Notes on River Segmentation from SAR Images
Jaya Sreevalsan Nair
Mar 1, 2024

A Three-Stage Deep Learning Model for Accurate Retinal Vessel Segmentation (2019, IEEE JBMI)
Yan et al.

Vessels < 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.

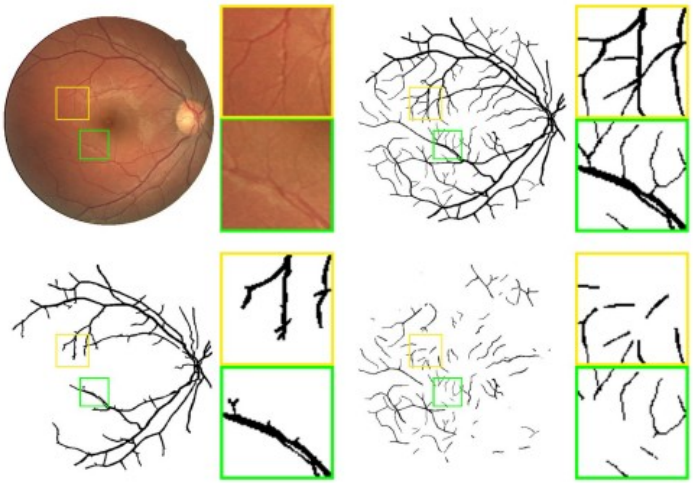


Fig. 1. Analysis of the retinal vessel segmentation problem. Row 1: from left to right, the fundus image and the enlarged patches, and the manual annotation and the annotations of the two fundus image patches. Row 2: from left to right, the manually annotated thick vessels and the annotated thick vessels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

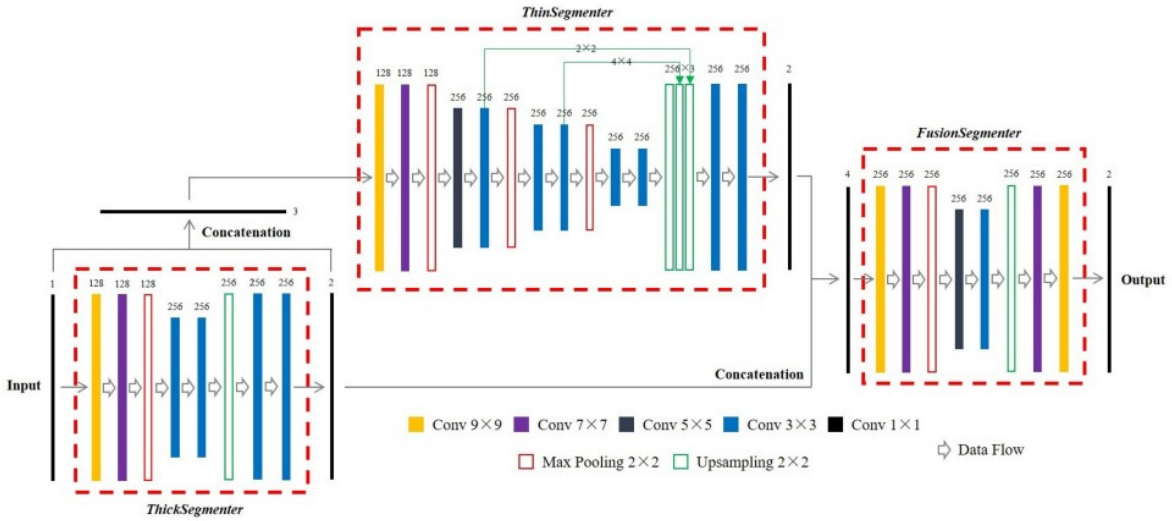


Fig. 2. The overview of the proposed three-stage deep learning framework. The framework consists of three separate models, namely *ThickSegmenter* for thick vessel segmentation, *ThinSegmenter* for thin vessel segmentation and *FusionSegmenter* for vessel fusion respectively.

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

Pai et al. - Ujjwal Verma,

In this paper, a robust methodology is proposed for an efficient and highly precise segmentation of surface river water and land. In addition, two different implementation of U-Net architecture is studied on SAR images, one in which U-net is trained from scracth (Vanilla U-Net) and other in which 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. However, transfer U-Net is able to identify very minute details in the image such as small rivers etc.

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

Pai et al. - Ujjwal Verma

Further improve the U-Net methodology by augmenting the dataset of manually annotated images through the use of Generative Adversarial Networks (GANs) to generate SAR images and corresponding masks.

The raw SAR images along with their corresponding (ground truth) masks are fed into the GAN where the generator has been given the latent space and the discriminator is fed real images and generated images. Once this GAN reaches an acceptable level of performance, we will obtain new SAR image patches along with their corresponding masks. These images, combined with the initial dataset are fed into the U-Net for further analysis.

Deep Convolutional GANs (DC-GANs) are used to create more patches of SAR images. The ground truth masks and raw SAR images are fed into the GAN (Fig. 2). The generator maps the underlying distribution of the SAR data as well as the corresponding masks to generate more images. To feed the GAN in the ¯rst place, heavy augmentation is required which might lead to a skewed output in which case the generated images will look like the training dataset to a higher degree than what is required. The DC-GAN is trained with batch size of 128, with batch normalization to avoid any large spikes within the batch. The Adam optimizer is used and a sparse softmax cross-entropy with logits loss function is used in the implementation with an initial learning rate of 0.0002.

. This methodology is augmented using GANs to create addition training data. The results show improvement in the accuracy across various parameters. Therefore, the large manual overhead of labeling SAR images for ground truth is removed because the GAN creates images that can be used instead. This opens the door for created larger datasets.

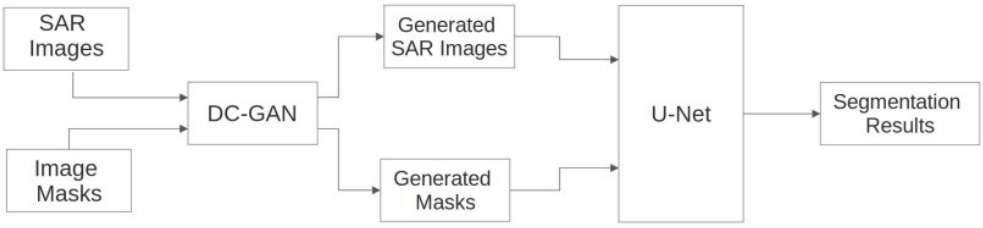
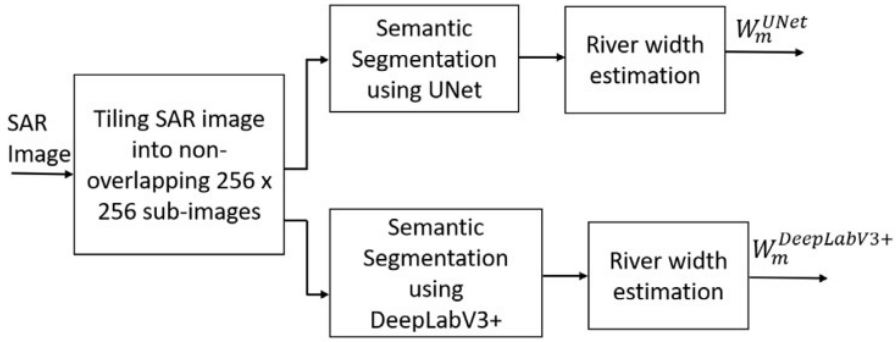


Fig. 2. Using DC-GANs to augment segmentation process of U-Net.

DeepRivWidth: Deep Learning Based Semantic Segmentation Approach for River Identification and Width Measurement (Computers and Geosciences, 2021) – U. Verma et al.



The SAR images were acquired for approximately three and half years (April 2017 to December 2020). In total, 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.

Following this, the 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.

consisting of the river pixels in the foreground class. 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. This is achieved in two steps: (1) computation of river centerline and (2) measuring the distance between two points on the riverbank along the orthogonal direction to the river centerline. First, the medial axis (topological skeleton) of the foreground object is computed. The centerline of the river is then assumed to be represented by this skeleton

The speckle noise is an

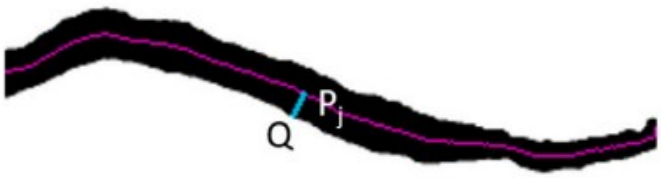


Fig. 3. The river centerline shown superimposed on the binary segmentation result obtained on a SAR sub-image. Also shown are the point on the river bank Q and the river centerline P_j.

undesirable effect of SAR image and there are several approaches for reducing the effect of speckle noise. In our study, it was found that the use of Lee filter for reducing the effect of speckle noise results in loss of image information

The average absolute error in measurement was 20.05 m using the segmentation results of U-Net and 31.3 m using the segmentation results of DeepLabV3+. Besides, the root mean square error (RMSE) of 24.9 m using U-Net segmentation map and 76.6 m using DeepLabV3+ segmentation map was obtained. The estimated river width using the U-Net segmentation results is more accurate than that of DeepLabV3+, mainly due to a more precise segmentation map obtained using U-Net

The source code used in this study is available at https://github. com/ArjunChauhan0910/DeepRivWidth.

