PDCA-FORMER: PRIOR-DIAGONAL CROSS ATTENTION-GUIDED TRANSFORMER FOR FLOOD MAPPING FROM SAR IMAGERY: A CASE IN KHARTOUM

Tamer Saleh[1,2], Mohamed Zahran[2], Shimaa Holail[1], Gui-Song Xia[1,3,*]

[1] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, China (tamer.mohamed@feng.bu.edu.eg)
[2] Geomatics Engineering Department, Faculty of Engineering at Shoubra, Benha University, Cairo, Egypt
[3] National Engineering Research Center for Multi-media Software, School of Computer Science and Institute of Artificial Intelligence, Wuhan University, China (guisong.xia@whu.edu.cn)

KEY WORDS: Flood Mapping, Sentinel-1 SAR, Deep Learning, Diagonal-Cross Attention Module (DCAM), Transformer, Remote Sensing, Khartoum-Sudan.

ABSTRACT:

Floods are considered one of the most serious crises, with severe consequences such as loss of life, destruction of infrastructure, and economic disruption. In recent years, deep learning has gained popularity for fast and accurate flood mapping from synthetic aperture radar (SAR) for damage assessment and proactive mitigation. However, due to the complex characteristics of SAR images, accurate flood mapping remains challenging. In this study, we propose a novel Prior-Diagonal Cross Attention-guided transformer (PDCA-Former) network for flood mapping from SAR images. Specifically, PDCA-Former adopts Prior Siamese Feature Extraction (PSFE) to extract multi-scale deep features from the input SAR images. Additionally, we propose a novel Diagonal Cross-Attention Module (DCAM) to capture relational information of all pixel positions on the entire image. DCAM is integrated into the Transformer to acquire contextual tokens with spatio-temporal information from prior features, resulting in immersed maps. To investigate the potential of SAR images and the proposed PDCA-Former for effective flood detection and estimation of the extent of damaged farmland around the confluence of two rivers, this study chose the Sudanese city of Khartoum as an experimental study area. The experimental results show that PDCA-Former outperforms the latest comparator methods in terms of F1 by 88.9% and IoU by 85.7%. We conclude that PDCA-Former offers a promising solution for accurate and efficient flood mapping from SAR imagery that can be quickly generalized to other regions. Therefore, it can significantly aid disaster management efforts on vulnerable communities.

1. INTRODUCTION

The long-term and devastating consequences of floods are multifaceted, encompassing loss of human lives, damage to critical infrastructure, disruption of economic activities, and significant financial losses (EM-DAT, 2008). Floods often occur as a result of heavy and prolonged rainfall that exceeds the natural capacity of the land to absorb water, leading to surface runoff in low-lying areas and exacerbating the risks of flooding. Vulnerable communities with limited resources and inadequate infrastructure, such as Sudan, are particularly impacted by floods (NASA, 2020). For instance, in the summer of 2020, Sudan experienced a catastrophic flood triggered by heavy rainfall in 17 Sudanese states, affecting over 650,000 people and resulting in over 100 fatalities. The cities of Khartoum and Omdurman were particularly hard-hit, with significant impacts on urban and heritage areas (OCHA, 2020). Satellite images in Figure 1 vividly illustrate the magnitude of the problem, revealing a flood near the Black Nile River around Omdurman, located approximately 310 miles northeast of Khartoum, with water levels reaching about 17.4 meters above normal levels. These images represent valuable contributions to the field of disaster management and provide an excellent opportunity for further research on the impact of floods on vulnerable communities. Given the substantial scale of the issue, it is imperative to implement real-time flood mapping using satellite-based Earth monitoring technologies to accurately assess the extent of the disaster and proactively mitigate its future impacts. Satellite technology is distinguished by unique advantages in flood mapping, as it offers multi-temporal, rapid, and wide coverage of the Earth's surface compared to costly and time-consuming ground-based survey scenarios. Optical remote sensing images have limitations in terms of their weak ability to penetrate clouds and their reliance on daylight operation . Synthetic aperture radar (SAR) images are an ideal choice for flood detection, as they can operate under adverse weather conditions . Previous studies have shown that Khartoum is a flood-prone area, as reported in (Chu and Lee, 2022); (Buğday and Buğday, 2019); (Elhaja et al., 2017); (Mahmood et al., 2017); These studies relied on analyzing digital elevation models (DEMs) and hydraulic modeling data, along with optical images, to assess the impact of floods. In addition, land cover maps derived from unsupervised classification were used as reference data. However, due to the limited spatial-temporal resolution of the data, the accuracy addressed by these studies does not indicate how similar the classification result is to the actual land cover. On the other hand, thresholding methods have demonstrated efficient flood detection using SAR images, as reported in (Wangchuk et al., 2022); (Chen et al., 2021); (Liang and Liu, 2020); (Qiu et al., 2021). However, accurate flood mapping using thresholding methods is extremely challenging due to the complex characteristics of SAR images. Moreover, they are susceptible to noise interference, lack spatial consistency, and are unable to handle complex nonlinear issues.

In recent decades, deep learning has become increasingly popular due to its significant advantages in representing deep features and modeling nonlinear problems, and has been widely used in flood detection. For example, (Zhang and Xia, 2022)

