
Article

Not peer-reviewed version

Deep Learning Integration of CNN-Transformer and UNet for Bi-Temporal SAR Flash Flood Detection

[Abbas Mohammed Noori](#) ^{*}, [Abdul Razzak T. Ziboon](#), Amjad Naser Al-hameedawi

Posted Date: 17 June 2025

doi: 10.20944/preprints202506.1153.v1

Keywords: Deep Learning; Flood Detection; SAR Imaging; U-Net; Transformer Models



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Deep Learning Integration of CNN-Transformer and UNet for Bi-Temporal SAR Flash Flood Detection

Abbas Mohammed Noori ^{1,2,*}, Abdul Razzak T. Ziboon ³ and Amjad N. AL-Hameedawi ¹

¹ Civil Engineering Department, University of Technology, Baghdad, Iraq

² Department of Surveying Engineering, Technical Engineering College of Kirkuk, Northern Technical University, Kirkuk 36001, Iraq

³ College of Engineering, Al-Esraa University, Baghdad, Iraq

* Correspondence: abbasnoori@ntu.edu.iq or bce.23.43@grad.uotechnology.edu.iq

Abstract: A new deep learning approach for bi-temporal flash flood detection in synthetic aperture radar (SAR) is proposed in this research. It combines a U-Net convolutional network with a Transformer model using a compact Convolutional Tokenizer (CCT) to improve the efficiency of long-range dependency learning. The hybrid model, namely CCT-U-ViT, that naturally combines the spatial feature extraction of U-Net and the global context capability of Transformer. The model significantly reduces the number of basic blocks as it uses the CCT tokenizer instead of conventional Vision Transformer tokenization, which makes it the right fit for small flood detection datasets. This model improves flood boundary delineation by involving local spatial patterns and global contextual relations. However, the method is based on Sentinel-1 SAR images and focuses on Erbil, Iraq, that experienced an extreme flash flood in December 2021. The experimental comparison results show that the proposed CCT-U-ViT outperforms multiple baseline models, such as conventional CNNs, U-Net, and Vision Transformer, obtaining an impressive overall accuracy of 91.24%. In addition, Furthermore, the model gets better precision and recall with F1-score of 91.21% and mIoU of 83.83%. Qualitative results demonstrate that CCT-U-ViT can effectively preserve the flood boundaries with higher precision and less salt-and-pepper noise compared with the state-of-the-art approaches. This study underscores the significance of hybrid deep learning models in enhancing the precision of flood detection with SAR data, providing valuable insights for the advancement of real-time flood monitoring and crisis management systems.

Keywords: deep learning; flood detection; SAR imaging; U-Net; transformer models

1. Introduction

Flood, and especially flash flood, is still one of the most harmful natural disasters resulting in death, population displacement and economic damage (Tanoue et al., 2020; Merz et al., 2021). But the impact is growing, due to climate change, urbanization and poor risk management. It is important to identify the location and spatial reach of floods in a timely and precise manner for effective disaster management, emergency response and flood mitigation purposes (Puttinaovarat & Horkaew, 2020). However, remote sensing is a crucial technique in the area of geoscience and particularly flood monitoring for it offers spatial information which is both large-scale and up-to-date (Ziboon et al. 2013, Ziboon et al. 2019, Shihab et al. 2020). In addition, SAR sensor systems (such as the ones on board of the Sentinel-1A) have unique capabilities for flood detection (Noori et al. 2024a).

However, SAR data provides advantages such as side-looking capabilities and the ability to collect surface information irrespective of weather and lighting conditions (Chini et al., 2021; Tsokas et al., 2022). Nonetheless, flood detection in urban settings presents difficulties due to significant double scattering between structures and neighboring ground surfaces (Li et al., 2019; Liu et al., 2024). In metropolitan regions, structures generate radar shadows and layover, obscuring substantial parts of the ground surface from Synthetic Aperture Radar (Mason et al., 2010; Tanguy et al., 2017). As a

result, numerous researches have focused on analyzing the discrepancies in backscatter between SAR pictures acquired prior to and after to floods (Schlaffer et al., 2015; Huang & Jin, 2022; Saleh et al., 2024a, 2024b, Noori et al. 2024c). To conclude, this research is based on the premise that double scattering caused by structures and adjacent floodwater in post-flood photos generally exceeds that from buildings and surrounding non-flooded terrain in pre-flood images (Wang et al., 2022).

In recent years, deep learning models have gained significant attention in remote sensing due to their capability to extract features and learn both local and global hierarchies autonomously. For flood detection with bi-temporal SAR images, deep learning techniques offer advantages like the ability to learn interactions between pre- and post-flood images. In contrast to traditional machine learning classifiers, deep learning methods provide enhanced feature representation that effectively addresses issues of speckle noise and complex land cover, particularly in SAR images. Furthermore, deep learning models can be tailored with modular computational components that specifically address the intricacies of the problem. Importantly, these models are often trained end-to-end, eliminating the need to address sub-problems (such as building shadows, layover, backscatter similarity in wetland, permanent water, and flooded regions) separately.

This research focuses on developing and rigorously evaluating a novel hybrid deep learning model that merges the strengths of the U-Net convolutional encoder-decoder structure with a Compact Convolutional Transformer (CCT) tokenizer. This combination allows for accurate and reliable flood detection using Sentinel-1 SAR imagery. The model is designed to effectively capture intricate spatial details and long-range global context, thus addressing the limitations of standalone CNN and Transformer frameworks. To achieve this, the model is trained and tested using benchmark datasets, including Sen1Floods11 and an additional dataset from the 2021 Erbil flood, facilitating a comprehensive performance assessment across various flood scenarios. The study aims to create a dual-path architecture that employs convolutional tokenization for efficient local-global feature integration, thereby boosting the model's classification accuracy and resilience to noise and variability often found in SAR data. It also benchmarks its performance against existing CNN, Transformer, and hybrid models, conducts ablation studies to evaluate the contribution of each component, and illustrates its potential for real-time flood monitoring and early warning systems.

This study's main contributions are summarized below:

1. The model combines a U-Net convolutional path for precise spatial and contextual feature extraction with a Transformer branch employing a CCT tokenizer to grasp sequential long-range dependencies and global context.
2. The model minimizes parameter needs and enhances generalization on small datasets commonly associated with flood detection by employing the CCT tokenizer instead of traditional Vision Transformer tokenization.
3. This innovative method merges global features derived from the convolutional decoder and the Transformer feature space, resulting in a thorough representation that enhances classification.

This work represents a significant advancement in remote sensing-based flood detection and monitoring. It tackles the limitations of current models through a novel architecture and tokenization strategy.

2. Related Works

Remote sensing-based flood mapping has been enriched by deep learning techniques including Convolutional Neural Networks (CNNs), U-Net architectures, Vision Transformers (ViTs), and hybrids which integrate these models. The development from old-style CNN networks to modern architectures is a kind of procedure consistently working for making local feature extraction and global context understanding into agreement.

2.1. CNN-Based Approaches

Convolutional Neural Networks (CNNs) have been long used as a building block in flood detection models because of their ability to extract local spatial features effectively using convolutional kernels that detect pixel neighborhoods and patterns (Wang et al., 2021, Eftekhari et al., 2023). Initial attempts, like patch-based CNN proposed by Aparna and Sudha (2022) also demonstrated showing flood detection on SAR imagery quite efficiently. Expanding on this concept, Sudiana et al. (2024) leveraged 3D CNNs to fuse the temporal information in the multitemporal SAR images and recovered the evolution of floods over time. Notwithstanding above achievements, a typical CNN architecture usually involves multi-stage pipelines, as demonstrated in Wang et al. (Jin et al., 2022), who firstly extracted water bodies and then conducted the change detection in temporal images. While this modular approach facilitates explainability, it is prone to segmentation artifacts (Sen et al., 2016; Sherrah et al., 2016), leading to suboptimal accuracy. Therefore, joint, one-stage approaches, where both segmentation and change detection are learned together, are increasingly favoured. An example of this is the work by Zagoruyko and Komodakis (2015), in which a model with a Siamese architecture is used to process the images of different time-steps and to extract a common feature representation to perform direct CD and prevent error propagation.