. The proposed method is validated by measuring the width of the rivers of the Mangalore–Udupi region of Coastal Karnataka (India), which is affected by frequent floods during monsoon and water scarcity in summer. The measured river width from SAR images would provide government authorities information for better planning and management of water resources of this ecologically sensitive region.

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

Accurate segmentation of topological tubular structures, such as blood vessels and roads, is crucial in various fields, ensuring accuracy and efficiency in downstream tasks. However, many factors complicate the task, including thin local structures and variable global morphologies. In this work, we note the specificity of tubular structures and use this knowledge to guide our DSCNet to simultaneously enhance perception in three stages: feature extraction, feature fusion, and loss constraint. First, we propose a dynamic snake convolution to accurately capture the features of tubular structures by adaptively focusing on slender and tortuous local structures. Subsequently, we propose 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. Experiments on 2D and 3D datasets show that our DSCNet provides better accuracy and continuity on the tubular structure segmentation task compared with several methods. Our codes are publicly available

https://github.com/YaoleiQi/DSCNet

However, it remains challenging due to the following difficulties:

(1) **Thin and fragile local structure**. As shown in Figure 1, 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.

(2) **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.

We employ three datasets containing two public and one internal dataset for validating our framework. In 2D, we evaluate the DRIVE retina dataset[25] and the Massachusetts Roads dataset[17]. In 3D, we used a dataset called Cardiac CCTA Data. Details concerning the experimental setup can be found in the supplementary material.

. On the Massachusetts Roads dataset, our DSCNet also achieves the best results. As shown in Table 1, our proposed DSCNet with TCLoss achieves the best segmentation results compared with other methods with Dice of 78.21%, RDice of 85.85%, and clDice of 87.64%. Compared with the results of the classical segmentation network UNet, our method achieves at most 1.31% Dice, 1.78% RDice, and 0.77% clDice improvements. The results show that our model also performs well for structurally complex and morphologically variable road datasets compared with other models.

All metrics were calculated for each image and averaged.

1. Volumetric scores: Mean Dice Coefficient (Dice), Relative-Dice coefficient (RDice)[22], CenterlineDice (clDice)[24], Accuracy (ACC) and AUC are used to evaluate the performance of the results

2. Topology errors: We follow [24, 28] and calculate the topology-based scores including the Betti Errors for Betti numbers β0 and β1. Meanwhile, to objectively verify the continuity of the coronary artery segmentation, the overlap until first error (OF) [23] is used to evaluate the completeness of the extracted centerline.

3. Distance errors: Hausdorff Distance (HD) [26] is also widely used to describe the similarity between two sets of points, which is recommended to evaluate the thin tubular structures

[4] Ozg ¨ un C¸ ic¸ek, Ahmed Abdulkadir, Soeren S. Lienkamp, et al. ¨ 3d u-net: Learning dense volumetric segmentation from sparse annotation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 424–432. Springer, 2016.

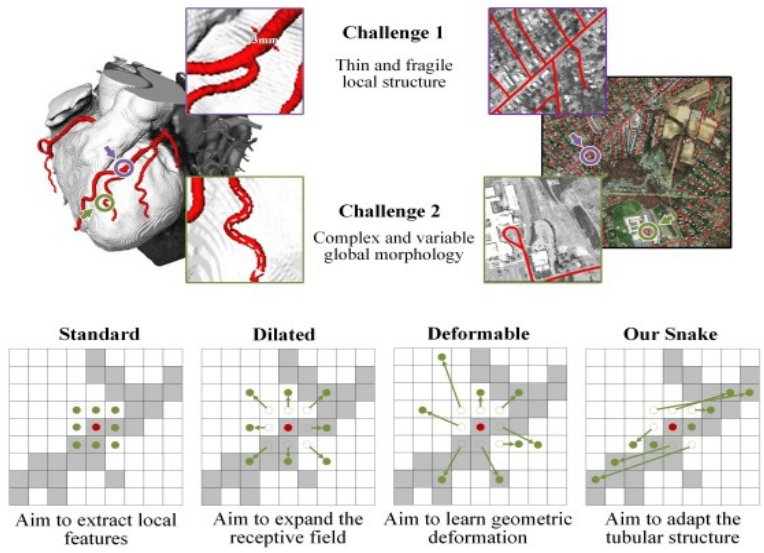


Figure 1. **Challenges.** The above figure shows a 3D heart vascular dataset and a 2D remote road dataset. Both datasets aim to extract tubular structures, but this task faces challenges due to fragile local structures and complex global morphology. **Motivation.** The standard convolutional kernel is 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.

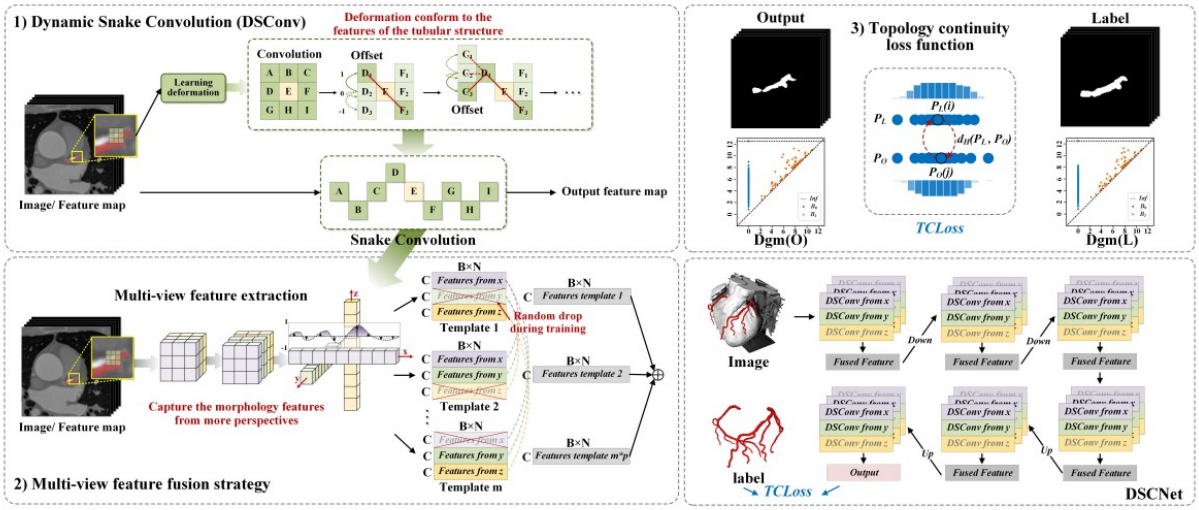
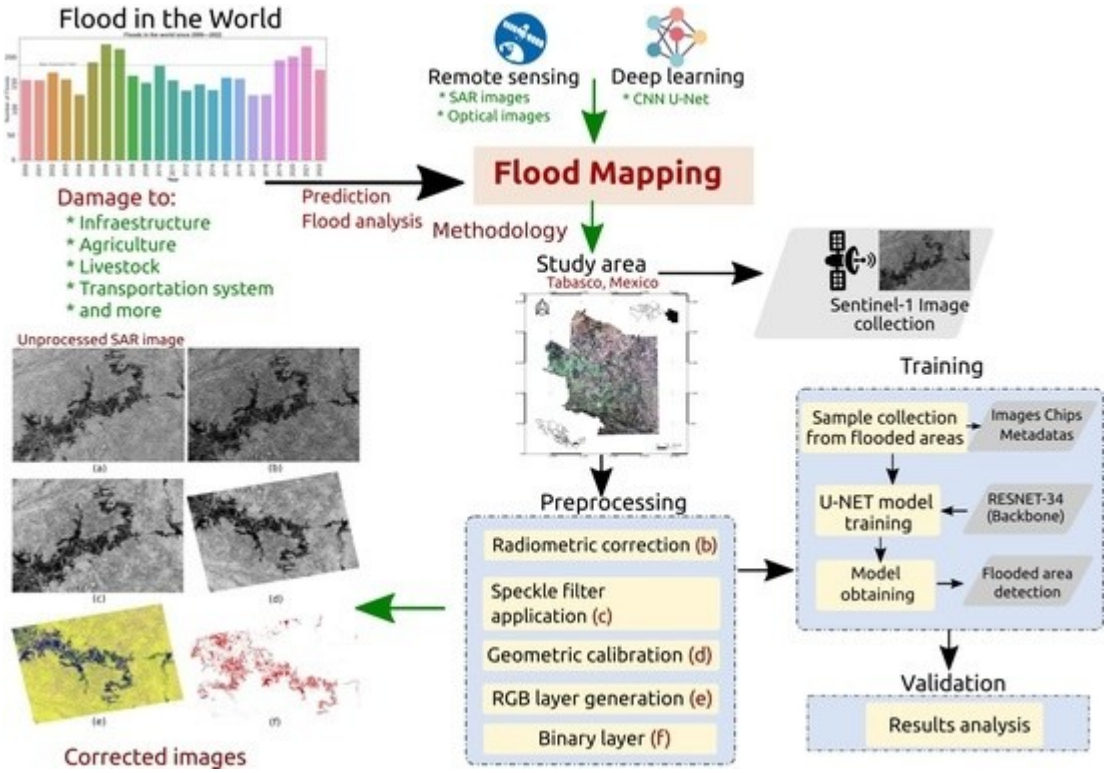


Figure 2. **Methodology.** Schematic overview of our proposed method illustrated on an example of the 3D coronary artery segmentation. Our method has three sections: (1) Dynamic snake convolution (DSConv), which learns the deformation according to the input feature map, adaptively focuses on the slender 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.