* Corresponding author

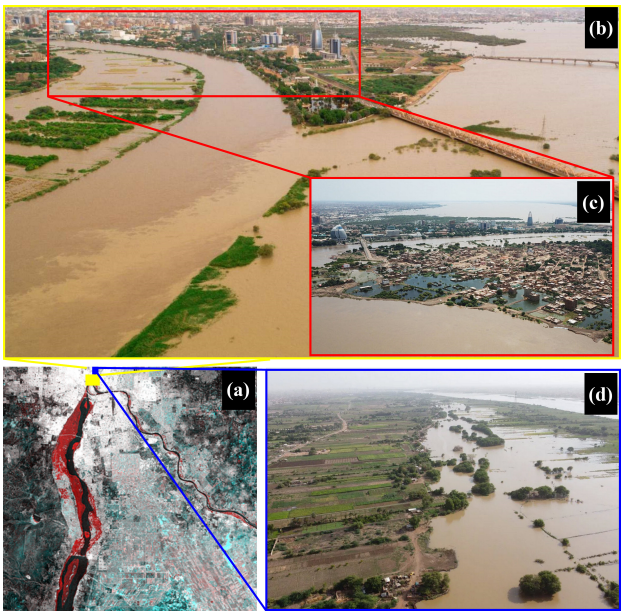


Figure 1. Depicts four images related to the floods that occurred in the confluence of the Blue and White Niles, Sudan. (a) Displays an RGB Sentinel-1 SAR image acquired on September 23, 2020. Panel (b) shows an aerial view of the extent of the flood's impact on the capital, Khartoum. Panel (c) provides another aerial view, revealing buildings and roads submerged by floodwaters near the Nile River. (d) Illustrates a flooded farmland after the water level of the Nile River rose in the El Halfaya region of Khartoum. It is worth noting that the awe-inspiring composition credits are attributed to Copernicus Sentinel-1© ESA, CC BY-SA 3.0; PreventionWeb© UNDRR, CC BY-ND 4.0; and REUTERS.

proposed a convolutional neural network (CNN) based on Unet to suppress speckle noise in GF-3 SAR images for flood detection in Lake Poyang, and the network achieved good results. To mitigate the impact of limited annotated training datasets, (Li et al., 2019) introduced an active self-learning CNN (A-SL CNN) framework for urban flood detection using multi-temporal TerraSAR-X data during Hurricane Harvey in August 2017. Various other deep learning networks based on CNN for flood detection can be found in the literature, such as CMCD-Net (He et al., 2023), CNN-Unet (Tavus et al., 2022), FWENet (Wang et al., 2022), attention Unet-MSL (Xu et al., 2022), and so on. Despite their significant achievements in detection, they still struggle to connect distant spatio-temporal concepts due to the limitation of the size of the receptive field. In recent years, transformers have achieved competitive performance in flood detection compared to their CNN counterparts, as they possess a large receptive field and model long-range dependencies in the data, benefiting from the multi-head self-attention mechanism . Examples of such methods include FloodTransformer (Roy et al., 2022), BiT-STANet-SNUNet (Dong et al., 2023), FL-Former (Le et al., 2023), DA-Transformer (Li et al., 2022), Trans-SANet+ (Zhou et al., 2022), DAM-Net (Saleh et al., 2023), and so on. However, these methods are not effective in capturing temporal information by ignoring the spatial interactions of neighboring pixels.

To address the aforementioned issues comprehensively, the current study selected Khartoum city as an experimental study area to investigate the potential of SAR images combined with CNN and Transformer models for effective flood detection and automatic mapping applications. In this research, we propose a novel Prior Diagonal-Cross Attention-based Transformer network (PDCA-Former) for flood mapping from SAR images. PDCA-Former comprises three main components, namely, Prior Siamese Feature Extraction (PSFE), Transformer Decoder Block (TDB), and Prediction Head (PH). PSFE extracts multi-scale deep features from parallel branches with shared weights based on the ResNet-50 network (He et al., 2016). We introduce a novel Diagonal-Cross Attention Module (DCAM) that captures contextual information for all pixel positions based on the complete image context, while reducing the complexity of the self-attention mechanism. The TDB further enhances image features by employing the CSWin Transformer (Dong et al., 2022), which leverages multi-head attention mechanisms to effectively capture rich and contextual information for detecting changing pixel units. Finally, the PH receives the output of the TDB to obtain the final results of flooded areas, employing weighted entropy and dice loss during network training to measure the similarity between the predicted change map and the ground truth.

The contribution of this paper is twofold. Firstly, we introduce a novel Prior Diagonal-Cross Attention-based Transformer network (PDCA-Former) for flood mapping from SAR imagery. PDCA-Former effectively collects contextual information from various receptive fields simultaneously and address issues of distortion and noise. Secondly, we propose a novel Diagonal-Cross Attention Module (DCAM) that captures contextual information for all pixel positions based on the complete image context, while reducing the complexity of the self-attention mechanism.

The remaining parts of this paper are organized as follows. In Section 2, the proposed method and evaluation metrics are listed. The study area, dataset, training details, and results obtained are described in Section 3. Finally, some conclusions are drawn in Section 4.

2. PROPOSED METHODOLOGY

2.1 Overview

Figure 2 presents an overview of the proposed method, PDCA-Former, which utilizes Siamese backbones, the diagonal-cross attention module (DCAM), and the CSWin Transformer block to accurately process pre- and post-flood input images ($t1$ and $t2$) for detecting flooded area maps through the following; (1) Siamese backbones simultaneously receive the input images of pre-flood ($t1$) and post-flood ($t2$) in two branches with shared weights. This generates multi-scale deep features ($F1$ and $F2$) for both images, which are then processed by a deep convolutional layer before being fed into attention modules; (2) The DCAM, a combination of the cross-attention module (CAM) and diagonal-attention module (DAM), captures contextual information of pixels based on the whole image dependencies. Specifically, $F1$ is fed into the CAM, while $F2$ is fed into the DAM, resulting in $H1$ and $H2$, respectively. A 3×3 convolutional layer is then applied to obtain the output $O1$ and $O2$, which are fused into $O12$ through a concatenation process; (3) $O12$ is further processed by a convolutional layer, and its output is divided into embedding tokens (denoted by $E12$). These tokens are then fed into the CSWin Transformer block (Dong et al., 2022), which leverages multi-head attention mechanisms to effectively capture context-rich, high-level embedded semantic