However, CNNs encounter the difficulties when applied to extensive flooding area since an increased receptive field results in high computational complexity. Moreover, they are only able to model short-range dependencies, making them unable to capture essential global semantic context information that is required for a practical semantic segmentation task (Doan & Le-Thi, 2025). To alleviate these challenges, spatial and channel attention mechanism has been introduced to CNNs to allocate computational resources to more relevant flood features. Huang et al. (2024) proposed WaterDetectionNet that incorporated the self-attention and the multiscale feature learning into an encoder-decoder architecture with the Xception backbone of DeepLabv3+, to facilitate a better flood mapping result. Similarly, Tahermanesh et al. (2025) by integrating spatial and channel attention mechanisms with the Inception v1 network and showing enhanced results for Sentinel-1 flood images under low data regime. These attention-augmented CNNs provide a potential direction of development with the penalty of computation effort. Beside flood detection, CNN architectures such as CE-Net (Gu et al., 2019) and dilated fully convolutional networks with active contour models (Hu et al., 2019) pushed the segmentation and boundary refinement in medical imaging. Their developments on context extraction and edge profiling have inspired similar progress on flood segmentation, suggesting the viability of introducing auxiliary modules with the intention to deal with the intrinsic limitation of the CNN.

2.2. U-Net and Its Variants

The U-Net is a well-known flood detection model because of its high-precision on pixel-level segmentation. With the pixel-level segmentation accuracy, U-Net has become a dominant flood detection model. Its encoder-decoder structure is symmetric with skip-connections to preserve spatial resolution and fine detail, which may otherwise be lost in typical CNN pooling-layers (Ronneberger et al., 2015). Such a feature reuse mechanism enables U-Net to better identify flood boundaries, which are critical to flood extent mapping. The state-of-the-art advancements of the original U-Net have improved boundary precision and multi-scale information. For instance, Qin et al. (2019) proposed BASNet which employs residual refinement modules and hybrid loss functions to enhance regional and boundary segmentation quality at the same time. Similarly, Bai et al. (2021) applied this method for the fusion of Sentinel-1 SAR with Sentinel-2 multi-spectral data, showing competitive mIoU scores greater than 50%, indicating the strength of cross-modal feature extraction. Moreover, U-Net++ (Zhou et al., 2018) further enhances the U-Net model by using nested dense skip connections which reduce the semantic gap between encoder and decoder features, leading to smoother transition of features and increasing the segmentation accuracy (Ghosh et al., 2024). This

recursive approach is particularly beneficial in the case of flood, where water accounts and shape can be confound.

In spite of these achievements, U-Net based architecture struggle in the presence of speckle noise and cluttered backgrounds often present in SAR images. Guo et al. (2021) proposed SA-UNet which contains spatial attention modules and structured dropout convolutional blocks for learning discriminating features as well as addressing overfitting on data-scarce cases. Similarly, Wang and Feng (2025) introduced multi-head attention into the U-Net architecture designed for SAR image characteristic that led to an incremental improvement of above 3% accuracy, precision compared to the baseline of U-Net in delineating the flood changes. It can be concluded from these attention-driven changes that (i) feature weighting should be emphasized when noise should be severe, and (ii) reducing noise is equally important for successful flood mapping. In addition, U2-Net (Qin et al., 2020) introduces a two-level nested U-structure with Residual U-blocks (RSUs) to gather multi-scale context information with affordable computation and proficiency, facilitating the training from scratch without the requirement of pretrained backbones. Tavus et al. (2022) further demonstrated the effectiveness of U-Net with ResNet50 backbones on Sentinel-1 flood data, with a uniform high F1-scores (approximately 0.82–0.83), which emphasized the robustness of the U-Net to different types of floods. Taken together, all these studies show a clear trend: although U-Net remains the backbone for flood segmentation, the introduction of attention and multi-scale fusion modules are necessary to adequately handle SAR image noise and complex flood morphologies.

2.3. Transformer Models

Transformers have drawn much attention towards flood detection, since they can model long range dependencies and the global context using self-attention (Vaswani et al., 2017). Vision Transformers (ViTs) Park et al. (2021); Touvron et al. (2021) and their hierarchical counterparts e.g., Swin Transformer Liu et al. (2021) address the fixed receptive fields left by the CNNs, therefore supporting a more comprehensive exploration into large flood events. A study by Chamatidis et al. (2024) demonstrated that ViT models combined with transfer learning on Sentinel-1 SAR and Sentinel-2 multispectral imagery, can outperform traditional CNN model accuracy by up to 15%. This demonstrates the potential role of transformers in recognizing global patterns that are important to understand for mitigating flooding risks, and for enabling emergency responses. However, ViTs typically need large-scale training data and struggle to model subtle local patterns due to the absent of convolutional inductive priors (Sharma & Saharia, 2024; Zhou et al., 2025).

To overcome these limitations, Zhou et al. (2025) proposed a hybrid transformer model, which is a mixture of a mixer transformer encoder with noise filtering and multiscale depth-wise convolution blocks. This architecture properly decouples the global context modeling and local spatial detail preserving, which tackles problems such as flood-background similarity and edge discontinuity, leading to the best performance on overall benchmark datasets. Other methods include CSWin-UNet (Liu et al., 2025) which uses cross shaped window self-attention and CASCADE upsampling to improve segmentation quality and computational complexity. AgileFormer (Qiu et al., 2024) improves the flexibility by incorporating deformable patch embeddings and spatially varied self-attention so that the irregular shaped flood regions can be accurately segmented, which is one of common problems in different shapes of flood landscapes. Furthermore, Fan et al. (2022) that successfully leveraged Swin Transformer layers in a U-Net pipeline for image restoration, showing that transformers have the potential of advancing high-resolution vision tasks, such as fine-grained flood mapping.

2.4. Hybrid CNN-Transformer Models

Recognizing that neither CNNs nor transformers alone sufficiently tackle flood detection issues, hybrid architectures merging their complementary advantages have become the state of the art. These models effectively combine CNNs' local spatial feature extraction with transformers' global context

modeling to capture intricate details and extensive dependencies. Recently, TransUNet (Chen et al., 2021) takes this integration to a new level by substituting the traditional U-Net encoder with a ViT module, which leverages CNNs to encode low-level local details and transformers to capture long-distance context information. Cao et al. (2022) extended this idea by creating SwinUNet, a transformer-only version of U-Net, introducing shifted window attention to improve computational efficiency while maintaining the segmentation performance. Wang et al. (2023) leverages a CNN-transformer hybrid encoder for U-Net (Ronneberger et al. 2015) architecture in combination with a non-linear double upsampling decoder to enhance the feature extraction and generalization for complex scenes, a task particularly relevant to flood segmentation. Yang et al. (2024) introduced the CvT-UNet which combines convolutional projections with multihead self-attention blocks, effectively balancing spatial localization and global context for an accurate weld pool segmentation, as of potential relevance also for flood boundary detection.

In contrast to transformer hybrids, which have strong demand for computation and pretraining resources (Zhang & Zhang, 2024), Zhang et al. (2025) tackled this problem with the development of FET-UNet, which includes the CNN (ResNet34) and Swin Transformer branches by high-level feature fusion and multi-scale upsampling and achieved more excellent results on the task of ultrasound segmentation—a domain with imaging difficulties similar to flood detection. Lightweight hybrid design for UNet models such as those in UNetFormer (Wang et al., 2022) with global-local attention mechanism made it feasible to provide robust accuracy in the real-time semantic segmentation which is required by flood monitoring. Furthermore, Sha et al. (2021) showed that a simple concatenation of transformer and U-Net (Transformer-UNet) outperforms the depth-wise U-Net models, with some trade-off in efficiency w.r.t. the depth of the backbone. Moreover, Doan and Le-Thi (2025) proposed a Siamese network using Swin-Transformer (SwinTrans) based on hierarchical feature extractors, the main power is computational efficiency that is accompanied with spatial connectivity to SAR flood image detection. Zhou et al. (2024) proposed the ViT-UNet for high-resolution coastal wetland classification, replacing convolutions with Vision Transformer blocks and introducing dual skip connections and the bilinear polymerization pooling to improve feature fusion, increasing the precision of the original U-Net by more than 4%. Together, these hybrid methodologies represent an evolving research trajectory that effectively combines the strengths of CNNs and transformers, tackling the challenges of flood detection—noise, scale variance, and complex boundaries—with enhanced accuracy and practical viability.