Sentinel-1 SAR Images and Deep Learning for Water Body Mapping (Remote Sensing MDPI, 2023)

Pech-May et al.



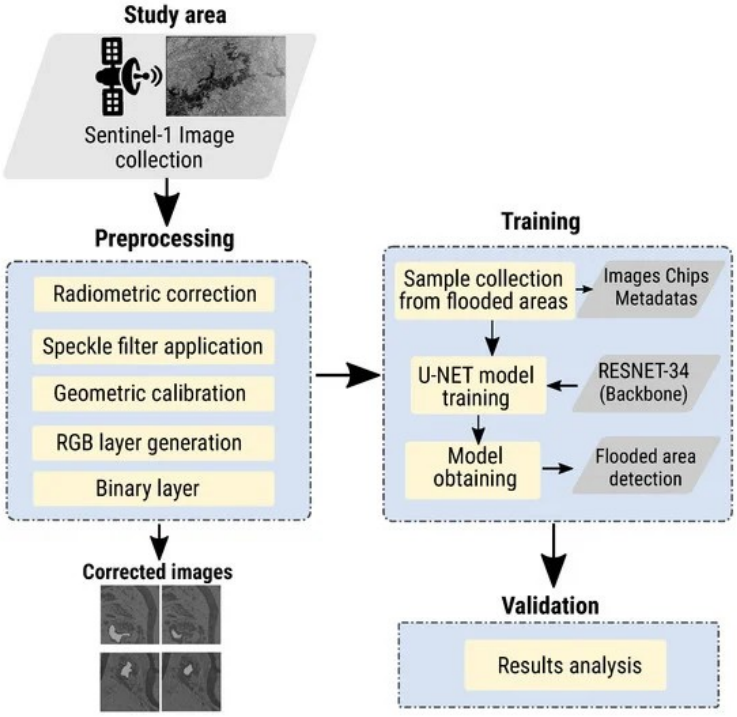
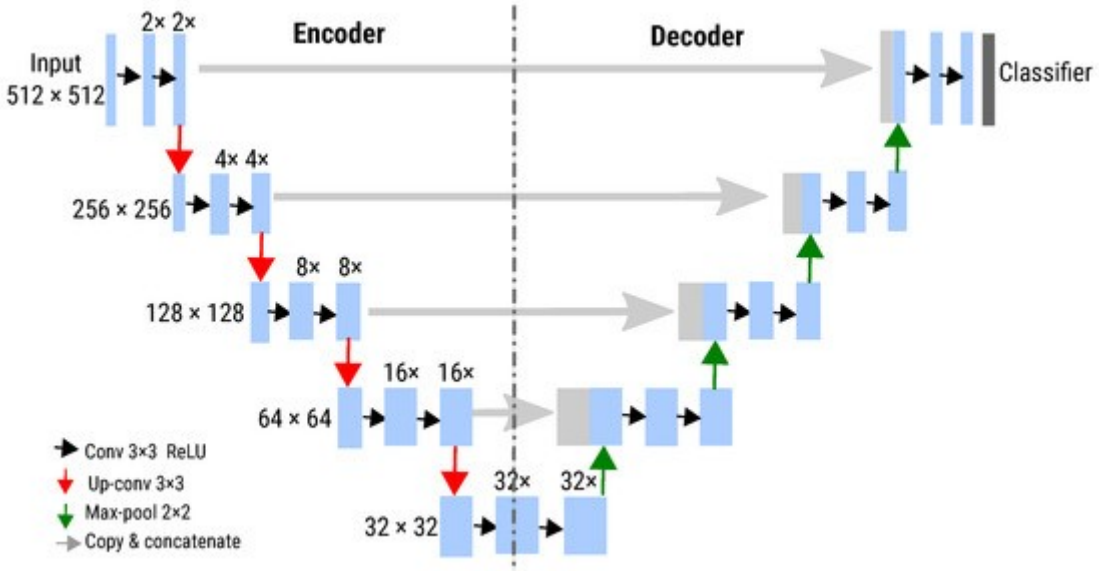


Figure 7 shows the U-Net structure for SAR image segmentation. It consists of two paths: encoder and decoder. The encoder is a pre-trained classification network (ResNet) where convolution blocks followed by max-pool downsampling are applied to encode the input image into feature representations at several levels. Each block is a convolution operation and follows a ReLU activation function. The red arrows indicate a 2 × 2 max-pooling layer.

Deep



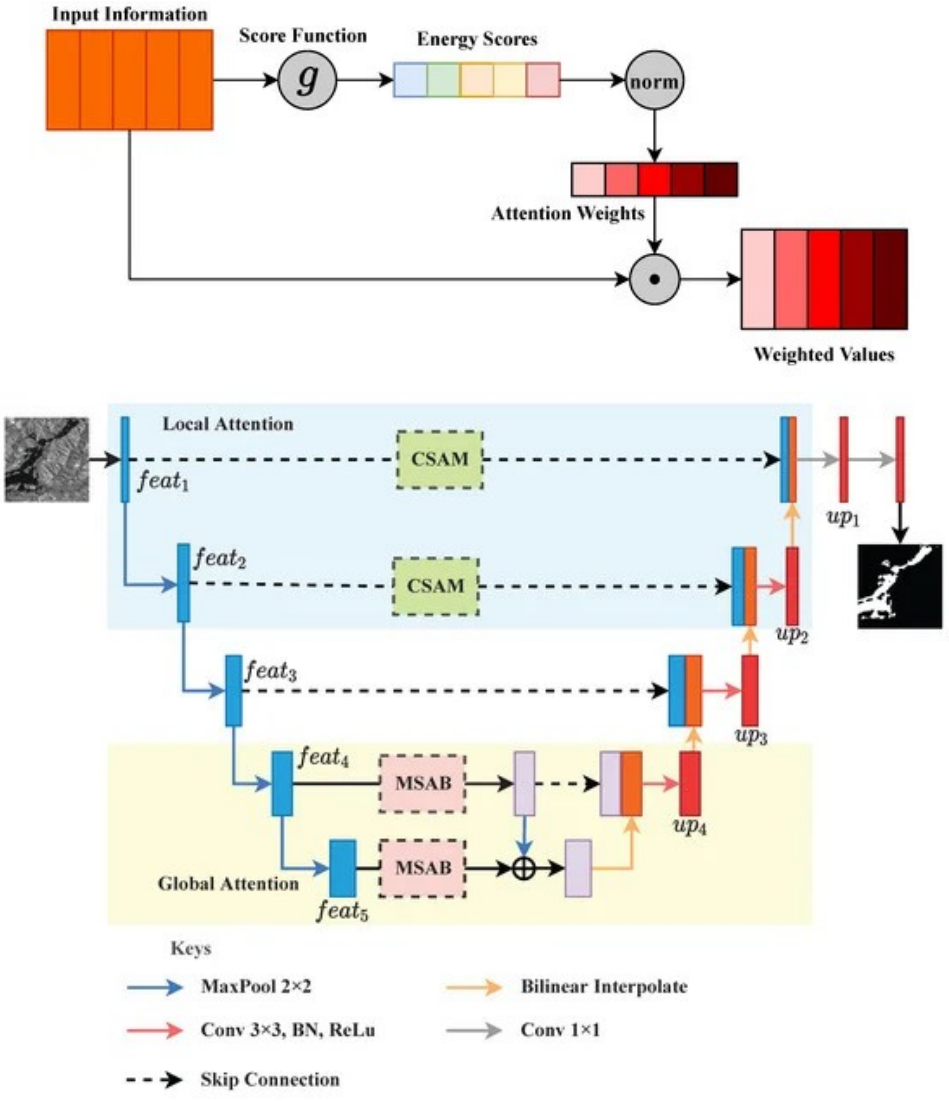
learning methods to obtain high precision for pixel classification require a rigorous training process. This process is achieved by integrating many samples and interactions in the neural network to ensure an adequate representation of the patterns present in the image. As shown in Table 7, training the model with 256 samples (chips) and 25 epochs resulted in precision of 82%, recall of 40%, and F1 of 53%. However, when training the model with 1036 samples and 100 epochs, the precision was 94%, recall 92%, and F1 93%.

It is important to note that the training process is not limited to integrating samples and iterations. It also requires many tests and evaluations to determine the effects of the different training and classification parameters on model performance. This allows for adjustment and optimizes the generated models, obtaining better results in pixel classification. It is essential to have powerful hardware resources to implement and train deep learning models that automate different cartographic tasks. This is because DL models require a large volume of data to train, and processing these data involves many computations and complex mathematical operations. By having powerful hardware resources, such as a high-capacity GPU or many processing cores, the training time can be reduced and a more significant amount of data can be processed, which improves the accuracy of the modeled tasks. In other words, the processing power can be improved and training times accelerated, enabling the better performance of DL models and higher accuracy in water body mapping tasks.