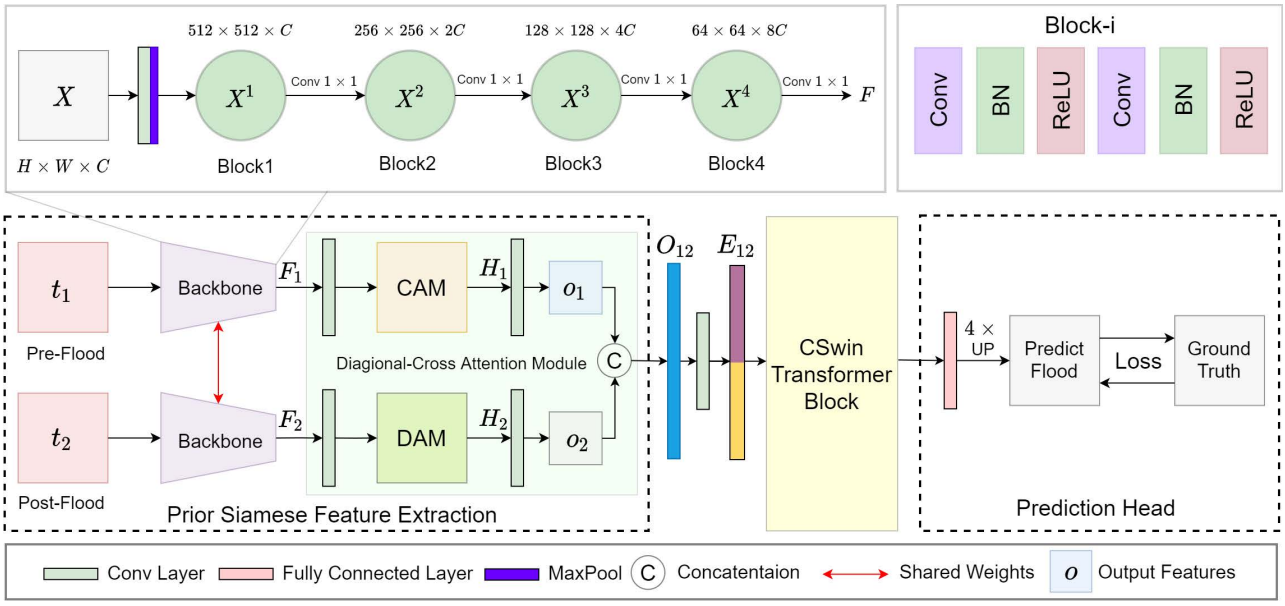


Figure 2. Overview of our proposed network for flood detection from SAR imagery.

information for detecting changed pixels; and (4) The prediction head receives the output of the Transformer decoder, compiles the deep features back to the original size, and obtains the final results of the flooded areas.

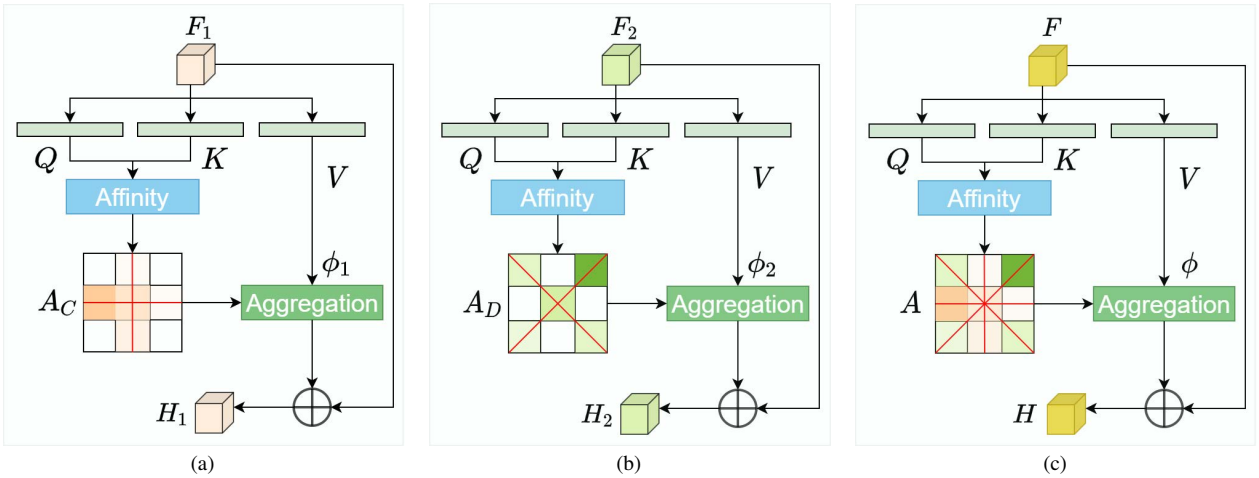
2.2 Prior Siamese Feature Extraction

The core component of the prior siamese feature extraction (PSFE) module comprises of two parallel backbones based on a pre-trained ResNet-50 network (He et al., 2016), which efficiently extract features. As shown in Figure 2 (top part), the backbone consists of four X^n blocks, where $n \in \{1, 2, 3, 4\}$. Each block encompasses a series of operations, including Convolution (Conv), Batch Normalization (BN), and Rectified Linear Unit (ReLU). Prior to entering the first block, the input image $\mathrm{X} \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times\mathrm{C}}$ undergoes a 7×7 Convolution layer followed by MaxPooling operation. Subsequently, downsampling operations are employed to generate multi-scale hierarchical features, resulting in the output of the final block as $X^4 \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 8C}$. To obtain the deep feature F, a 1×1 convolutional layer is applied to each block. To further explore contextual information in F, we propose the DCAM by blending the CAM with the DAM. The CAM captures contextual information of pixels in horizontal and vertical directions from the input F_1, which greatly reduces the number of weights, resulting in H_1. On the other hand, the DAM captures the contextual information of pixels in diagonal directions from the input F_2, obtaining H_2. Then, the attention features are condensed for both H_1 and H_2 using a 3×3 convolutional layer. To learn deep features that can effectively represent submerged regions with changing scope, the O_1 and O_2 features are combined through a concatenation process, which are the outputs of the CAM and DAM, respectively, to obtain the feature encoder O_{12}. As a result, the DCAM provides additional contextual information for all pixel positions based on the full image dependencies, while reducing the complexity of the self-attention factor. The proposed DCAM is illustrated in Figure 3. The local feature map $F \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 8C}$, obtained from the backbone and convolutional layers, is used to generate attention feature maps H from two self-attention modules, each shown in Figure 3. In the CAM, three feature maps Q, K, and V are generated by three convolutional layers with F_1, where $\{\mathrm{Q}, \mathrm{K}\} \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times\mathrm{C}'}$ and $\{V\} \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times\mathrm{C}}$, and C' represents the channel number that is reduced from C to lower the dimensionality. Subsequently, the attention map $A_C \in \mathbb{R}^{(H+W-1)\times W\times H}$ is generated from Q and K using an affinity operation, calculating the correlation between pixels in vertical and horizontal directions. Next, $\phi_1 \in \mathbb{R}^{(\mathrm{H}+\mathrm{W}-1)\times\mathrm{C}}$ is computed as a set of feature vectors in V that are in the same row or column as the position. The aggregation process in equation 1 is then applied to obtain H_1, where the contextual information is collected through summation of $A_C^i \phi_1^i$ with F_1 for $i \in \{0, 1, ..., H + W - 1\}$. The same process is repeated for the DAM to obtain H_2. Finally, H_1 and H_2 are combined to obtain the feature map H, which contains aggregated contextual information from all directions for each position.