3. Study Area and Flood Detection Datasets

3.1. Description of the Study Area

Erbil, located in northern Iraq, serves as the focus of this study. Its geographic coordinates are 36°11'28"N and 44°0'33"E, as illustrated in Figure 1. The region features a broad plain, interspersed with hills to the east that reach elevations up to 426 meters above sea level (Al-Hameedawi, 2014). The landscape is predominantly covered by Quaternary sediments that have accumulated due to the weathering and erosion of the neighboring highlands. Significantly, in the northern and northeastern parts of the study area, the Quaternary sediments lie atop the Bai Hassan formation, composed of molasses-type rock formations. The central section of Erbil is generally flat, whereas the northeastern and eastern areas present a more rugged landscape (Ahmed et al., 2023).

Erbil's climate is classified as semiarid, exhibiting apparent seasonal shifts in humidity. Summers are marked by low humidity, with temperatures often rising above 45 °C, whereas winters experience moderate humidity, with temperatures frequently falling below 0 °C. The area encounters a cool and wet climate, receiving over 400 mm of average annual precipitation (Ali & Mawlood, 2023). Rainfall usually starts in mid-October and lasts until May.

On December 17, 2021, Erbil and the Kurdistan Region of Iraq faced an unusually severe rainfall following one of the driest years in recent history (Noori et al. 2024b). Starting at 4 a.m., the rain led to extensive flooding in several districts of Erbil, such as Dara Too, Qush Tappa, Shamamk, Zhyan,

Roshinbiri, and Bahrka, especially in the northern and eastern areas of the city. This heavy downpour caused substantial destruction, damaging homes, buildings, and vehicles, and tragically resulting in loss of life (Sissakian et al., 2022).

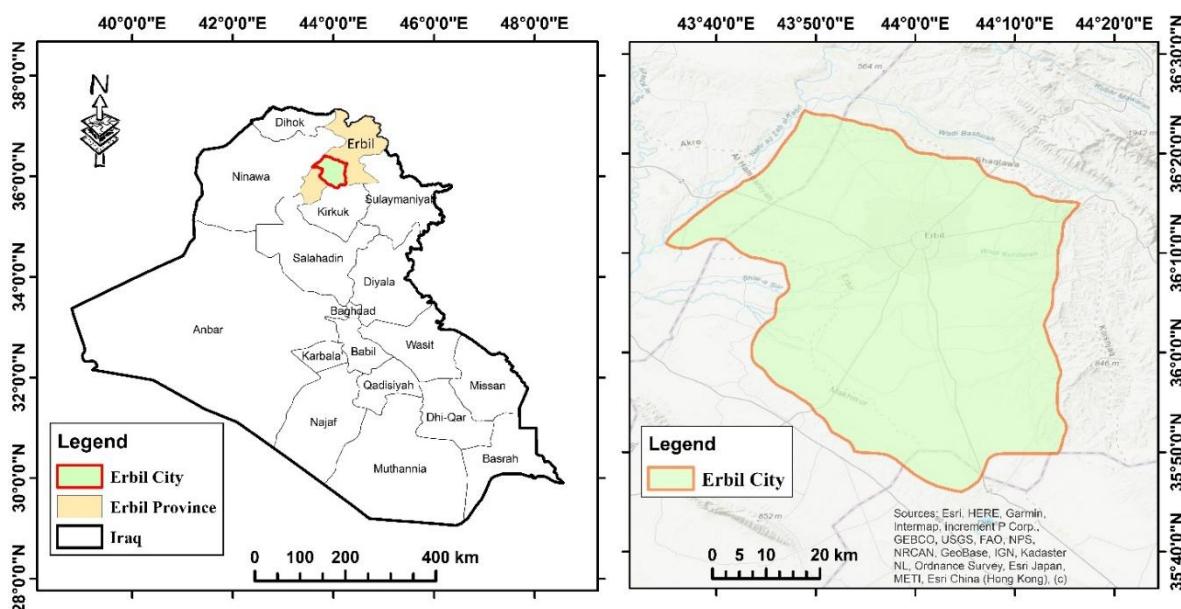


Figure 1. Map of the study area.

3.2. Flood Detection Datasets

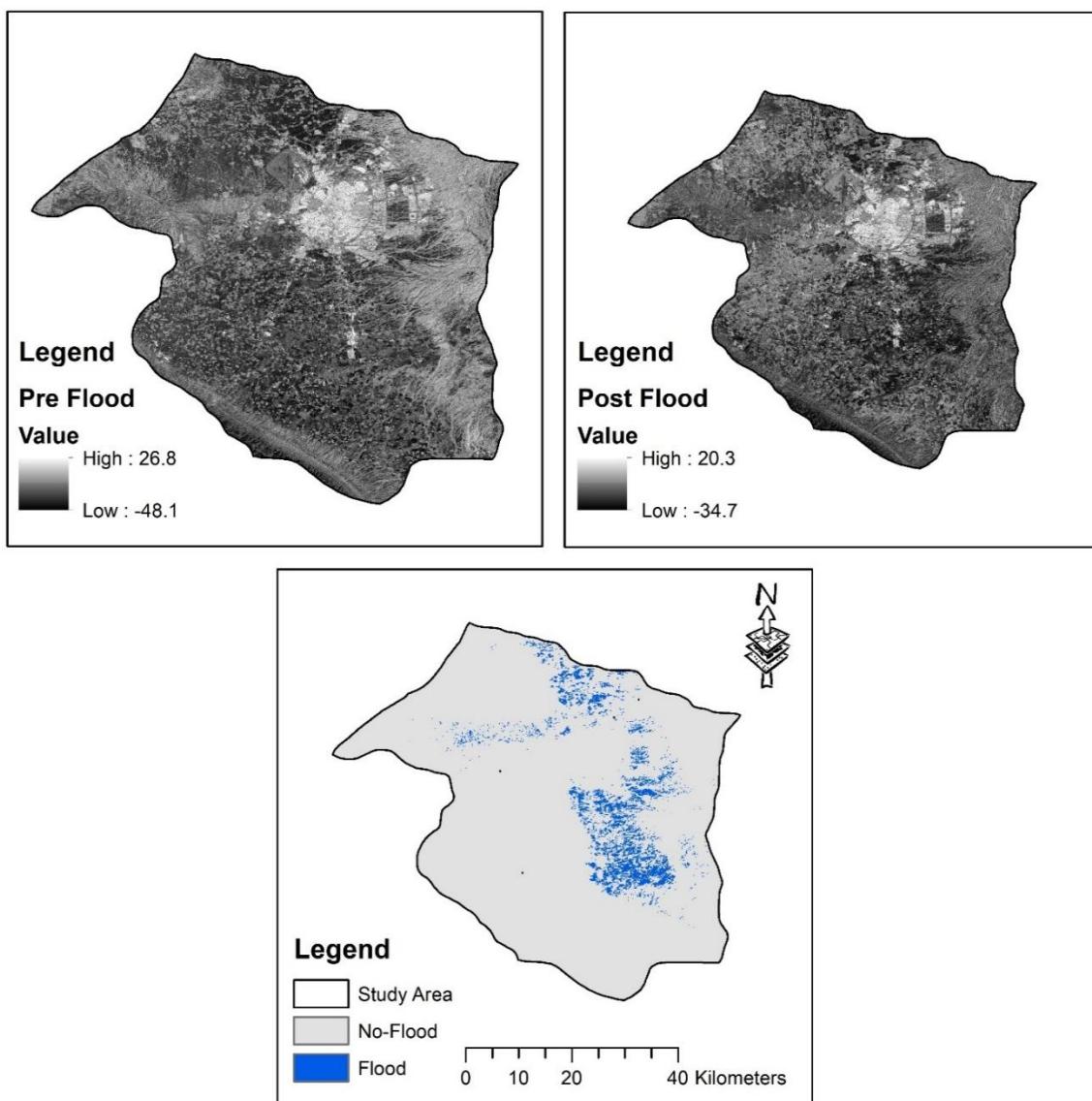
3.2.1. Erbil Flood Dataset

The satellite images from Sentinel-1 mission were employed in the current research for flood delineation. Based on the Copernicus program of the European Space Agency, the Sentinel-1 offers high resolution SAR images and has been including the possibility of using it for flood monitoring. Differing from the optical images, SAR is insensitive to cloud or illumination, and is suitable for the long period monitoring of the flood (Zhao et al., 2024). The study location is Erbil City due to flash flood that hit the city in December 2021 which resulted in remarkable losses in lives and properties. In the present study the flood impinged land surface alteration maps were obtained from the pre- and post-flood SAR scenes (Figure 2). These images were sourced through Google Earth Engine (GEE), a cloud computing platform that specializes in processing geospatial data sets (Vijayakumar, Privacy enabled remote sensing imagery processing and analysis at scale using Google Earth Engine, 2024).