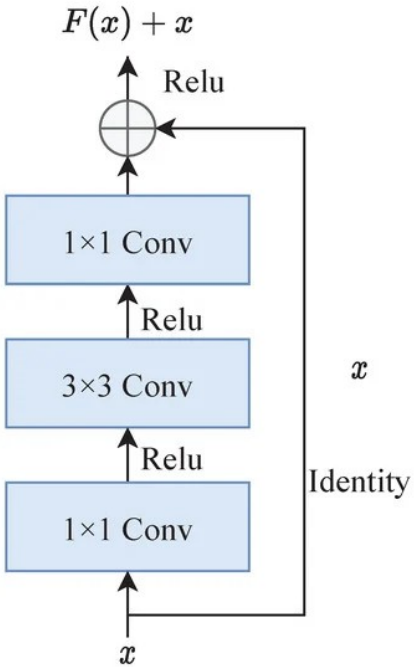
On the other hand, repeatedly using the same DL model in the same study area may be associated with limitations and errors. This is because floods can alter the topography and terrain characteristics. Therefore, if the model is trained with pre-flood images and used to classify post-flood images, it may not capture terrain changes and new features that may emerge. Another issue is that floods can vary in magnitude and extent over time. If the model is trained on historical data and applied to more recent imagery, there may be significant differences in flood conditions. This can lead to a lack of model adaptability and decreased classification accuracy. On the other hand, if the training data used for the model have biases or limitations, such as limited coverage of flood events or a lack of diversity in lighting conditions and scale, the model may not be able to generalize, producing incorrect or biased results.

HA-Unet: A Modified Unet Based on Hybrid Attention for Urban Water Extraction in SAR Images (Electronics, MDPI, 2022) – Song et al.

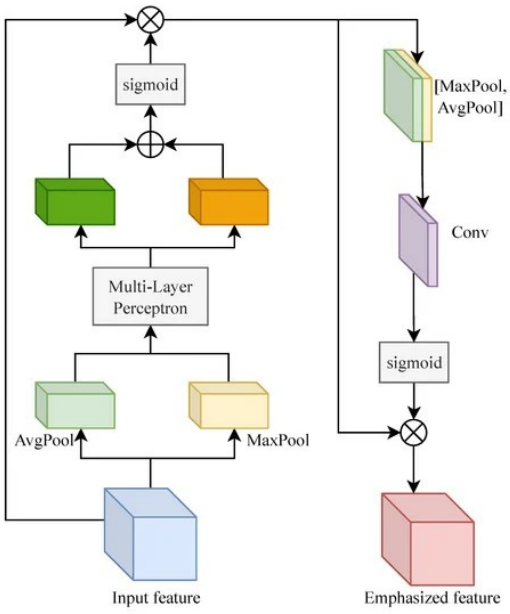
Urban water plays a significant role in the urban ecosystem, but urban water extraction is still a challenging task in automatic interpretation of synthetic aperture radar (SAR) images. The influence of radar shadows and strong scatters in urban areas may lead to misclassification in urban water extraction. Nevertheless, the **local features captured by convolutional layers in Convolutional Neural Networks (CNNs)** are generally redundant and cannot make effective use of global information to guide the prediction of water pixels. To effectively **emphasize the identifiable water characteristics and fully exploit the global information of SAR images, a modified Unet based on hybrid attention mechanism is proposed** to improve the performance of urban water extraction in this paper. Considering the feature extraction ability and the global modeling capability in SAR image segmentation, the Channel and Spatial Attention Module (CSAM) and the Multi-head Self-Attention Block (MSAB) are both introduced into the proposed Hybrid Attention Unet (HA-Unet). In this work, Resnet50 is adopted as the backbone of HA-Unet to extract multi-level features of SAR images. 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 [33]. The key idea of Resnet is the deep residual learning framework, and the deep residual learning framework in Resnet is presented in Figure 6. 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 [32]. Taking into account the number of parameters and training difficulty, Resnet50 is adopted for water feature extraction in this work and the structure of Resnet50 backbone is presented in Table 2.

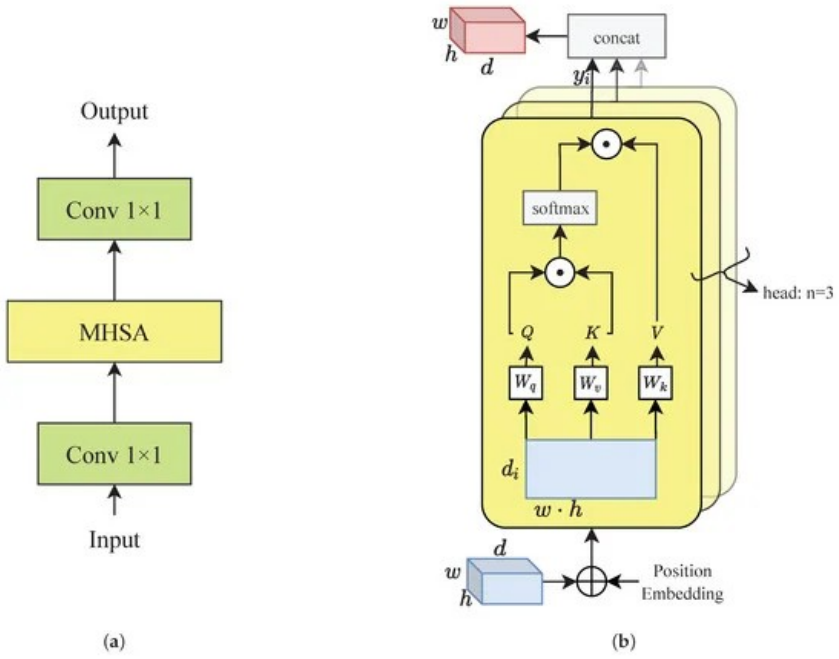


Identifiable and representative feature representations are essential in high accuracy segmentation. However, these features extracted by CNNs are redundant, especially at low stage, and they may influence the implicit representation of CNNs [34]. Thus, a feature selecting approach is necessary for urban water extraction. In the proposed HA-Unet, 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. The schematic of CSAM is presented in **Figure 7**.



After 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 dimension and spatial dimension, while the unnecessary ones are suppressed.

Segmentation is a task that requires accurate pixel-level predictions. Not only fine-grained features, but also long-range dependencies are crucial to resolving the ambiguities of local pixel prediction [35]. In large-scene SAR images, intrinsic correlations among pixels are beneficial to improve classification accuracy, especially for small regional segmentation [21]. Nevertheless, 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 shown in **Figure 8** 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.



As shown in **Figure 8**, the input feature *x* of MHSA layer is appended with positional embedding. With the parallel execution of *n* heads, the MHSA layer is able to learn the richer non-local context. Considering the high computational complexity in MHSA, only three MHSA layers with four heads are used to construct MSAB in our experiment. 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.

The loss function is an essential parameter for CNN training. 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 in this work, which is defined as
$L=Lce+Ld$,

The urban water extraction results generated by DeeplabV3+ [36], original Unet [22], and the proposed HA-Unet.

The proposed HA-Unet is structurally innovative, employing the channel and spatial attention mechanism and the multi-head self-attention mechanism at the same time. First, the low semantic features are enhanced by CSAM to improve the expression of water features. Then, the deep feature maps after MSAB can capture more feature references for local predictions. The urban water extraction results and the quantitative indexes indicate that the proposed HA-Unet is more effective than the original Unet as well as

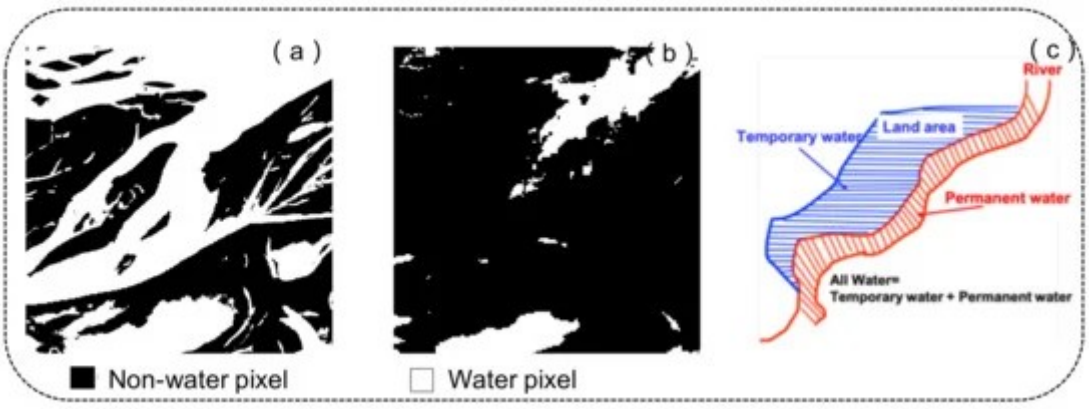
DeeplabV3+ in urban water extraction. Additionally, three enlarged regions A, B and C intuitively show that HA-Unet with the hybrid attention has fewer omission errors and commission errors even in complex scenes. Furthermore, the ablation experiment and visualization results vividly show the important role of HA-Unet in urban water extraction. In comparing the two attention modules, either of the two modules can improve the performance of original Unet gradually in urban water extraction. Thanks to the identifiable water features emphasized by CSAM and the long-range dependency, the proposed HA-Unet can better understand the characteristics of water boundaries, locations, and shapes to improve the urban water extraction accuracy.

there are still limitations and shortcomings to our work. Since MSAB can model long-range dependencies of SAR images, the massive number of parameters in the multi-head self-attention mechanism makes it difficult to meet the requirements of real-time inferencing, in high-resolution SAR images. In this work, MSAB is performed only in the last two stages in Resnet50 to achieve a balance between efficiency and accuracy. In the future, further research will be continued to explore a more efficient attention mechanism to improve the efficiency of the water extraction algorithm.