$$\mathrm{H}_1 = \sum_{\mathrm{i}=0}^{\mathrm{H}+\mathrm{W}-1} \mathrm{A}_{\mathrm{C}}^{\mathrm{i}} \phi_1^{\mathrm{i}} + \mathrm{F}_1 \tag{1}$$

2.3 Transformer Decoder Block

The Transformer decoder is responsible for further optimizing the image features from E_{12} to output E'_{12}, as depicted in Figure 4. The input to the Transformer decoder is E_{12}, which is obtained by combining O_1 and O_2 features and then split into trainable embedding tokens that contain embedded high-level semantic information. The main block in the Transformer decoder is the shifting window multi-head cross-attention (SW-MHCA), which aims to integrate stronger spatio-temporal tokens of E_{12} into O_{12} deep features. More specifically, in SW-MHCA, the query (Q) is generated from E_{12}, while the key (K) and value (V) are generated from H, denoted by equation 2. Layer normalization (LN) is applied before the SW-MHCA and MLP output to normalize the activations and stabilize the training process. Residual connections are also used, allowing the model to retain information from previous layers and avoid the vanishing gradient problem. Next, the contextual representations are passed through the MLP, which consists of three linear layers and a Gaussian Error Linear Unit (GeLU),



(a) (b) (c)

Figure 3. The cross-attention module (CAM) represents in (a), (b) represents the diagonal attention module (DAM), and the combined DCAM represents in (c).

allowing the model to learn more complex patterns and representations. Finally, E'_{12} is inserted into the prediction head to generate prediction flood maps.

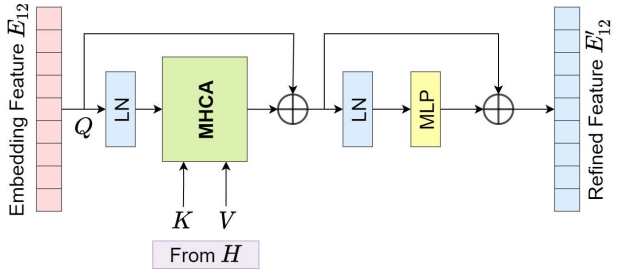


Figure 4. The transformer decoder block.

$$\begin{aligned} \mathrm{Q_{SW}}, \mathrm{K_{SW}}, \mathrm{V_{SW}} &= \mathrm{E_{12}P^{Q}_{SW}}, \mathrm{H_1P^{K}_{SW}}, \mathrm{H_2P^{V}_{SW}} \\ (\mathrm{Q_{SW}}, \mathrm{K_{SW}}, \mathrm{V_{SW}}) &= \sigma\left(\frac{\mathrm{Q_{SW}K_{SW}}^{\mathrm{T}}}{\sqrt{\mathrm{d}}}\right)\mathrm{V_{SW}} \end{aligned} \tag{2}$$

where P^Q_{SW}, P^K_{SW}, and P^V_{SW} = the weights of the three linear layers in the SW-MHCA to obtain Q_{SW}, K_{SW}, and V_{SW} respectively

d = the dimension of Q_{SW}, K_{SW}, and V_{SW}, which is used to avoid division by 0

σ = the SoftMax function

2.4 Prediction Head

After extracting and integrating the previous fragments through PSFE and the Transformer decoder, an improved bi-temporal feature, denoted as E'_{12}, is obtained for flood map detection. E'_{12} is first passed through a fully connected layer, and the output is enlarged four times using bi-linear interpolation before being fed into the Sigmoid prediction function. Since the distribution of change and non-change classes is often highly unbalanced in the detection task, the loss function may be biased towards the class with larger samples during training, resulting in lower class recognition accuracy for the class with smaller samples. Therefore, we use the weighted entropy function to train the network, which is defined as in equation 3. Furthermore, the dice loss is employed to measure the similarity between the predicted change map Pr and the ground truth GT, and its value ranges from 0 to 1. The dice loss is defined as in equation 4. The total loss used to train our method is defined as in equation 5.

$$\begin{aligned} \mathcal{L}_{\mathrm{bce}}(\mathrm{Pr}, \mathrm{GT}) = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}} [&-\omega_c (\mathrm{GT}_i) \log (\mathrm{Pr}_\mathrm{i}) \\ &-\omega_{\mathrm{uc}} (1 - \mathrm{GT}_\mathrm{i}) \log (1 - \mathrm{Pr}_\mathrm{i})] \end{aligned} \tag{3}$$

$$\mathcal{L}_{\mathrm{dice}}(\mathrm{Pr}, \mathrm{GT}) = 1 - \frac{2|\,\mathrm{Pr} \cap \mathrm{GT}|}{|\,\mathrm{Pr}\,| + |\mathrm{GT}|} \tag{4}$$

$$\mathcal{L}_{\mathrm{T}}(\mathrm{Pr}, \mathrm{GT}) = \mathcal{L}_{\mathrm{bce}}(\mathrm{Pr}, \mathrm{GT}) + \mathcal{L}_{\mathrm{dice}}(\mathrm{Pr}, \mathrm{GT}) \tag{5}$$

where Pr = prediction
GT = ground truth
ω_c = weights of the changed pixels
ω_{uc} = weights of the unchanged pixels
i = pixel index

