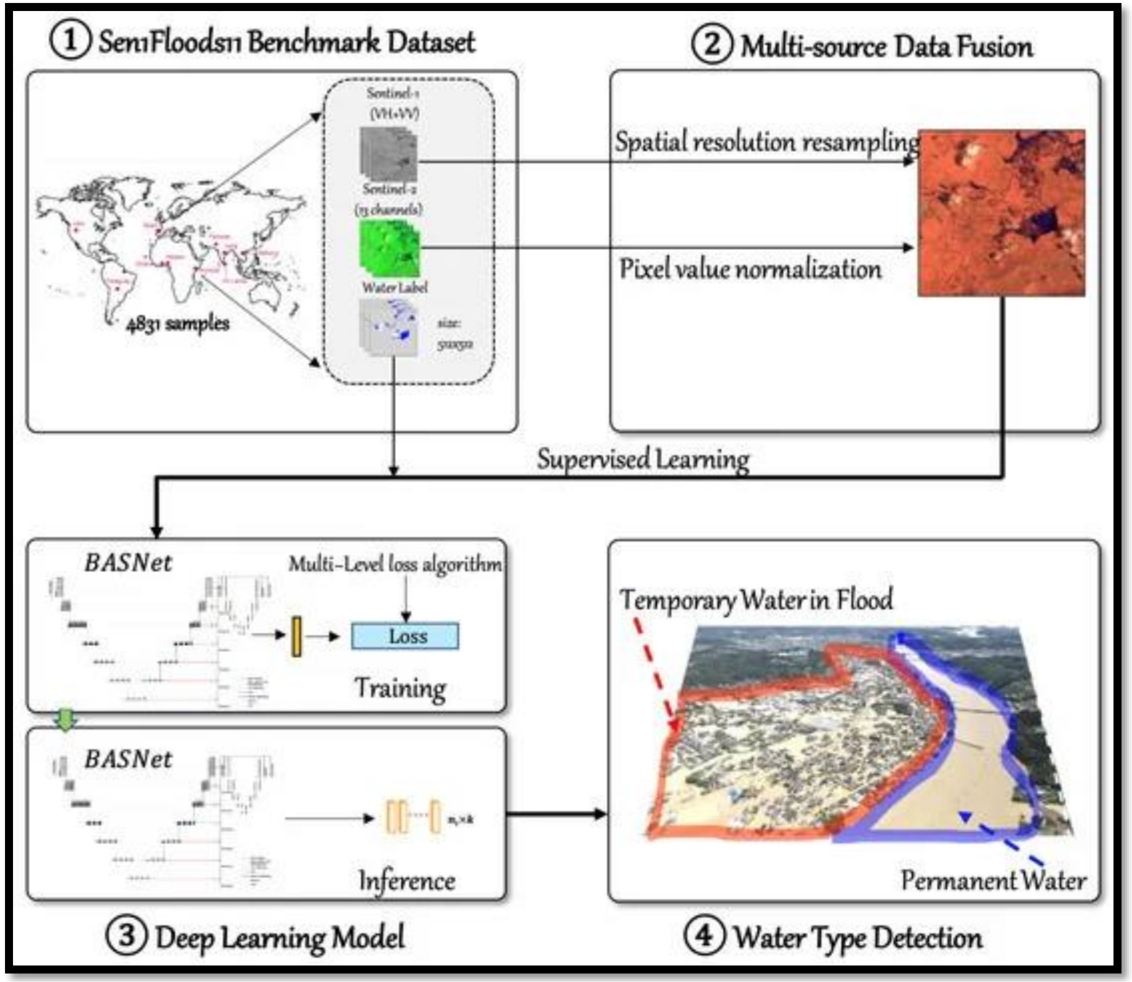
Salient Parts

- Utilize focal loss to deal with the class (water/non-water) imbalance problem.
- Due to the high cost of hand labels, 4370 tiles are not hand-labeled and exported with annotations automatically generated by the Sentinel-1 and Sentinel-2 flood classification algorithms, which can serve as weakly supervised training data. The remaining 446 tiles are manually annotated by trained remote sensing analysts for high-quality model training, validation and testing. The weakly supervised data contain two types of surface water labels. One is produced by the histogram thresholding method based on the Sentinel-1 image; the other is generated by the Normalized Difference Vegetation Index (NDVI), MNDWI and thresholding method based on the Sentinel-2 image.
- Hand labels include all water labels and permanent water labels. For all water labels, analysts exploited Google Earth Engine to correct the automated labels using Sentinel-1 VH band, two false color composites from Sentinel-2 and the reference water classification from Sentinel-2 by removing uncertain areas and adding to the water classification.
- The Sen1Floods11 dataset shows the highly imbalanced distribution between flooded and unflooded area. water pixels account for only 9.16%, and non-water pixels account for 77.22%, which is about eight times the number of surface water pixels. The percentages of water pixels and non-water pixels in permanent waters are 3.06% and 96.94%, respectively, and the number of non-water pixels is about 32 times that of non-water pixels.