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 Sensing, MDPI, 2021)

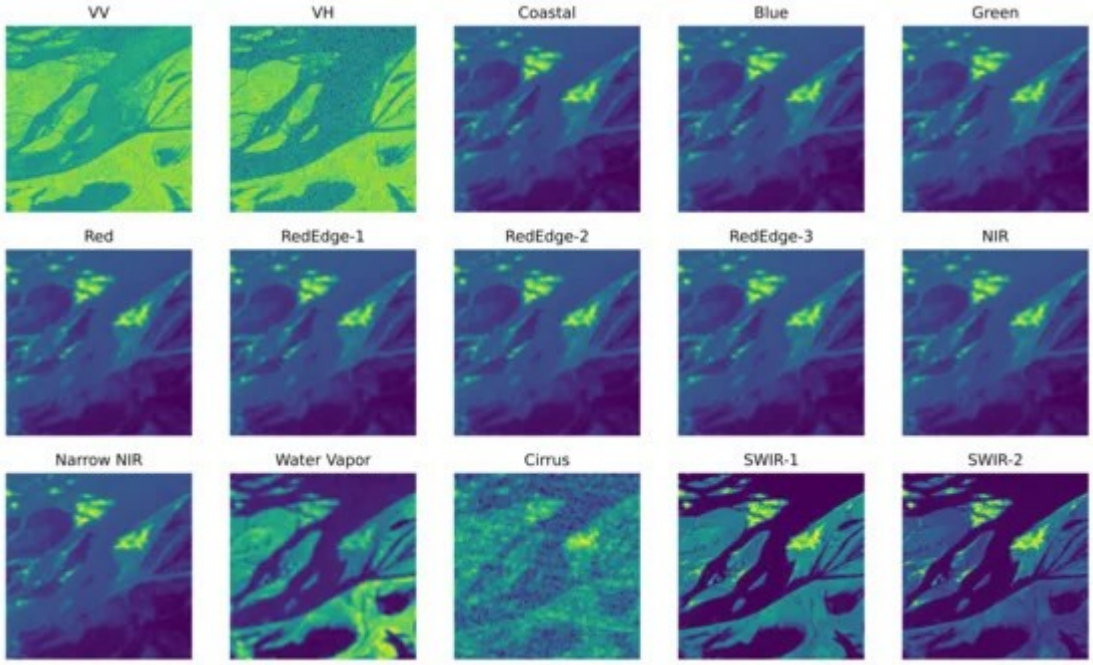
Identifying permanent water and temporary water in flood disasters efficiently has mainly relied on change detection method from multi-temporal remote sensing imageries, but estimating the water type in flood disaster events from only post-flood remote sensing imageries still remains challenging.

Here, we present 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. Specifically, we proposed an automatic segmentation method for surface water, permanent water, and temporary water identification, and all tasks share the same convolutional neural network architecture. We utilize focal loss to deal with the class (water/non-water) imbalance problem. Thorough ablation experiments and analysis confirmed the effectiveness of various proposed designs. In comparison experiments, the method proposed in this paper is superior to other classical models. Our model achieves a mean Intersection over Union (mIoU) of 52.99%, Intersection over Union (IoU) of 52.30%, and Overall Accuracy (OA) of 92.81% on the Sen1Flood11 test set. On the Sen1Flood11 Bolivia test set, our model also achieves very high mIoU (47.88%), IoU (76.74%), and OA (95.59%) and shows good generalization ability.



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. All cloud and cloud shadow pixels were masked and excluded from training and accuracy assessments. 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. For the permanent water label, with the help of the JRC (European Commission Joint Research Center) surface water data set, Bonafilia et al. [5] labeled the pixels that were detected as water at both the beginning (1984) and end (2018) of the dataset as permanent water pixels. The pixels never observed as water during this period are treated as non-water pixels. The remaining pixels are masked. Examples of water label are visualized in Figure 3.



Like most existing studies [42], the Sen1Floods11 dataset shows the highly imbalanced distribution between flooded and unflooded area. As shown in **Table 1**, for all water, 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.

The Sen1Floods11 we utilized in this study can be reached at **https://github.com/cloudtostreet/Sen1Floods11** (accessed on 10 November 2020).

Figure 4 depicts a flowchart of our work. In this work, the benchmark Sen1Floods11 dataset that contains 4831 samples with 512 × 512-pixel size from both Sentinel-1 and Sentinel-2 imagery were utilized to develop the algorithm. The spatial resolution resampling and pixel value normalization were adopted to fuse sentinel-1 and sentinel-2 imagery. The model input is a stack of fused image bands with permanent water and temporary water annotations. The network used is BASNet proposed by Qin et al. [43] is used. BASNet architecture is shown in Figure 5. The network combines a densely supervised encoder–decoder network similar to U-Net and a new residual refinement module. The encoder–decoder produces a coarse probability prediction map from the image input, and the Residual Refinement Module is responsible for

learning the residuals between the coarse probability prediction map and the ground truth. We apply the network to remote sensing data sets, adapt, train, and optimize it to better predict flood areas.

The **encoder–decoder network** can fuse abstract high-level information and detailed low-level information and is mainly responsible for water body segmentation. The encoder part contains an input convolution block and six convolution stages consisting of basic res-blocks. The input convolution block comprises a convolution layer with batch normalization [44] and Rectified Linear Unit (ReLU) activation function [45]. The size of the convolution kernel is 3 × 3, and the stride is 1. This convolution block can convert an input image of any number of channels to a feature map of 64 channels. The first four convolution stages directly use the four stages of ResNet34 [46]. Except for the first residual block, each residual block will double the feature map's channels. The last two convolution stages have the same structure, and both consist of three basic res-blocks with 512 filters and a 2 × 2 max pooling operation with stride 2 for downsampling.
Compared with traditional convolution, atrous convolution can obtain a larger receptive field without increasing parameter amount and capture long-range information. In addition, atrous convolution can avoid the reduction of feature map resolution caused by repeated downsampling and allow a deeper model [47,48]. To capture global information, Qin et al. [43] designed a bridge stage to connect the encoder and the decoder. This stage is comprised of three atrous convolution blocks. Each atrous convolution block consists of a convolution layer with 512 atrous 3 × 3 filters with dilation 2, a batch normalization, and a ReLU activation function.

In the decoder part, each decoder stage corresponds to an encoder stage. As shown in the Formulas (1) and (2), gi is the merge base, fi is the feature map to be fused, hi is the merged feature map, and $conv3\times3$ is a convolution layer followed by batch normalization and ReLU activation, RRM represents the Residual Refinement Module and the operator [·;·] represents concatenation along the channel axis. There are three convolution layers with batch normalization and a ReLU activation function in each decoder stage. The feature map from the last stage is first fed to an up-sampling layer to get $gi+1$, and then concatenated with the current feature map fi. In order to alleviate overfitting, the last layer of the bridge stage and each stage of the encoder is fed to a 3×3 convolution layer followed by a bilinear up-sampling layer and sigmoid activation function to generate a prediction map and then supervised by the ground truth:

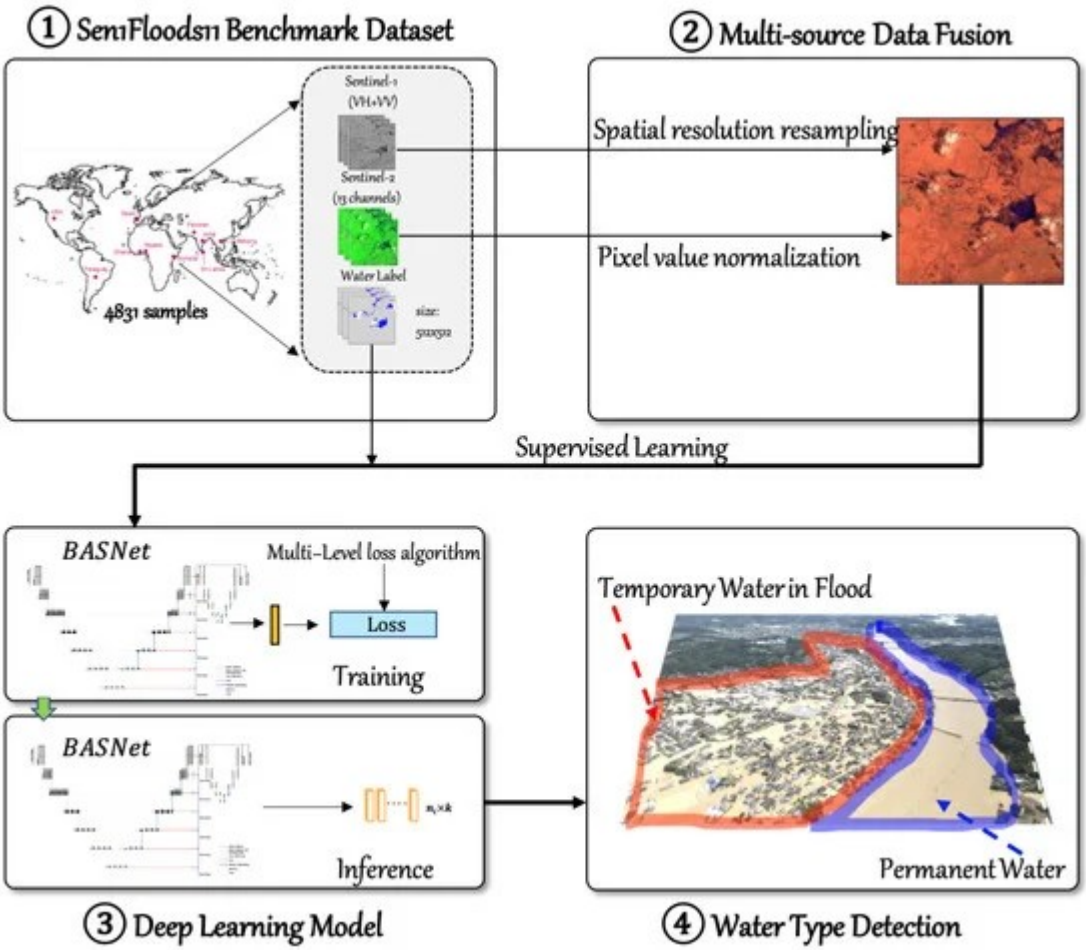
The **residual refinement module** learns the residuals between the coarse maps and the ground truth and then adds them to the coarse maps to produce the final results. By fine-tuning the prediction results, the fuzzy and noisy boundaries can be made sharper. The probability gap between water and non-water pixels can be increased. Compared to the encoder–decoder network, the residual refinement module has a simpler architecture containing an input layer, a four-stage encoder–decoder with a bridge, and an output layer. Each stage has only one convolution layer followed by a batch normalization and a ReLU activation function. The convolution layer has 64 3 × 3 filters with stride 1. In addition, down-sampling and up-sampling are performed through non-overlapping 2 × 2 max pooling and bilinear interpolation in the encoder–decoder network.

Training loss is defined as the summation of all outputs' losses. Each loss is comprised of three parts: focal loss [49], Structural SIMilarity (SSIM) loss [50], and Intersection over Union (IoU) loss [51].
We replace Binary Cross Entropy (BCE) loss in Qin et al. (2019) [43] with focal loss.
Comparison with other loss functions: Besides focal loss, distributional ranking (DR) loss [57] and normalized focal loss [58] are also proposed to address the class imbalance problem.
DR loss [57] treats the classification problem as a ranking problem and improves object detection by distributional ranking supplementary. The distributional ranking model ranks the distributions between positive and negative examples in the worse-case scenario. As a result, this loss can handle a class imbalance problem and the problem of imbalanced hardness of negative examples as well as maintaining efficiency. In addition, it can separate the foreground (water) and background (non-water) with a large margin.

After fusing Sentinel-2 optical imagery and Sentinel-1 SAR imagery, results improve significantly on all tasks, which demonstrates that optical imagery can provide useful supplementary information on water segmentation.

These methods include the Otsu thresholding method based on the VH band [5], FCN-ResNet50 [5], Deeplab v3+ [59], and U2-Net [60].

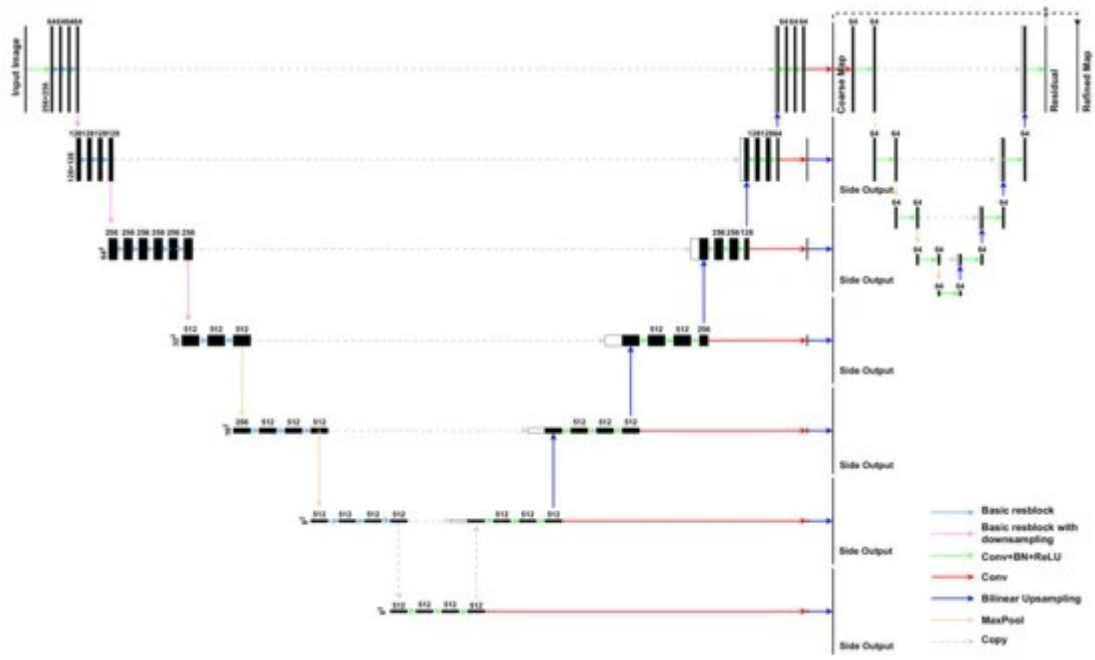
The Otsu thresholding method [5,15] is a widely used method in water body extraction. It can degenerate a grayscale image into a binary image with the best threshold to distinguish the two different types of pixels. The between-class variance is calculated according to the specific algorithm corresponding to Otsu [15]. Then, find the threshold corresponding to the largest between-class variance as the best threshold. This method is unsupervised, simple, and fast. ResNet [46] is utilized as a standard backbone in most networks. Bonafilia et al. [5] use the fully convolutional neural networks (FCNN) model with a ResNet50 backbone to map floods. Here, we compare our method with it. Our tasks can also be regarded as segmentation tasks. Chen et al. [48] proposed Deeplab for semantic segmentation. Here, we try to use it to map floods and compare it with our model. We use the latest version of Deeplab(Deeplabv3+ [59]) as our comparison. We replace its Aligned Xception with Resnet-50 [46] to decrease the parameters amount and computational complexity. Moreover, to account for the relatively small batch size, we convert all of the batch normalization layers to group normalization layers. Considering water bodies as the salient object, we can solve our problems by SOD models. U2-Net [60] is a SOD network and has mainly two advantages compared with previous architectures. First, it allows training from scratch rather than from existing pre-trained backbones, which avoids the problem of distributional differences between RGB images and satellite imagery; second, it can achieve deeper architecture while maintaining high-resolution feature maps at a low memory and computation cost.



Data augmentation can improve the generalization ability of the model, especially when there is only a little training data. Since the Bolivia test set consists of completely unknown flood event images, the gains indicate that the model's generalization ability has been improved. Focal loss [49] is designed to deal with the extreme imbalance between water/non-water, difficult/easy pixels during training. The ablation experimental results show the effectiveness of focal loss in dealing with a sample imbalance problem. Optical imagery contains information on the ground surface's multispectral reflectivity, which is widely used in water indices and thresholding methods. Image fusion aims to use optical image data to assist SAR image prediction. Our experimental results demonstrate that optical imagery can provide useful supplementary information on water segmentation.

However, all water and temporary water have poor mIoU scores. There are two reasons to explain this phenomenon. One reason is the training process, as the image of all water and temporary water data have more water pixels and fewer non-water pixels. The difference in the sample size leads to differences in the learning effect. From OMISSION and COMMISSION, we can see that the all water and temporary water tasks perform better than permanent water on water pixels, and perform worse on non-water pixels. On the whole, non-water pixels dominate our data. The poorer prediction of non-water pixels leads to worse overall results. The other reason is the difference in image characteristics, and the images in all water and temporary water data contain more small tributaries and scattered areas from newly flooded areas. These areas are usually more challenging to identify.

With the help of hybrid loss, our model pays more attention to boundary pixels and increasing the confidence of the prediction. As a result, our method can not only produce richer details and sharper boundaries but also distinguish water and non-water pixels with a larger probability gap. The excellent feature extraction ability of deep learning model enables our model to deal with some challenging scenes.