2.5 Evaluation Metrics

To compare the ground truth and predicted change map, five metrics were utilized to validate the accuracy and effectiveness of the proposed method. These metrics include Recall (R), Precision (P), IoU, Overall Accuracy (OA); and F1-score. The formulas for these metrics are provided below:

$$\begin{aligned} \mathrm{Recall} &= \frac{TP}{TP + FN} \\ \mathrm{Precision} &= \frac{TP}{TP + FP} \\ \mathrm{OA} &= \frac{TP + TN}{TP + FN + TN + FP} \\ \mathrm{IoU} &= \frac{TP}{TP + FP + FN} \\ \text{F1-score} &= \frac{2 \times \mathrm{Recall} \times \mathrm{Precision}}{\mathrm{Recall} + \mathrm{Precision}} \end{aligned} \tag{6}$$

where TP = total number of true positives for all classes
FN = false negatives
FP = false positives
TP = true positives

3. EXPERIMENTAL RESULTS

3.1 Study Area

Khartoum, the capital of Sudan, was chosen as the study area due to the flood season and rainfall between June and September, with an average of approximately 135 mm over the years 2017 to 2020. It is also the confluence of two major tributaries of the Nile River, where the Blue Nile flows from Tana Lake in Ethiopia, and the White Nile flows from the African Great Lakes around the East African Rift. Khartoum is located in the eastern center of the country between latitudes 15 and 16 degrees north and longitudes 32 and 33 degrees east, and is spatially characterized by an elevation of 386 meters above mean sea level, as shown in Figure 5. In September 2020, Khartoum experienced a severe storm caused by heavy rains, resulting in significant impact on the agricultural areas and villages along the banks of the Nile River.

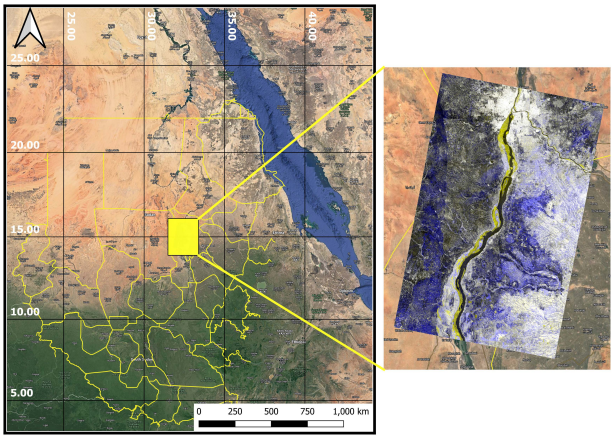


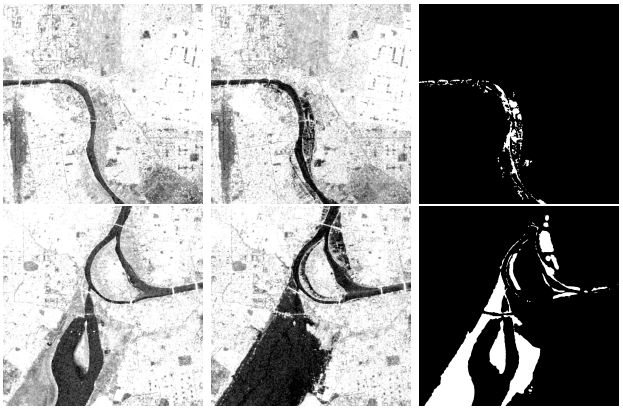
Figure 5. Location of the study area with the RGB of the Sentinel-1 SAR data.

3.2 Dataset

In this study, two Sentinel-1 Ground Range Detected (GRD) images covering the Blue Nile and White Nile region of Khartoum city were used. Level 1 GRD data was acquired in Interferometric Wide (IW) mode with a swath of 250 km and a range and azimuth resolution of 10 m. Only the VV (vertical transmit and vertical receive) polarized channel was used. Sentinel-1 data were downloaded from the Alaska Satellite Facility (ASF) for the period between July 13 and September 23, 2020. Furthermore, the Digital Elevation Model (DEM) from the Shuttle Radar Topographic Mission (SRTM) 3 arc second was used as ancillary data for geometric correction. Table 1 provides information about the data used in this paper. Figure 6 shows a sample of Sentinel-1 and label datasets for training deep learning models. These datasets have dimensions of 512 × 512. The label dataset contains two types of pixels: no-flood and flood. The dataset consisted of 145 pairs of images for training, 42 pairs for validation, and 39 pairs for testing. All images are projected onto WGS84 at a ground resolution of 10 meters.

	Pre-flood	Post-flood
Satellite	Sentinel-1B	Sentinel-1B
Date of Acquisition	13-Jul-20	23-Sep-20
Mode/Swath	IW/GRDH	IW/GRDH
Polarization	VV	VV
Orbit	22447	23497
Number of Images	1	1
Pass	Descending	Descending
Size/Pixels	13558 × 18260	13560 × 18260
DEM	SRTM 3Sec	SRTM 3Sec

Table 1. Remote Sensing Data used for 2020 Khartoum Flood Mapping.



(a) (b) (c)

Figure 6. Khartoum dataset. (a) Image acquired in July 2020 as a pre-event reference. (b) Image taken on September 23, 2020, after the flooding of Sudan. (c) The ground truth.

3.3 Implementation Details

We conducted all experiments on our dataset utilizing a virtual machine desktop with a 64-bit Windows 10 Pro operating system. The software configuration included the Python programming language, PyTorch 1.7.2, CUDA 10.1, and cuDNN 7.6.1. Additionally, we employed an NVIDIA GRID RTX8000-8Q with 8GB memory, an Intel(R) Xeon(R) CPU E5-2687W v4 @ 3.00GHz, and 32.0 GB of GPU memory to ensure optimal hardware capabilities. To ensure model convergence, we trained all methods for 300 epochs employing the Adam optimizer with an initial learning rate of 10e-5. The input images were sized at 512 × 512 with the EfficientNetv2 (Tan and Le, 2021) decoder, while they were 448 × 448 with the CSwin (Dong et al., 2022) decoder, and the batch size for both was set to 4. Due to the Sudan dataset's small size, we applied data augmentation during training to the input image patches. This augmentation included random flip, random rotation, histogram matching, Gaussian blur, and clipping techniques. Lastly, we conducted validation after each training epoch to ensure the effectiveness of the model. The best model on the validation set was saved and subsequently utilized for evaluation on the test set.