Table 1 presents a summary on the pre- and post-flood preparation process for Sentinel-1 images for assessing the impact of the flash flood. The acquisitions of the data were collected in the Interferometric Wide Swath (IW) beam mode, which provides wide swath at medium resolution and is considered appropriate for the large-scale disaster monitoring. The high resolution (HR) data (10m \times 10m spatial resolution) can ensure the precision monitoring of waters, land sink and infrastructure damages. This method, with a swath-width of 250 km per swath, can scale large regions with flooding disaster.

Table 1. Satellite data from Sentinel-1 used for flood analysis.

Satellite	Acquisition date	Processing level	Polarization	Spatial resolution (meters)
Sentinel-1	Pre-flood 25-09-2021	Level 1	Single - VH	10
Sentinel-1	Post-flood 20-12-2021	Level 1	Single - VH	10

**Figure 2.** Maps of the pre-flood, post-flood, and flood ground truth datasets.

The raw radar data from Sentinel-1 goes through a pre-processing phase, which involves standard procedures to prepare for further analysis. The first step is processing this raw data, including radiometric calibration to adjust for sensor and atmospheric influences, ensuring that the signal strength accurately reflects surface characteristics. Afterward, geometric correction aligns the data with the Earth's surface, compensating for distortions caused by the satellite's movement and the curvature of the Earth. Speckle filtering is applied to reduce noise from the radar signal interacting with the surface, enhancing the visibility of features such as bodies of water or buildings. Speckle noise commonly affects SAR images, complicating visual analysis. Lee (5x5) speckle filtering is implemented to mitigate this noise while preserving image details. Additionally, orthorectification corrects distortions caused by terrain variations, ensuring that the resulting images accurately

represent the Earth's surface. The data may also be subsetted based on specific use cases, such as masking flooded areas for targeted data analysis by researchers. This pre-processing is crucial for improving the quality and applicability of satellite data in monitoring events like flash floods. All preprocessing steps of the Sentinel-1 dataset were implemented using JavaScript in GEE.

3.2.2. S1GFloods Dataset

The benchmarking data of our model consists of the data in Saleh et al. (2024), which used the Sentinel-1 SAR databased satellite images. This radar imaging system, designed by the European Space Agency (ESA), is capable of creating high-resolution images regardless of weather, time of day or atmosphere opacity. The dataset covers common and high-impact cause of flooding such as heavy rainfall, riverine flooding, dams and levee failure, tropical storms and hurricanes. Its geographic variation ensures the adaptability of the flood monitoring method to a wide range of environmental scenarios, such as rural, mountainous, urban communities, vegetated zones, rivers, ponds, lakes and reservoirs.

The dataset contains 4,830 image sets consisting with pre-flood, post-flood and change labeled images. Each image is of size 256×256 pixels and has 3 channels that represent the Red (R), Green (G), and Blue (B) in RGB images. The labels are provided as a single channel; each pixel is assigned as flood (0) or non-flood (1). In order to better train the flood detection model, the dataset is divided into training set, which has 4,300 image sets (90% in total), and test set, which consists of 530 image sets (10% in total). Training sets are further divided into two smaller training subset and validation subset to train the model with the right parameters. A model which has this division allows a good generalization that reduce the chance of overfitting and enhances the learning.

4. Methodology

4.1. Network Architecture Overview

This research introduced the Compact Convolutional Tokenizer-based Hybrid U-Net and Vision Transformer Model (CCT-U-ViT) (Figure 3). This innovative model features a hybrid architecture designed for Synthetic Aperture Radar (SAR)-based flood detection, integrating U-Net's spatial feature extraction with the Transformer's global context modeling. The system processes pre- and post-flood SAR images utilizing a U-Net encoder-decoder framework with skip connections; the encoder methodically extracts hierarchical features using 8, 16, 32, and 64 filters across four levels. Each encoder level corresponds to a decoder layer, facilitating the reconstruction of spatial details. At the bottleneck layer, equipped with 128 filters, the model employs three vital enhancement modules: a Compact Convolutional Transformer (CCT) tokenizer that transforms 2D pooled features into 1D tokens for sequential processing, position embedding for adding spatial awareness (a dense 128-dimensional structure), and transformer blocks that implement multi-head attention mechanisms along with MLP and skip connections to capture long-range dependencies. The U-Net's spatial features are merged with the global features derived from the transformer via a feature fusion module, which concatenates, densifies (to 64 channels), and applies a dropout rate 0.2 for regularization. Ultimately, a binary classification head with dense layers differentiates between flood and non-flood pixels, enabling the model to utilize local spatial patterns through the U-Net path and global contextual relationships via the transformer path to delineate flood boundaries precisely.

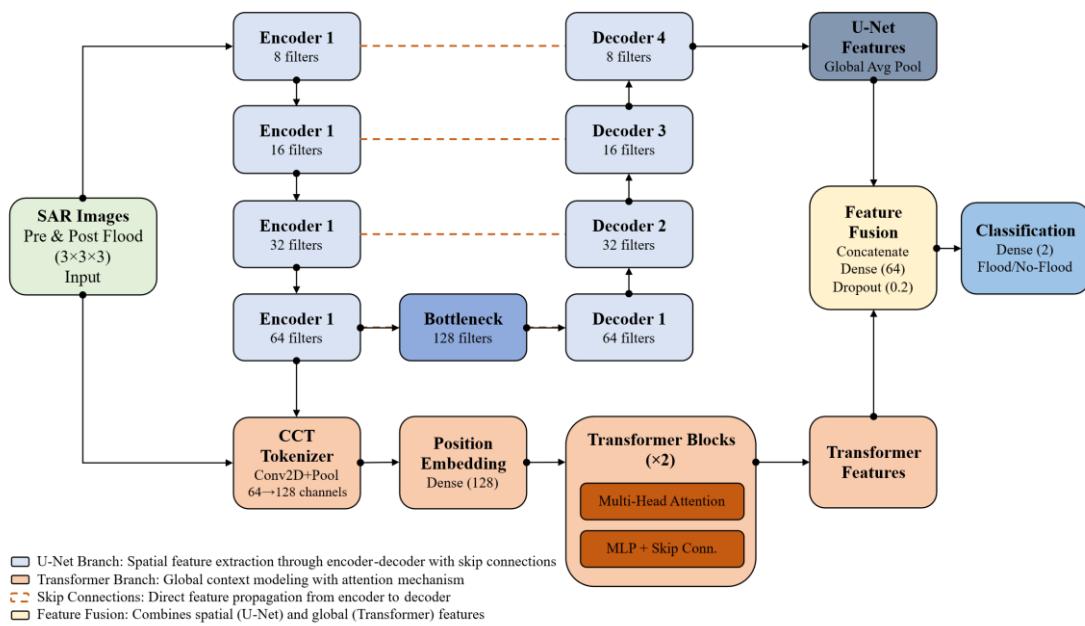


Figure 3. Architecture of the proposed CCT-U-ViT model.