S1S2-Water: A Global Dataset for Semantic Segmentation of Water Bodies From Sentinel- 1 and Sentinel-2 Satellite Images
Wieland et al. (JSTARS, 2024)

This study introduces the S1S2-Water dataset—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 Sentinel-1 and Sentinel-2 images with quality-checked binary water mask. Samples are drawn globally on the basis of the Sentinel-2 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. On the basis of this dataset, we carry out performance evaluation of CNN architectures to segment surface water bodies from Sentinel-1 and Sentinel-2 images. We specifically 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. We also evaluate the performance impact when a model is trained on permanent water data and applied to independent test scenes of floods.

The S1S2-Water dataset follows recent guidelines for the construction of remote sensing benchmark datasets as much as possible. A stratified random sampling was performed to be representative of a variety of climatic, atmospheric and land cover conditions, to cover different seasons and to ensure that samples always include water areas in addition to other land cover classes (Figure 1). In addition, a fixed division into training, validation and test splits is defined to ensure the transparency and repeatability of experiments. The dataset follows Open Source standards regarding data formats and structure. Each sample consists of Sentinel-1 Ground Range Detected (GRD) Interferometric Wide (IW) swath data and Sentinel-2 L1C images with associated quality-controlled binary water mask annotations, elevation and slope layers as well as metadata (Figure 2). Each sample is aligned to the Sentinel-2 tile-grid and covers an extent of 100 x 100 km with 4 a pixel-spacing of 10 m.

The final dataset consists of 65 samples (each 100 x 100 km in size) that are spread across 29 countries and cover an area of approximately 650,000 km². The samples cover 18 pre-dominant landcover types and show a wide distribution across elevation and slope. Images have been acquired between 2018-05-21 and 2020-11-26 and are distributed across nearly all months of the year (no samples are available for December). The difference between Sentinel-1 and Sentinel-2 acquisitions per sample is 1 day in average with a standard deviation of 4 days. We decided against tiling the samples into smaller patches, but provide a Python package along with the data to do so instead. This way, the user has greater flexibility to prepare the samples according to requirements of the desired network architecture and hardware setup. A tiling of S1S2-Water into non-overlapping 256 x 256 pixels patches, results in a total of 50,000+ patches for training and 25,000+ patches for validation and testing respectively. Dataset and preparation package are released openly alongside this publication [reference will be provided once peer-reviewed version of this article has been published].

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, we developed an independent reference dataset S1S2-Flood that covers 12 major flood events across the globe (Table I). Similar to S1S2-Water, each sample of this S1S2-Flood dataset consists of Sentinel-1 GRD and Sentinel-2 L1C images with associated quality-controlled binary water mask annotations, elevation and slope layers as well as metadata (Figure 3). Maximum time difference between acquisition dates of Sentinel-1 and Sentinel-2 images has been limited to one day to ensure spatial and temporal consistency of the flood masks. We used the same procedure as for S1S2-Water samples to annotate the satellite images. First, water bodies are roughly identified using a threshold procedure based on NDWI and then manually refined for each sensor specifically in multiple

Similar to S1S2-Water, each sample of this S1S2-Flood dataset consists of Sentinel-1 GRD and Sentinel-2 L1C images with associated quality-controlled binary water mask annotations, elevation and slope layers as well as metadata (Figure 3). Maximum time difference between acquisition dates of Sentinel-1 and Sentinel-2 images has been limited to one day to ensure spatial and temporal consistency of the flood masks. We used the same procedure as for S1S2-Water samples to annotate the satellite images. First, water bodies are roughly identified using a threshold procedure based on NDWI and then manually refined for each sensor specifically in multiple iterations using extensive quality checks and corrections.

The models that we test in this study are based on the widely-used U-Net architecture [33], which has proven to deliver highly accurate results for water segmentation tasks in high-resolution satellite images at relatively low computational complexity [8], [15]. In combination with the U-Net architectures we compare different encoders, namely MobileNet-V3, ResNet-50, EfficientNet-B0 and EfficientNetB4. We selected these for our experiments since they show a good trade-off between number of model parameters and Imagenet Top-1 accuracy [34].

For Sentinel-2 we use only spectral bands that are available across different satellite sensors (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. In this experiment we test the influence of different data augmentation techniques on the segmentation performance. Specifically, we apply random contrast, brightness, scale and image flipping.

In this experiment we aim at answering the question, whether a model trained on images depicting normal water can be transferred to flood images. Compared to normal water, flood water is mostly characterized by higher contents of sediment and debris. Flooded vegetation, infrastructure or vehicles may further interact with the water surface and modify the reflectance characteristics of the target class in the satellite images. As baseline, we evaluate the performance of sensor-specific models that have been trained solely on normal water (S1S2-Water) and apply them to the test images of our independent flood water dataset (S1S2-Flood). We then compare these results to models that have been trained on a joint dataset that includes normal and flood water images (S1S2-Water + S1S2-Flood).

We observe an improvement of 0.10 IoU compared to using VV polarization alone and 0.06 IoU compared to using VH polarization alone (Table III). In our experimental setup, 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 [29], [38]. 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, our results also underline the theoretical assumption that a combination of both polarizations works best in practice. For Sentinel-2, 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 (Table IV).

Similar to the findings of Bonafilia et al. (2020) [11] our experiments confirm that adding flood water samples to the training data supports model transfer and increases test scores on the independent S1S2-Flood dataset.

Training a model with a joint training dataset that contains normal and flood water samples (TM-1) outperforms fine-tuning a pre-trained model on S1S2-Water with additional flood samples (TM-2) on Sentinel-1. On Sentinel-2 images both approaches show the same improvement of test scores. James et al. (2021) [6] analyzed a similar model transfer for water segmentation in Sentinel-2 images. Even though their transfer experiment targeted geographical differences, their results emphasize the added value of retraining with limited samples of the target domain.

. Across several experiments we identified the superior performance of U-Net Efficientnet-B0 models, which show good generalization ability across varying environmental conditions and produce high accuracies at high throughput in both SAR and multi-spectral images. In this context, not only the choice of model architecture and encoder is relevant, but also the sensor-specific input feature space and the way training data are augmented.

By successfully applying the model to six flood events (S1S2-Flood) and independent test splits of other benchmark datasets (**Sen1Floods11 and WorldFloods**), we could highlight the usefulness of this work for rapid mapping activities to support situational awareness in emergency response. This underlines the findings of previous work of the authors that it is possible to train a model that is able to cope with highly diverse data availability scenarios in disaster situations [39].

---

https://www.iceye.com/blog/how-to-build-an-icecube-for-supervised-machine-learning-with-time-series-sar-images

https://eo4society.esa.int/projects/sar2cube/

https://www.youtube.com/watch?v=dnrDfLDciGQ&t=9s

TV-Net: A Structure-Level Feature Fusion Network Based on Tensor Voting for Road Crack Segmentation
Zheng et al., (IEEE TITS, 2024)

Segmentation using traditional convolutional neural network training can lose feature information when the network depth goes larger, which makes accurate prediction a challenging topic. To address these issues, we propose a new approach that features an enhanced tensor voting module and a customized pixel-level pavement crack segmentation network structure, called TV-Net. We optimize the tensor voting framework and find the relationship between tensor scale factors and crack distributions. A tensor voting fusion module is introduced to enhance feature maps by incorporating significant domain maps generated by tensor voting. Additionally, we propose a structural consistency loss function to improve segmentation accuracy and ensure consistency with the structural characteristics of the cracks obtained through tensor voting.

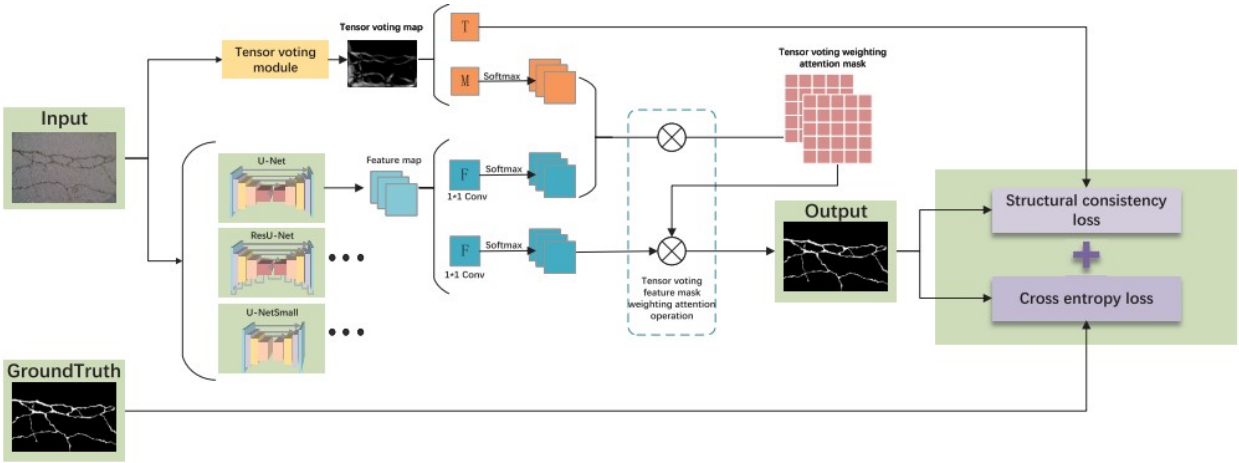


Fig. 4. Structure diagram of the proposed TV-Net.