3.4 Results and Analysis

We evaluated the effectiveness of PDCA-Former, our proposed network for flood detection, on the Khartoum dataset through a series of experiments, and the quantitative results are presented in Table 2. The results clearly indicate that PDCA-Former outperforms the other methods in most measures. Specifically, PDCA-Former achieved the highest overall accuracy OA, F1-score, and IoU of 98.2%, 88.9%, and 85.7%, respectively, when using a ResNet-50 encoder and CSwin decoder. The high IoU

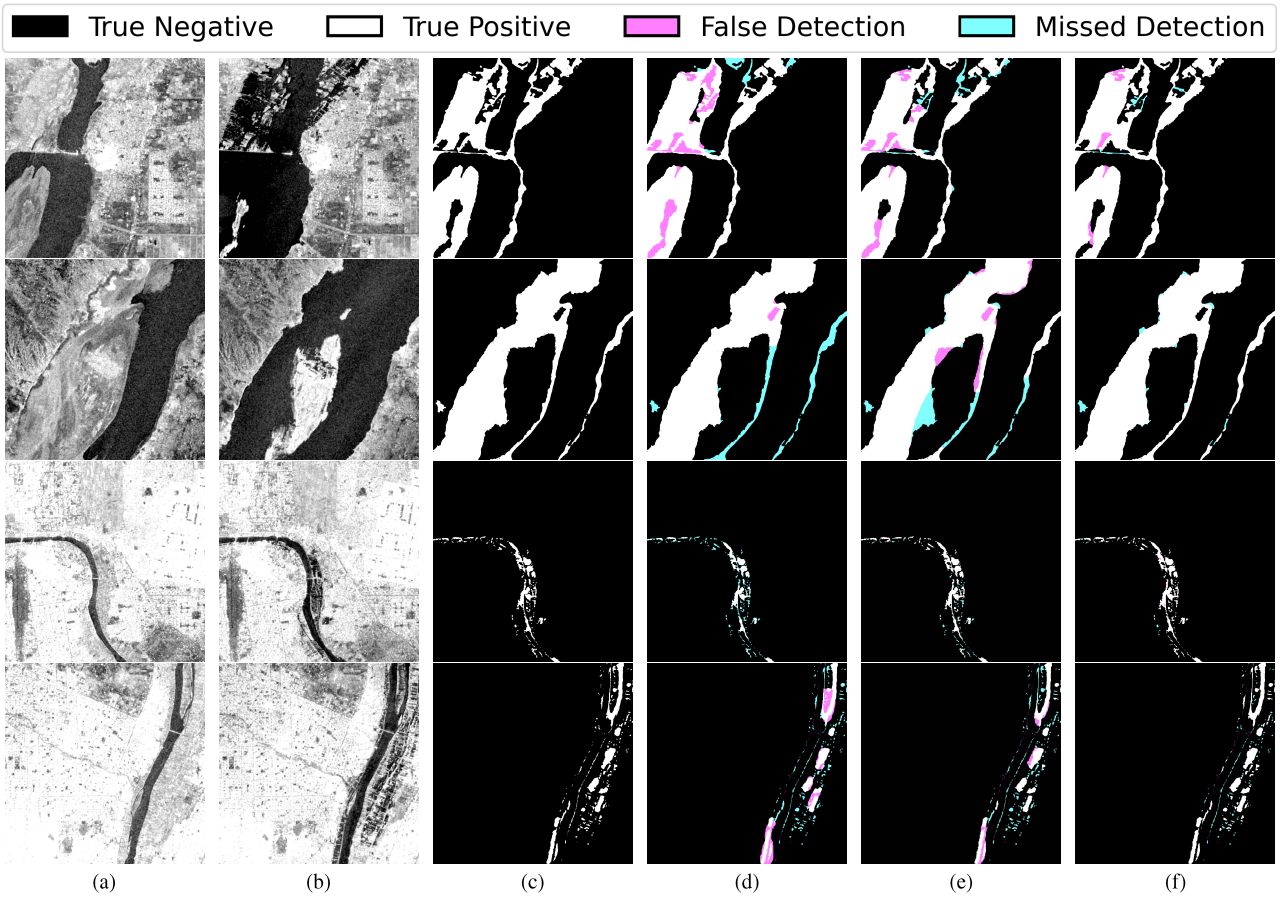


Figure 7. A visualized comparison of flood detection methods on the Khartoum dataset, where (a,b) represent original bi-temporal images, (c) represents the ground truth, and (d) to (f) represent the results of CNN, BiT, and PDCA-Former (our method) respectively. Changed areas are shown in white, and unchanged areas in black. Cyan indicates false positives (FPs), while magenta indicates false negatives (FNs).

value signifies that our model has excellent overlap between the predicted and ground truth regions, indicating the accurate detection of boundary edges and complete objects, as demonstrated in Figure 7. PDCA-Former with the CSwin decoder is found to be an effective method for flood detection, as it outperformed other methods, including the BiT with the CSwin decoder, which had the second-best performance with an IoU of 82.7%. The CNN method with ResNet-50 encoder and CSwin decoder performed the worst, with an F1-score of 77.1% and an IoU of 79.9%. These results highlight the importance of choosing an appropriate decoder in the performance of PDCA-Former, and that our approach has the potential to improve the accuracy of flood detection systems and aid in effective disaster response and management.

To intuitively showcase the performance of our proposed method for flood detection using bi-temporal images, we selected three typical scenes for visualization, as depicted in Figure 7. Our choice of representative scenarios was based on the need to demonstrate the versatility and robustness of our method under different conditions. The first row of visualizations clearly shows that our PDCA-Former method outperforms other methods in detecting flooded areas. Our method achieves near-perfect performance compared to the ground truth in identifying a flooded island within a river. Moreover, our method exhibits fewer false alarms (indicated by the magenta color) and fewer missed detections (indicated by the cyan color) compared to CNN and BIT. This finding suggests that our method is more accurate and reliable in detecting flooded areas than the other methods tested. In the second row of visualizations, we observed the emergence of a previously nonexistent island within the river, caused by the flooding. Despite this challenging scenario, our PDCA-Former method accurately detected the edges of the island. This finding demonstrates the superior ability of our method to recognize changing boundaries of the river caused by the rise in water level. In the last two rows of visualizations, we tested our method in detecting changes in complex farmland regions. Our method demonstrated remarkable performance in detecting changes, while the other two methods resulted in disorderly outputs due to numerous missed detections. This finding suggests that our proposed method is better equipped to detect changes in complex environments, making it a valuable tool for flood monitoring and prevention.