4.2. U-Net Architecture

The U-Net module utilizes an encoder-decoder architecture frequently used in segmentation tasks. The encoder path extracts hierarchical features from the input image, while the decoder reconstructs the image or generates high-level representations. Comprising multiple blocks, the encoder contains convolutional layers, which are followed by batch normalization and ReLU activation. As the blocks progress, they reduce the spatial resolution of the feature maps, capturing more abstract representations. At the end of each encoder block, max pooling layers downsample the feature maps. Conversely, the decoder path employs transposed convolutions to enhance the resolution of feature maps, along with concatenation with corresponding encoder features via skip connections. This methodology enables the model to utilize fine-grained spatial information from earlier layers, improving prediction accuracy. In both the encoder and decoder paths, each convolutional block features two convolutional layers with 3x3 kernels, followed by batch normalization and ReLU activation. Ultimately, the U-Net decoder path produces a high-level feature representation of the input image.

4.2.1. U-Net Encoder with Hierarchical Convolutional Blocks

The encoder path comprises a series of convolutional blocks, each containing convolutional layers that utilize batch normalization and ReLU activation (Figure 4). As these blocks proceed, they incrementally increase the filter count while decreasing spatial dimensions using max-pooling, which helps extract increasingly abstract hierarchical features. Skip connections from each encoder stage preserve high-resolution spatial details for subsequent reconstruction.

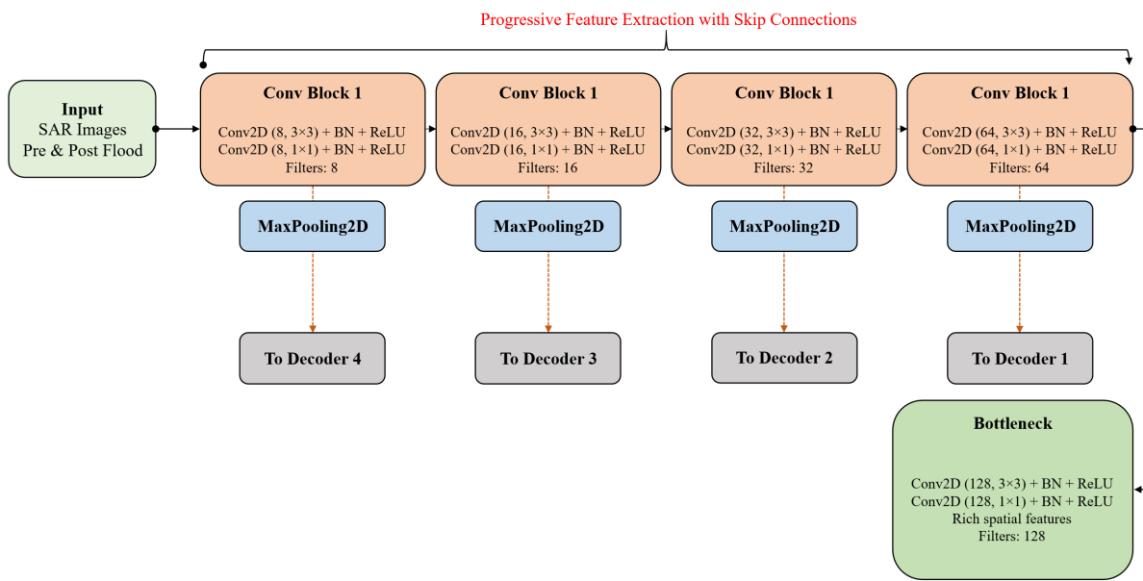


Figure 4. Architecture of the U-Net Encoder model.

4.2.2. U-Net Decoder with Transposed Convolutions and Skip Connections

The decoder reflects the encoder's architecture but uses transposed convolutional layers for upsampling (Figure 5). In every decoder phase, features from the related encoder block are concatenated via skip connections, restoring fine-grained spatial details diminished during downsampling. Additional convolutional blocks enhance the upsampled feature maps, leading to better spatial reconstruction.

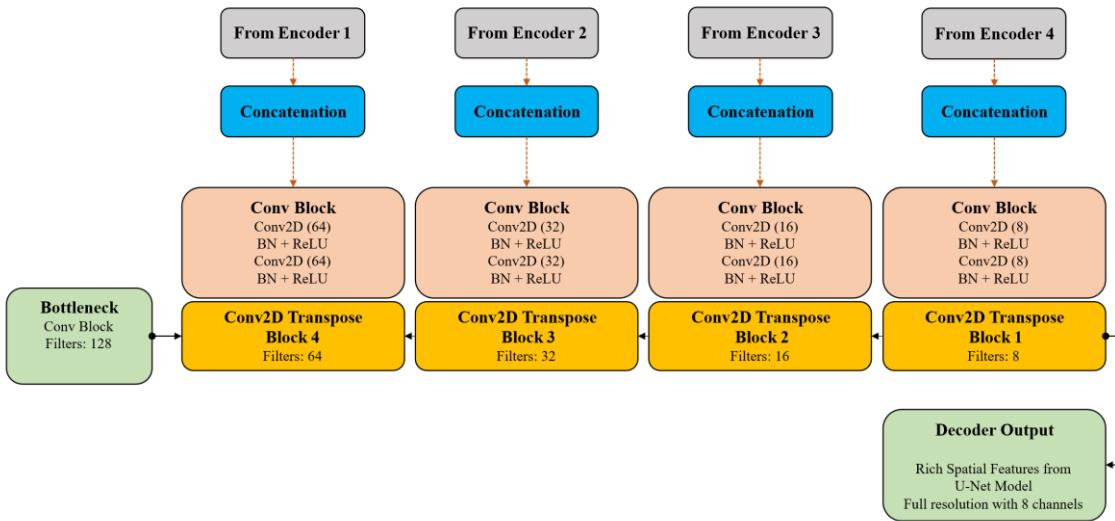


Figure 5. Architecture of the U-Net Decoder model.

4.3. CNN-Based Tokenizer for Transformer Input (CCTTokenizer)

Rather than employing standard patch extraction, the transformer branch uses a CNN-based tokenizer. This approach yields richer and more informative tokens by learning spatial features before sending sequences into the transformer, thus improving the quality of tokens for subsequent attention mechanisms. The CCTTokenizer layer acts as the CNN tokenizer, converting the image into

a sequence of patches through a succession of convolutional and pooling layers. Initially, the convolutional layers, which use 3x3 kernels, capture low-level features, followed by max-pooling layers that decrease spatial dimensions. Finally, the output from this segment is reshaped into a sequence of tokens for processing by the Transformer. The architecture of this module is presented in Figure 6.

Spatial information is integrated into the tokenized patches by applying positional embeddings to the CNN tokenizer's output. These embeddings are trained and then merged with the tokenized patches. This allows the Transformer to understand the relative positions of patches in the image.