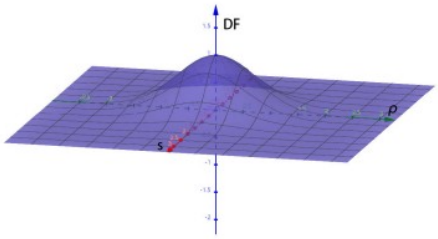


Fig. 5. Tensor voting field function diagram.

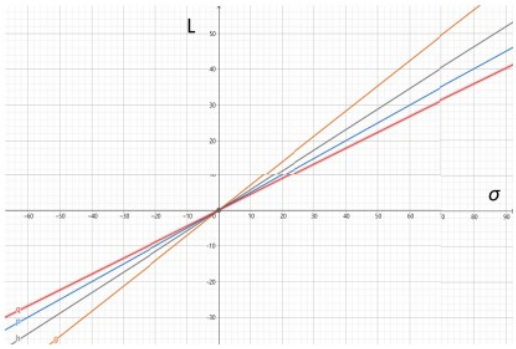


Fig. 6. Graph of L as a function of σ.

Crack Segmentation Methods

Image segmentation holds a significant position in the field of image recognition, with its methods being extensively applied across various domains, particularly in the segmentation of road defects. One of the classical segmentation methods is the Fully Convolutional Network (FCNet) [23], which is based on the classic Convolutional Neural Network (CNN) and has been successfully applied to semantic segmentation tasks. FCNet adopts a skip connection structure, integrating features from different network layers to ensure the accuracy of segmentation, and uses convolutional upsampling (deconvolution) for the fusion of high-level and low-level features. Building upon the FCNet, PSP-Net [41] proposes a pyramid pooling module by

integrating context information from different regions, effectively utilizing global information. PSP-Net introduces more context information based on the FCNet algorithm through global average pooling operations and feature fusion, resulting in a pyramid structure of features.

Building upon the FCNet, PSP-Net [41] proposes a pyramid pooling module by integrating context information from different regions, effectively utilizing global information. PSP-Net introduces more context information based on the FCNet algorithm through global average pooling operations and feature fusion, resulting in a pyramid structure of features.

The U-Net [39] algorithm framework mainly consists of an encoder and a decoder. The encoder uses convolutional layers and pooling layers to progressively reduce the size and dimensions of the feature map while increasing the number of channels, extracting high-level features of the input image. The decoder then uses deconvolution layers (or upsampling) and convolution layers to gradually restore the size and dimensions of the feature map, resulting in a feature map of the same size as the original image. U-Net also introduces a skip mechanism, connecting the features of the corresponding layer in the encoder with those in the decoder, helping to retain more spatial information and detailed features. This skip mechanism allows U-Net to utilize feature information from different levels to enhance the accuracy and robustness of image segmentation. Based on U-Net, ResU-Net [37] adds a skip connection structure.

The DeepLabV3+ [42] model is mainly divided into encoding and decoding modules. In the encoding module, high and low-level feature semantic maps are extracted from the input images through the backbone network, and the extracted high-level feature semantic maps are entered into the Atrous Spatial Pyramid Pooling (ASPP) module. This module includes parallel 1 × 1 convolution, atrous convolution with multiple rates, and a global average pooling layer. The feature map obtained after ASPP processing maintains the same channel dimension as the low-level feature semantic map in the decoding module, and they are spliced and fused together. Through upsampling, the spatial information of the image is gradually recovered, capturing the boundaries of the target cracks more clearly

==

It can be seen that, due to the characteristics of pavement image, namely the presence of noise and changing reflectance, U-Net will lose the structural feature map information of cracks as the training level deepens in the process of training.

To address this issue, this paper introduces a tensor voting module into the network architecture. Tensor voting is a perceptual grouping method that infers significant structures from images through the tensor representation of features and nonlinear voting. This allows the network to highlight key areas and reduce the impact of noise on the segmentation results. The introduction of the tensor voting module helps It can be seen that, due to the characteristics of pavement image, namely the presence of noise and changing reflectance, U-Net will lose the structural feature map information of cracks as the training level deepens in the process of training. To address this issue, this paper introduces a tensor voting module into the network architecture. Tensor voting is a perceptual grouping method that infers significant structures from images through the tensor representation of features and nonlinear voting. This allows the network to highlight key areas and reduce the impact of noise on the segmentation results. The introduction of the tensor voting module helps to overcome the problem of feature information loss due to the monotonous semantic information and high levels of noise interference in road surface images.

Model Architecture

TV-Net: The basic structure of TV-Net is described as follows 4. TV-Net is a plug-and-play structure, and the backbone network can be composed of U-Net, ResU-Net, U-NetSmall, etc. The image size of input and

output is 512*256. During the up-sampling part, the network adopts convolution instead of the max-pooling method to gradually reduce the spatial dimension of input data and extract highdimensional features. Convolution is also used to replace the max-pooling method in the down-sampling part. After each convolution operation, one Batch Normalization (BN) layer and one Rectified Linear Units (ReLu) activation function are added. To standardize the characteristics of each layer in the network. BN layer can make the feature distribution of each layer more reasonable, expand the fault-tolerant ability of the network, and shorten the convergence time of the model. The Relu activation function can achieve better gradient descent and back-propagation.

The feature map generated by the backbone network is convolved with the kernel value of 1 before the softmax operation. The Tensor voting map generated by the Tensor voting module also operates softmax. After the first tensor voting feature mask weighted attention operation, a tensor voting weighted attention mask is obtained. Then, using the probability map of backbone network output, the second tensor voting feature mask weighted attention operation is carried out, and finally, the output is obtained. The loss function is divided into two parts. The first part is the structural consistency loss between the probability graph generated by the tensor voting graph and the output. The second part is the cross entropy loss between the ground truth and the output. After multiplying and adding their respective weight coefficients, the final loss function is obtained.

Tensor Voting Module: According to the tensor voting theory, the size of the tensor field is determined by the scale factor σ. A larger tensor field will include more data in the voting process, which can suppress noise but may also lose some structural details. A smaller tensor field will include less data in the voting process, which can retain more structural details but may also be less robust to noise. Furthermore, the recoverable gap size between road crack breakpoints is related to the sparsity of crack pixels in road images and the optimal scale factors required for different data distributions can be calculated as follows.

==

The specific steps are as follows: 1) The road image is grayed and coded into a triplecoordinate x, coordinate y, and direction angle. Fig. 6. Graph of L as a function of σ. 2) The initial tensor coding is performed for each triplet pixel, and the coding is a spherical tensor. 3) The initial tensor direction and salient features are obtained by sparse sphere tensor voting. 4) According to the tensor theory, when the difference between λ1 and λ2 is particularly large, highly significant linear target pixels can be retained through threshold segmentation. Through these steps, the feature map of tensor voting can be obtained, as shown in the figure7.

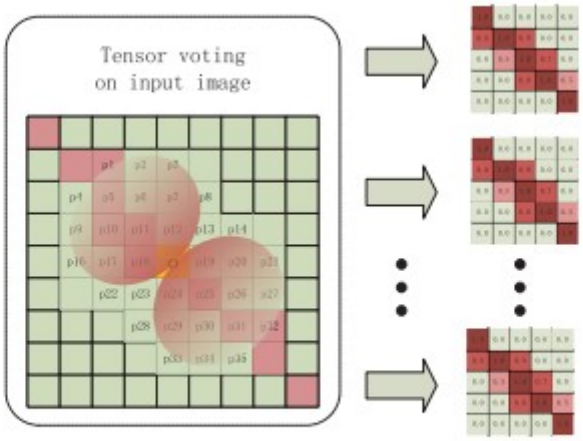


Fig. 7. The feature map of tensor voting.

Tensor Voting Weighted Attention Module: The cracks are characterized by strong structures and prominent ground lineaments. The method of tensor voting can well extract the prominent linear structure attention map. Figure 8 shows the weighted attention operation of the tensor voting feature. The last layer features of the model are weighted with the features extracted by the tensor voting module.

Loss Function According to the characteristics of the road crack image, most of them are crack-free areas, and the road crack area only accounts for a small part of the entire image. Therefore, the binary cross entropy loss function is used to process the loss function between the road crack data set, the image label, and the image segmentation target.

Here n represents the number of image pixels, p represents the predicted pixel value, and q represents the corresponding label pixel value. According to the principle of tensor voting, the salient feature of the target obtained by tensor voting highlights the main structural target and retains the details of the target. Through experimental verification, its recall value reaches 0.98, so the salient feature map obtained by tensor voting is used as the second type of pseudo label, and the structural consistency loss function of the second type of pseudo label and segmentation target is proposed. The tensor voting graph is used as a pseudo-label. During training, the loss function only adopts the prediction result of the positive region and the cross entropy of the pseudo-label as a loss function. In training, it can play the purpose of regularizing the penalty function to punish pixels beyond the feature region of the tensor voting structure.

Comparing the U-Net with the tensor voting module fusion and U-Net, it can be seen from Figure 10 that the U-Net model with the tensor voting module fusion can detect the broken cracks, and the U-Net usually takes the broken part of the cracks as the background. Moreover, the U-Net model incorporating the tensor voting module can have no structural noise interference, which indicates that the tensor voting module can detect the structural characteristics of cracks.