3.5 Flood Inundation Extraction

The present study focuses on the assessment of flooding in the confluence of the two rivers in Sudan, following the disaster that occurred on September, 2020. The proposed PDCA-Former method was utilized to extract the flooded area, which is depicted in Figure 8. The total area of inundation was estimated to be approximately 348.253 sq-km. The proposed method provides a reliable means of quantifying the extent of inundation, allowing for informed decision-making by relevant authorities.

Method	Encoder	Decoder	P	R	OA	F1-score±σ	IoU
CNN (Zhang and Xia, 2022)	ResNet-50	EfficientNetv2	**0.955**	0.646	0.979	0.771± 0.026	0.799
		CSwin	0.943	0.627	0.964	0.753± 0.028	0.783
BiT-Former (Dong et al., 2023)		EfficientNetv2	0.905	0.663	0.963	0.765± 0.029	0.801
		CSwin	0.910	0.782	0.976	0.841± 0.024	0.827
Ours (PDCA-Former)		EfficientNetv2	0.944	0.778	0.968	0.853± 0.022	0.810
		CSwin	0.952	**0.834**	**0.982**	**0.889± 0.018**	**0.857**

Table 2. The average quantitative results obtained by different methods on the test set. The best results are highlighted in bold font, and the second-best results are underlined. σ represents the standard deviation associated with the quantitative results. All values are reported as percentages (%).

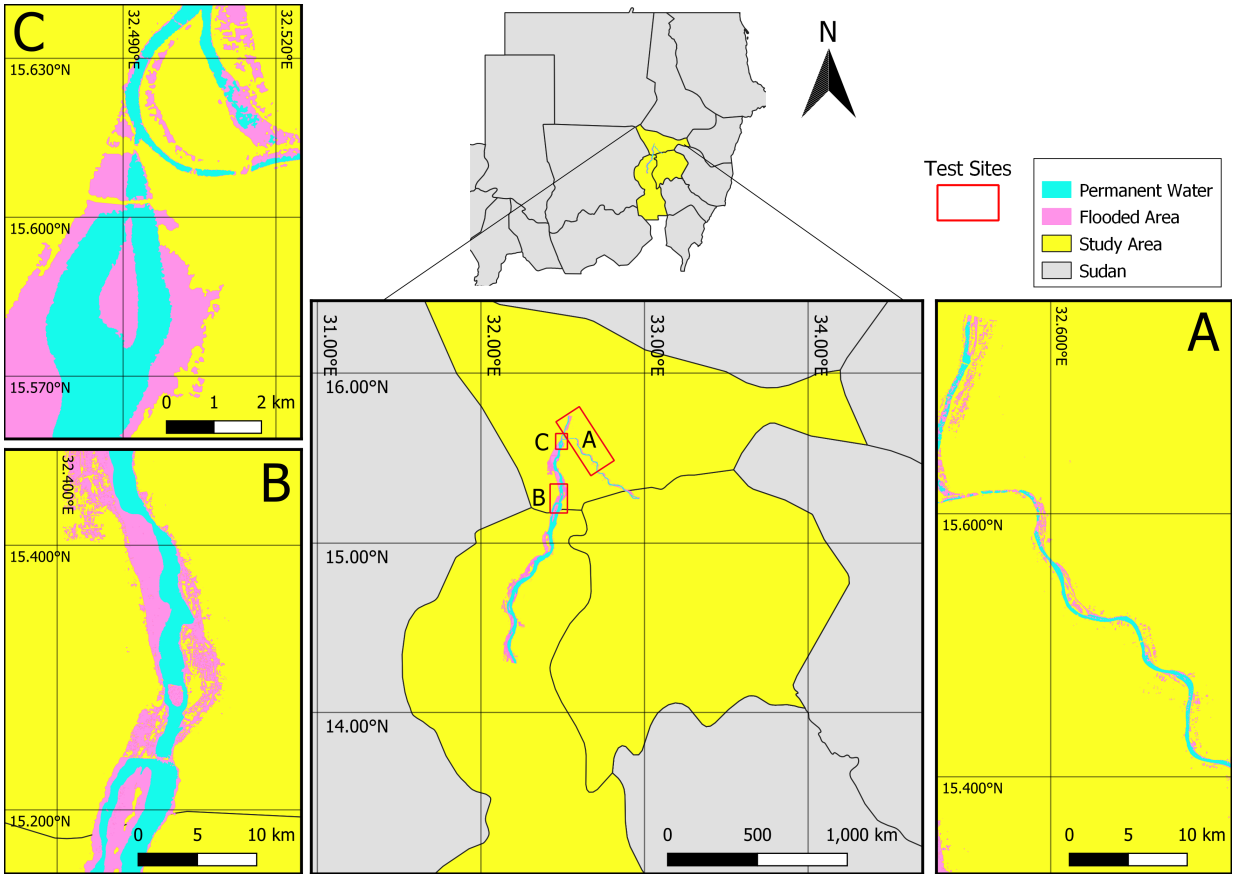


Figure 8. The color-coded maps of flood detection results.

4. CONCLUSIONS

This paper presents the Prior-Diagonal Cross Attention-Guided Transformer (PDCA-Former), a novel network for flood mapping from dual-time Synthetic Aperture Radar (SAR) images. The PDCA-Former network includes the Prior Siamese Feature Extraction (PSFE) and Diagonal Cross-Attention Module (DCAM). The PSFE module extracts multi-scale depth features, while the DCAM module captures relational information for all pixel positions across the entire SAR image. We assessed the performance of the PDCA-Former network on the Khartoum dataset of SAR images from the Sudan flood in 2020 using both qualitative and quantitative methods. The results were compared with those of several other methods, and the findings show that the PDCA-Former network effectively reduces missed detections in flood detection tasks and has superior ability in recognizing changing river boundaries, outperforming the other methods. Future work will include developing a more diverse dataset that includes different categories of floods, such as open and urban floods, to evaluate the PDCA-Former network's performance on different flood types. The proposed PDCA-Former network has potential applications in improving the accuracy of flood mapping, which can aid in disaster management and mitigation efforts.