https://github.com/cloudtostreet/Sen1Floods11


(a)
(b)
(c)
River
Temporary water
Land area
Permanent water
All Water=
Temporary water + Permanent water
■ Non-water pixel
□ Water pixel


VV
VH
Coastal
Blue
Green
Red
RedEdge-1
RedEdge-2
RedEdge-3
NIR
Narrow NIR
Water Vapor
Cirrus
SWIR-1
SWIR-2


① Sen1Floods11 Benchmark Dataset
② Multi-source Data Fusion
4831 samples
Sentinel-1 (VH+VV)
Sentinel-2 (13 channels)
Water Label
size: 512x512
Spatial resolution resampling
Pixel value normalization
Supervised Learning
BASNet
Multi-Level loss algorithm
Loss
Training
BASNet
Inference
Temporary Water in Flood
Permanent Water
③ Deep Learning Model
④ Water Type Detection

Paper-09

S1S2-Water: A Global Dataset for Semantic Segmentation of Water Bodies From Sentinel- 1 and Sentinel-2 Satellite Images (2024)

- A global reference dataset for training, validation, and testing of convolutional neural networks (CNNs) for semantic segmentation of surface water bodies in publicly available Sentinel-1 and Sentinel-2 satellite images.
- The dataset consists of 65 triplets of S1 and S2 images with quality-checked binary water mask. Samples are drawn globally on the basis of the S2 tile-grid (100 km × 100 km) under consideration of predominant landcover and availability of water bodies.
- Each sample is complemented with metadata and digital elevation model (DEM) raster from the Copernicus DEM.
- Use CNN architectures to segment surface water bodies from S1 and S2 images, and evaluate the influence of image bands, elevation features (slope) and data augmentation on the segmentation performance and identify best-performing baseline-models.
- The model for Sentinel-1 achieves an Intersection over Union (IoU) of 0.845, Precision of 0.932, and Recall of 0.896 on the test data. For Sentinel-2 the best model produces an IoU of 0.965, Precision of 0.989, and Recall of 0.951, respectively.
- This work also evaluates the performance impact when a model is trained on permanent water data and applied to independent test scenes of floods.

https://ieeexplore.ieee.org/document/10321672

Independent Flood Water Dataset

- The S1S2-Water dataset covers normal water bodies and does not specifically consider anomalously flooded areas.
- To evaluate the performance impact when a model is trained on normal water data (S1S2-Water) and applied to floods, an independent reference dataset S1S2-Flood is developed that covers 12 major flood events across the globe.
- Similar to S1S2-Water, each sample of this S1S2-Flood dataset consists of S1 GRD and S2 L1C images with associated quality-controlled binary water mask annotations, elevation and slope layers as well as metadata.
- Maximum time difference between acquisition dates of S1 and S2 images has been limited to one day to ensure spatial and temporal consistency of the flood masks. The same procedure as for S1S2-Water samples is used to annotate the satellite images.

Salient Parts

- U-Net architecture has been proven to deliver highly accurate results for water segmentation tasks in high-resolution satellite images at relatively low computational complexity.
- Compared with different encoders, namely MobileNet-V3, ResNet-50, EfficientNet-B0 and EfficientNetB4, which are selected since they show a good trade-off between number of model parameters and Imagenet Top-1 accuracy
- For Sentinel-2, only spectral bands that are available across different satellite sensors are used (e.g., Landsat OLI) to ensure a high degree of transferability of the trained models.
- Water shows low reflectance in the NIR and SWIR wavelengths as it absorbs more energy, while non-water generally has a higher reflectance. This leads to a high contrast in reflectance values between water and non-water landcover classes in the NIR and SWIR spectral bands compared to the visible R, G and B bands.
- Satellite images are affected by changes in landcover, atmospheric conditions, seasonality and other scene and image properties such as sun elevation or radiometric resolution. Due to this very large variability of influencing factor, even large reference datasets may not cover all possibilities that may occur in real-world applications. To this regard, data augmentation enables a network to learn invariance to changes in the augmented domains to a degree that may go beyond what is present in the raw training image
    - Data augmentation with random contrast, brightness, scale and image flipping

Regarding Polarization in S1 & Bands in S2

- An improvement of 0.10 IoU compared to using VV polarization alone and 0.06 IoU compared to using VH polarization alone for S1 SAR images.
- VH polarization seems to have a larger positive impact on the test scores of the water segmentation than VV polarization. This is contradicting with several studies that focus solely on VV polarization for water segmentation
- These studies show that by using solely VV polarization land and water can be distinguished very well.
- While this is true in particular for smaller water bodies, VV polarization is sensitive to wind-induced roughening effects and hence prone to cause false-negatives over larger open water bodies.
- VH polarization on the contrary is known to be less sensitive to roughening effects and can aid in reducing false-negatives in such situations.
- Therefore, a combination of both polarizations works best in practice.
- For S2, combining the NIR spectral band with the R-G-B bands provided an improvement of 0.09 IoU compared to using R-G-B bands alone