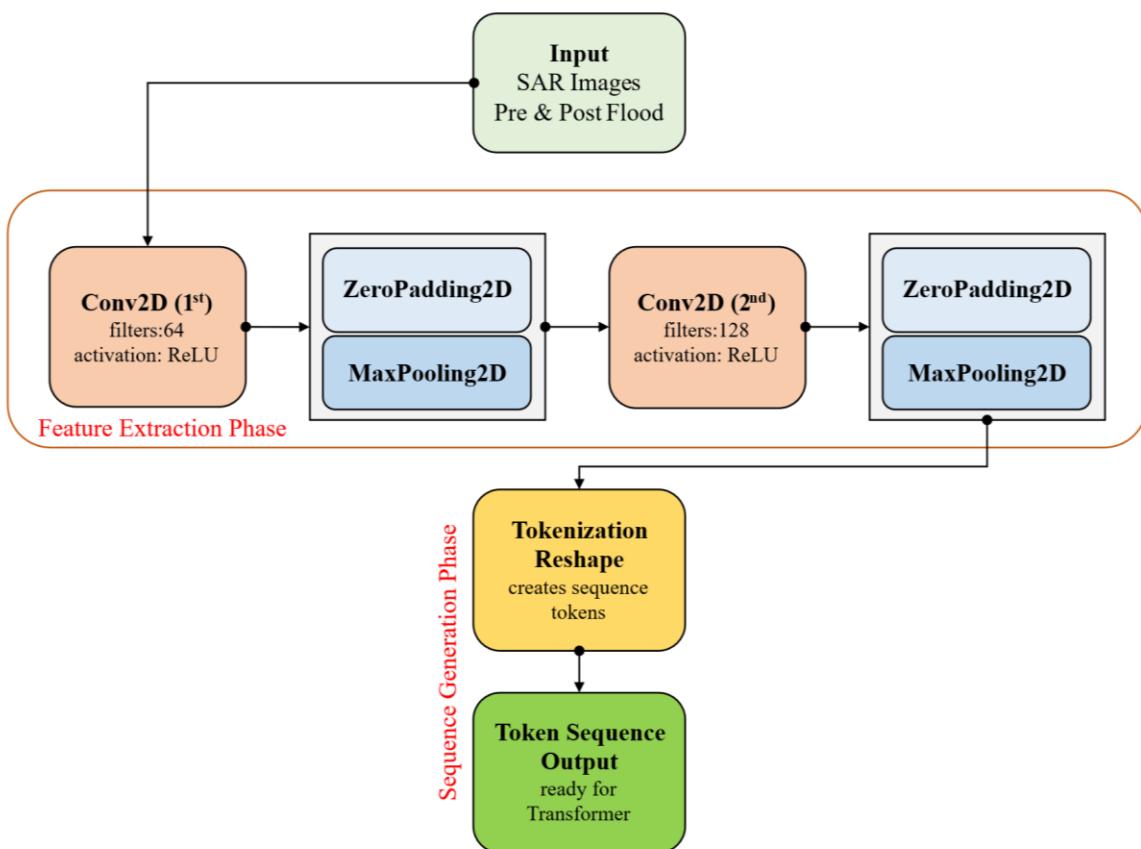


Figure 6. Architecture of the CCT Tokenizer Module.

4.4. Transformer Blocks

The transformer branch includes several encoder layers, each incorporating layer normalization, multi-head self-attention with adjustable head counts and projection sizes, and position-wise feed-forward MLPs that utilize GELU activations and dropout for regularization (Figure 7). Residual skip connections between the layers improve gradient flow and support stable training. Learned positional embeddings are combined with token embeddings to offer spatial context.

Each Transformer layer incorporates a Multi-Head Attention mechanism that operates on the tokenized patches to capture long-range dependencies between image regions. Num_heads specifies the number of attention heads, while the attention space's dimensionality is indicated by projection_dim. After each attention operation, a residual connection is included in the output to aid training. This is followed by a feedforward neural network that consists of two dense layers utilizing GELU activations and dropout for regularization. Layer normalization is applied after both the attention and feedforward layers to enhance training stability. The Transformer block's output is improved through multiple layers, enabling the model to capture local and global contextual information.

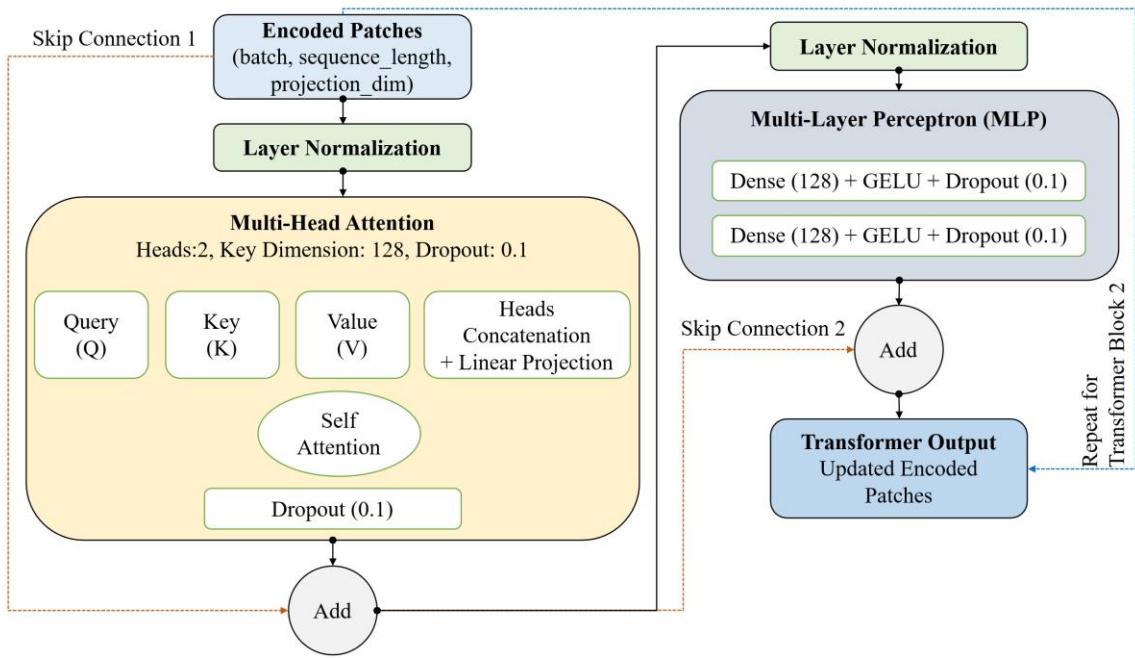


Figure 7. Architecture of the Transformer branch in CCT-U-ViT.

4.5. Feature Fusion and Classification Head

After processing independently, the global average pooled features from the U-Net decoder are concatenated with the transformer outputs to form a detailed feature vector that captures local spatial and global contextual information. This combined vector is passed through fully connected layers with dropout regularization and ReLU activation, leading to a final dense layer that yields logits for binary classification. A fully connected layer featuring 64 units and ReLU activation is applied to this integrated feature vector. To avoid overfitting, dropout regularization is incorporated with a dropout rate of 0.2. Ultimately, the final output is produced by a dense layer with two units (representing binary classification) and no activation function, intended for use with a binary cross-entropy loss function. Figure 8 presents this feature fusion and classification head module.

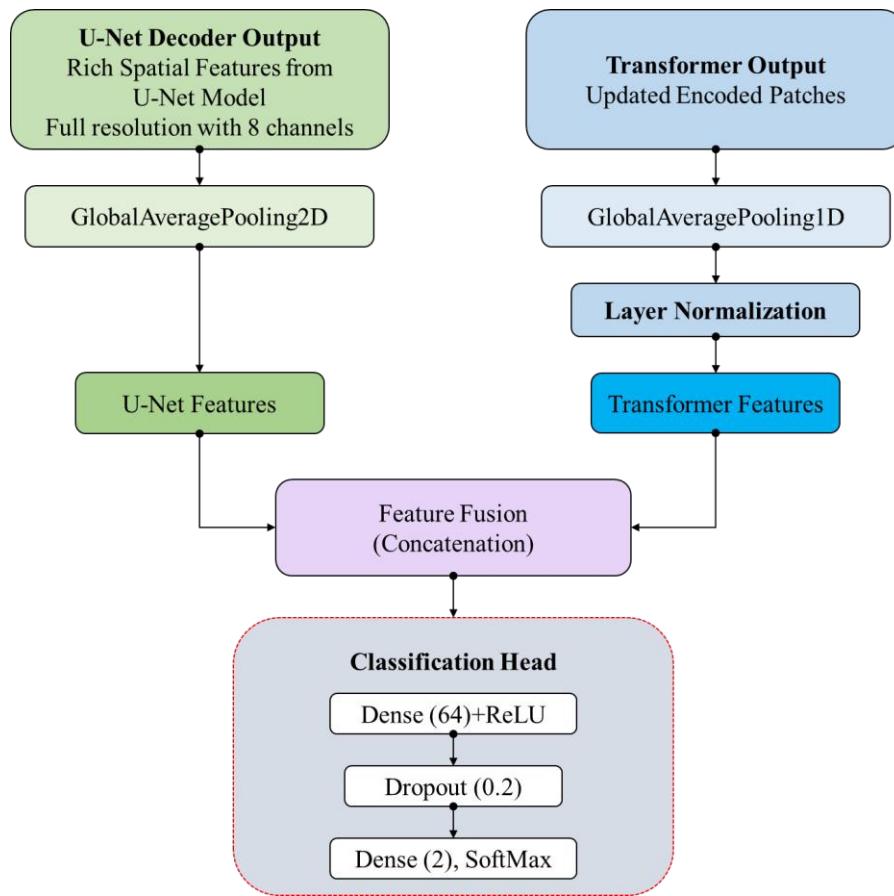


Figure 8. Architecture of the feature fusion and classification head of the proposed model.