REFERENCES

Buğday, E., Buğday, S. E., 2019. Modeling and simulating land use/cover change using artificial neural network from remotely sensing data. *Cerne*, 25, 246–254.

Chen, S., Huang, W., Chen, Y., Feng, M., 2021. An Adaptive Thresholding Approach toward Rapid Flood Coverage Extraction from Sentinel-1 SAR Imagery. *Remote Sensing*, 13(23), 4899.

Chu, Y., Lee, H., 2022. Performance of Random Forest Classifier for Flood Mapping Using Sentinel-1 SAR Images. *Korean Journal of Remote Sensing*, 38(4), 375–386.

Dong, X., Bao, J., Chen, D., Zhang, W., Yu, N., Yuan, L., Chen, D., Guo, B., 2022. Cswin transformer: A general vision transformer backbone with cross-shaped windows. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12124–12134.

Dong, Z., Liang, Z., Wang, G., Amankwah, S. O. Y., Feng, D., Wei, X., Duan, Z., 2023. Mapping inundation extents in Poyang Lake area using Sentinel-1 data and transformer-based change detection method. *Journal of Hydrology*, 129455.

Elhaja, M. E., Csaplovics, E., Abdelkareem, O. E., Adam, H. E., El, A., Khalifa, K., Ibrahim, K., Eltahir, M., 2017. Land use land cover changes detection in White Nile State, Sudan using remote sensing and GIS techniques. *Int. J. Environ. Monit. Prot*, 4, 14–19.

EM-DAT, 2008. EM-DAT: The international disaster database. `http://www.emdat.be/Database/Trends/trends.html` (accessed on 10 April 2023).

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, X., Zhang, S., Xue, B., Zhao, T., Wu, T., 2023. Cross-modal change detection flood extraction based on convolutional neural network. *International Journal of Applied Earth Observation and Geoinformation*, 117, 103197.

Le, Q.-C., Le, M.-Q., Tran, M.-K., Le, N.-Q., Tran, M.-T., 2023. Fl-former: Flood level estimation with vision transformer for images from cameras in urban areas. *MultiMedia Modeling: 29th International Conference, MMM 2023, Bergen, Norway, January 9–12, 2023, Proceedings, Part I*, Springer, 447–459.

Li, T., Wang, C., Wu, F., Zhang, H., Tian, S., Fu, Q., Xu, L., 2022. Built-Up Area Extraction from GF-3 SAR Data Based on a Dual-Attention Transformer Model. *Remote Sensing*, 14(17), 4182.

Li, Y., Martinis, S., Wieland, M., 2019. Urban flood mapping with an active self-learning convolutional neural network based on TerraSAR-X intensity and interferometric coherence. *ISPRS Journal of Photogrammetry and Remote Sensing*, 152, 178–191.

Liang, J., Liu, D., 2020. A local thresholding approach to flood water delineation using Sentinel-1 SAR imagery. *ISPRS journal of photogrammetry and remote sensing*, 159, 53–62.

Mahmood, M. I., Elagib, N. A., Horn, F., Saad, S. A., 2017. Lessons learned from Khartoum flash flood impacts: An integrated assessment. *Science of the Total Environment*, 601, 1031–1045.

NASA, 2020. A pair of niles and deltas, NASA. `https://earthobservatory.nasa.gov/images/146824/a-pair-of-niles-and-deltas` (accessed on 20 March 2023).

OCHA, 2020. Sudan floods snapshot, OCHA. `https://www.humanitarianresponse.info/ru/operations/sudan/infographic/sudan-floods-snapshot-6-oct-20202023`).

Qiu, J., Cao, B., Park, E., Yang, X., Zhang, W., Tarolli, P., 2021. Flood monitoring in rural areas of the Pearl River Basin (China) using Sentinel-1 SAR. *Remote Sensing*, 13(7), 1384.

Roy, R., Kulkarni, S. S., Soni, V., Chittora, A. et al., 2022. Transformer-based Flood Scene Segmentation for Developing Countries. *arXiv preprint arXiv:2210.04218*.

Saleh, T., Weng, X., Holail, S., Hao, C., Xia, G.-S., 2023. DAM-Net: Global Flood Detection from SAR Imagery Using Differential Attention Metric-Based Vision Transformers. *arXiv preprint arXiv:2306.00704*.

Tan, M., Le, Q., 2021. Efficientnetv2: Smaller models and faster training. *International conference on machine learning*, PMLR, 10096–10106.

Tavus, B., Can, R., Kocaman, S., 2022. A CNN-based flood mapping approach using sentinel-1 data. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 3, 549–556.

Wang, J., Wang, S., Wang, F., Zhou, Y., Wang, Z., Ji, J., Xiong, Y., Zhao, Q., 2022. FWENet: a deep convolutional neural network for flood water body extraction based on SAR images. *International Journal of Digital Earth*, 15(1), 345–361.

Wangchuk, S., Bolch, T., Robson, B. A., 2022. Monitoring glacial lake outburst flood susceptibility using Sentinel-1 SAR data, Google Earth Engine, and persistent scatterer interferometry. *Remote Sensing of Environment*, 271, 112910.

Xu, C., Zhang, S., Zhao, B., Liu, C., Sui, H., Yang, W., Mei, L., 2022. SAR image water extraction using the attention U-net and multi-scale level set method: flood monitoring in South China in 2020 as a test case. *Geo-Spatial Information Science*, 25(2), 155–168.

Zhang, L., Xia, J., 2022. Flood detection using multiple chinese satellite datasets during 2020 china summer floods. *Remote Sensing*, 14(1), 51.

Zhou, Y., Yang, K., Ma, F., Hu, W., Zhang, F., 2022. Water-land Segmentation via Structure-Aware CNN-Transformer Network on Large-scale SAR data. *IEEE Sensors Journal*.