4.6. Benchmark Models

4.6.1. CNN 2D – 1 Layer

This model employs a sequential convolutional neural network architecture, starting with a 2D convolutional layer with 32 filters and a 2×2 -pixel kernel size. It uses Rectified Linear Unit (ReLU) activation to introduce non-linearity, processing input tensors of shape $(3, 3, 2)$. The feature maps are then flattened into a one-dimensional vector, connecting to a fully connected layer with 16 hidden units that also use ReLU activation. The network ends with a dense layer of 2 units, applying Softmax activation to generate probability distributions for target classes. This design allows the network to capture hierarchical features through convolution while keeping a lightweight parameter count due to its limited filters and hidden units.

4.6.2. CNN 2D – 2 Layers

This model is a sequential convolutional neural network with two layers. The first layer has 64 filters and a 2×2 kernel size, using ReLU activation on input tensors of shape $(3, 3, 2)$. The second layer comprises 32 filters and a 1×1 kernel with ReLU activation. This architecture allows for hierarchical feature extraction, as the first layer captures local patterns while the second refines features. After the convolutional layers, feature maps are flattened into a one-dimensional vector and processed through a fully connected layer with 16 hidden units and ReLU activation. The network concludes with a dense output layer featuring two units and Softmax activation for probability distributions across target classes.

4.6.3. CNN 3D – 1 Layer

3DCNNs model The model is based on 3D convolutional neural networks taking (3,3,2) as input tensor size and expanded to (3, 3, 2, 1) with the 3D convolutional neural network that can handle batch, spatial input data. It begins with a 3D version of convolutional layer, containing 32 $2 \times 2 \times 1$ pixel filters to perform 2D convolutions through the input channels while preserving channel dimension. However, non-linearity is introduced with ReLU activation functions. The output is feature map set reshaped to 1D-vector that flows through a fully connected layer with 16 ReLU hidden units. The network also has a final dense output layer with 2 units and Softmax activation that provides probability distributions of the target classes.

4.6.4. CNN 3D – 2 Layers

The model is built on the Keras framework and utilises a two-layer 3D convolutional neural network. It takes a tensor input with size $[p \times p \times \text{channels}]$ and produces a reshaped output of size $[p \times p \times \text{channels} \times 1]$. The first layer has 64 $2 \times 2 \times 1$ pixel kernel filters and the second layer has 32 $2 \times 2 \times 1$ pixel kernel filters, both layers have used ReLU activation. In this structure, the hierarchical feature extraction process is enabled, at the same time, the channel dimension is preserved. The feature maps from the second layer are flattened to form a 1D vector, which is fed into a fully connected layer of 16 hidden units with ReLU activation. The network is topped off with a dense output layer having [output_shape] nodes with Softmax activation, used to generate probabilities across all of the target classes.

4.6.5. Hybrid CNN

Model architecture: Our merge uses a hybrid CNN (both 3D and 2D convolutions) following the Keras functional API. Input tensors of dimensions $[p \times p \times \text{channels}]$ are reshaped to $[p \times p \times \text{channels} \times 1]$. It is composed of two 3D convolutional layers, the first with 64 filters (kernel size $3 \times 3 \times 2$) and the second with 32 filters (kernel size $3 \times 3 \times 1$), all using ReLU activation. After these layers, feature maps are reshaped into a single dimension by combining the dimensions of filter and feature-map channels along which we will perform 2D convolutions. A 2D convolutional layer using 16 filters (3×3 kernel size) and the rectified linear activation is then applied. This architecture is designed such that the network is able to learn spatiotemporal patterns through 3D convolutions and treat fused features using 2D convolutions. Finally, the feature maps are flattened to a one-dimensional vector and go through a fully connected layer with 16 hidden units and ReLU activation. A dropout layer with 0.5 rate for dropout is used for avoiding the overfitting before the output layer which is having [output_shape] units with Softmax activation, and returns the probability values of classes.

4.6.6. U-Net

The model architecture uses a modified U-Net CNN with an encoder-decoder structure and skip connections via the Keras functional API. It has three components: the encoder path, a bottleneck, and a decoder path, each using specialized convolutional blocks. However, these blocks perform a 3×3 convolution followed by batch normalization, ReLU activation, and a 1×1 convolution with the same normalization and activation. This dual-convolution technique supports spatial feature extraction and channel refinement. The encoder path comprises four blocks, combining a convolutional block with 1×1 max pooling, with filter counts increasing (8, 16, 32, 64) for hierarchical feature extraction. Each encoder block retains features before pooling through skip connections for the decoder. The bottleneck contains a convolutional block with 128 filters, linking the encoder and decoder while managing abstract features. The decoder mirrors the encoder with four blocks, starting with a 1×1 transposed convolution (upsampling) and concatenating corresponding skip features, merging low-level spatial data with high-level semantics. Filter counts decrease (64, 32, 16, 8), reconstructing spatial resolution and reducing feature complexity. After the decoder, global average pooling condenses dimensions to 1×1 , summarizing information across the feature map. The network

ends with a dense output layer with [output_shape] units and sigmoid activation for binary classification.

4.6.7. Vision Transformer (ViT)

The ViT model is a patch-based technique that processes input tensors, utilizing a transformer architecture for feature extraction and classification. The Patches layer segments the input image (3×3) into non-overlapping 1×1 patches, which are projected into a higher-dimensional space (projection_dim=64) via the PatchEncoder layer with learnable position embeddings. The transformer backbone includes four layers with: a multi-head self-attention mechanism (4 heads), layer normalization (epsilon=1e-6) with residual connections, and a multilayer perceptron (MLP) with GELU activation, consisting of two dense layers sized [128, 64] and a dropout rate of 0.1. Following the transformer layers, the architecture normalizes the final encoded representation, flattens spatial dimensions, applies dropout regularization at 0.5, and includes a final MLP head with two dense layers sized [64, 32] and a dropout rate of 0.5, finishing with an output layer using Softmax activation for classification. The model uses the Adam optimizer with a learning rate 0.001 and weight decay of 0.0001. Binary cross-entropy is the loss function, and performance is evaluated using binary accuracy metrics. Data is processed in batches of 32 samples.

This Vision Transformer architecture preserves the original ViT principles, tailored for smaller images by adjusting patch sizes and hyperparameters. Dropout regularization (0.1 in transformer layers and 0.5 in final layers) and layer normalization mitigate overfitting, while multi-head attention helps the model understand complex spatial relationships.

4.6.8. CNN-Transformer

The model's architecture integrates a hybrid Convolutional Neural Network (CNN) with a Transformer framework, optimized for effective visual data processing. Initially, input images are processed through a custom CCTTokenizer layer that performs convolutional operations and pooling to extract spatial features, reshaping the data into token sequences that the transformer can process. Positional embeddings can be included to represent the spatial relationships among the image patches. The model's core comprises stacked transformer layers equipped with multi-head self-attention, layer normalization, stochastic depth for regularization, and a feedforward network. Following the transformer layers, a SequencePooling layer utilizes soft attention on the token sequences to generate a pooled representation, which is forwarded to a fully connected output layer to facilitate either classification or regression. Various regularization techniques, such as dropout and stochastic depth, are implemented to reduce overfitting, while the model is optimized using binary cross-entropy loss and label smoothing for classification purposes.

4.7. Performance Metrics

A set of common evaluation metrics are adopted to evaluate the performance of deep learning models according to their accuracy of detecting floods in SAR images. We calculate these measures to investigate how well the model distinguishes between the flood and non-flood pixels. The performance measures used in this study are:

Accuracy measures the proportion of correctly classified pixels, including flood and non-flood categories, compared to the total pixel count in the dataset. It is defined as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

where:

- TP = True Positives (correctly detected flood pixels)
- TN = True Negatives (correctly detected non-flood pixels)
- FP = False Positives (non-flood pixels classified as flood)

- FN = False Negatives (flood pixels classified as non-flood)

Cohen's Kappa coefficient evaluates the alignment between predicted and actual classifications, factoring in chance agreement. Its calculation is as follows:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \quad (2)$$

where:

- P_o = Observed agreement (the actual accuracy of the model)
- P_e = Expected agreement by chance (the agreement expected if the model were to classify randomly)

A higher Kappa value signifies improved agreement beyond what would be anticipated by chance.

The F1 Score represents the harmonic mean of Precision and Recall. This metric is particularly valuable for handling imbalanced datasets because it balances the trade-off between false positives and false negatives:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN} \quad (3)$$

where:

- Precision is the proportion of detected flood pixels that are correctly classified as flood
- Recall is the proportion of actual flood pixels that are correctly detected

The mean Intersection over Union (mIoU) metric measures the overlap between predicted and actual flood areas. It represents the average Intersection over Union for both flood and non-flood classes:

$$mIoU = \frac{1}{2} \left(\frac{TP}{TP + FP + FN} + \frac{TN}{TN + FN + FP} \right) \quad (4)$$

This metric evaluates the model's ability to correctly distinguish between flood and non-flood areas by determining the ratio of the intersection of predicted and actual regions to their union.

5. Results and Discussions

5.1. Experimental Setup and Parameter Settings

The trials were performed on a high-performance computing workstation equipped with an Intel architecture CPU including 24 physical and 32 logical cores, together with 32 GB of RAM, which exhibited a 69.0% usage during training. However, the NVIDIA GeForce RTX 4070 Ti GPU enhanced this system's computing power, providing 28.3 GB of VRAM and facilitating mixed-precision training with the Ampere architecture, essential for managing large flood detection datasets and complex model architectures. Furthermore, The deep learning framework was developed with Python 3.10, utilizing TensorFlow 2.9.1 and the Keras API, chosen for its efficiency in GPU resource utilization and suitability for computer vision applications. The model's architecture employed Sentinel-1 SAR images, utilizing 3x3 spatial patches to facilitate temporal comparisons of circumstances pre- and post-flood, hence enabling flood detection via a binary classification approach.

For training, the Adam optimizer was employed with standard adaptive learning rates and key parameters set as Beta1 = 0.9, Beta2 = 0.999, and Epsilon = 1e-7, which aids in achieving stable convergence even amidst noisy data. We selected the binary cross-entropy loss function to ensure appropriate gradient signals for binary classification tasks. The dataset contained 54,733,367 labeled pixels, highlighting a significant class imbalance (with a 1:31 ratio of flood to non-flood pixels) and was split into 70% for training, 10% for validation, and 20% for testing. Data preprocessing involved normalizing pixel values and extracting 3x3 patches from co-registered SAR imagery, explicitly avoiding data augmentation to preserve the integrity of radar backscatter characteristics. The model

underwent training for 100 epochs with a batch size of 64, and it was evaluated using various metrics, including Overall Accuracy (OA), Kappa Coefficient, F1-Score, and mean Intersection over Union (mIoU), to ensure a comprehensive performance assessment despite the existing class imbalance. Additionally, memory management techniques were implemented to optimize GPU memory usage, with batch processing tailored for the RTX 4070 Ti, and random seed initialization was performed to ensure the reproducibility of results.

$$\begin{aligned} BCELoss(x) &= -1 \times [y \cdot \ln(x) + (1-y) \cdot \ln(1-x)] \\ &= \begin{cases} -\ln(1-x), & y = 0 \\ -\ln x, & y = 1. \end{cases} \quad (5) \end{aligned}$$

5.2. Results of Erbil Flood Detection

The evaluation results in Table 2 highlight the superior performance of the proposed CCT-U-ViT architecture across all established evaluation metrics, providing empirical support for the model's effectiveness in flood detection tasks. The CCT-U-ViT model attained the highest overall accuracy (OA) of 91.24%, marking a 1.45% increase over the second-best model, CNN-Transformer, which scored 90.79%. This improvement holds significant practical relevance due to the extensive scope of satellite imagery analysis. The performance boost is attributed to the effective integration of convolutional compact transformers with the U-Net architecture, allowing the model to capture local spatial dependencies through convolutional operations and long-range contextual relationships using self-attention mechanisms. With a Kappa coefficient of 0.8248, the CCT-U-ViT demonstrates excellent inter-rater reliability. Its performance greatly surpasses chance agreement, showcasing a statistically significant improvement over baseline methods for flood detection applications.

The comparative analysis reveals apparent performance differences among various architectural paradigms. Traditional 2D CNN methods (90.41-90.56% OA) perform well due to their ability to harness spatial correlations and the translation invariance properties inherent in satellite imagery. On the other hand, 3D CNN models show comparatively lower performance (87.41-88.92% OA) because of their increased complexity and the risk of overfitting when temporal dimensions are added without sufficient training data. The ViT achieves competitive results (90.64% OA) by utilizing global attention mechanisms, but the quadratic computational complexity of self-attention and limited inductive biases for spatial data hinders its performance. The U-Net architecture (89.53% OA), effective for general segmentation tasks thanks to its encoder-decoder structure with skip connections, also encounters challenges in addressing the spectral complexity and spatial variability often seen in flood detection scenarios. These insights are further corroborated by the F1-score and mIoU metrics, with CCT-U-ViT achieving 0.9121 and 0.8383, respectively, reflecting a superior balance of precision and recall, as well as intersection-over-union performance, which are critical for accurately defining flood boundaries.

Table 2. Accuracy assessment of the proposed and benchmark models for flood detection on the Erbil dataset.

Model	Metric			
	OA	Kappa	F1 score	mIoU
CNN 2D - 1 Layer	0.9041	0.8082	0.9031	0.8234
CNN 2D - 2 Layers	0.9056	0.8113	0.9044	0.8254
CNN 3D - 1 Layer	0.8741	0.7482	0.8642	0.7609
CNN 3D - 2 Layers	0.8892	0.7784	0.8845	0.7929
Hybrid CNN	0.9008	0.8016	0.8985	0.8158
U-Net	0.8953	0.7907	0.9013	0.8204
ViT	0.9064	0.8127	0.9086	0.8326
CNN - Transformer	0.9079	0.8158	0.9076	0.8308
CCT-U-ViT (ours)	0.9124	0.8248	0.9121	0.8383

The qualitative assessment of flood detection results in Figures 9 and 10 highlights differences in spatial pattern recognition abilities among the architectures evaluated, affecting operational flood monitoring applications. The ground truth reference (Figure 9c) displays intricate flood patterns with irregular boundaries and fragmented water bodies throughout the study area, posing significant challenges for automated detection algorithms. Visual analysis reveals that CCT-U-ViT generates the most spatially coherent flood maps with excellent boundary preservation, thanks to the hierarchical feature extraction capabilities of the compact convolutional transformer and the multi-scale representation learning of the U-Net decoder pathway.

Transformer-based models, such as ViT, CNN-Transformer, and CCT-U-ViT, preserve flood boundary integrity and minimize salt-and-pepper noise artifacts frequently observed in pixel-wise classification methods. This improvement is due to the self-attention mechanism, which effectively captures long-range spatial dependencies and contextual relationships, facilitating superior differentiation between spectrally similar but spatially distinct land cover types. In contrast, traditional CNN architectures face challenges with fine-scale boundary details; 2D CNNs tend to produce excessively smoothed boundaries attributable to successive pooling operations, while 3D CNNs demonstrate significant fragmentation, likely stemming from difficulties in learning optimal spatio-temporal filters with limited training data.

The U-Net architecture demonstrates moderate performance and exhibits typical encoder-decoder artifacts. This includes occasional missed detections in smaller flooded areas and minor over-segmentation in transition zones. Such behavior is consistent with U-Net's optimization for biomedical image segmentation, where its skip connections, while preserving spatial resolution, may struggle with the spectral complexity and radiometric variations found in multi-spectral satellite imagery. The superior capabilities of attention-based models—especially in sustaining spatial coherence—highlight the advantages of explicitly modeling spatial relationships through attention mechanisms, compared to purely convolutional methods in complex earth observation tasks.

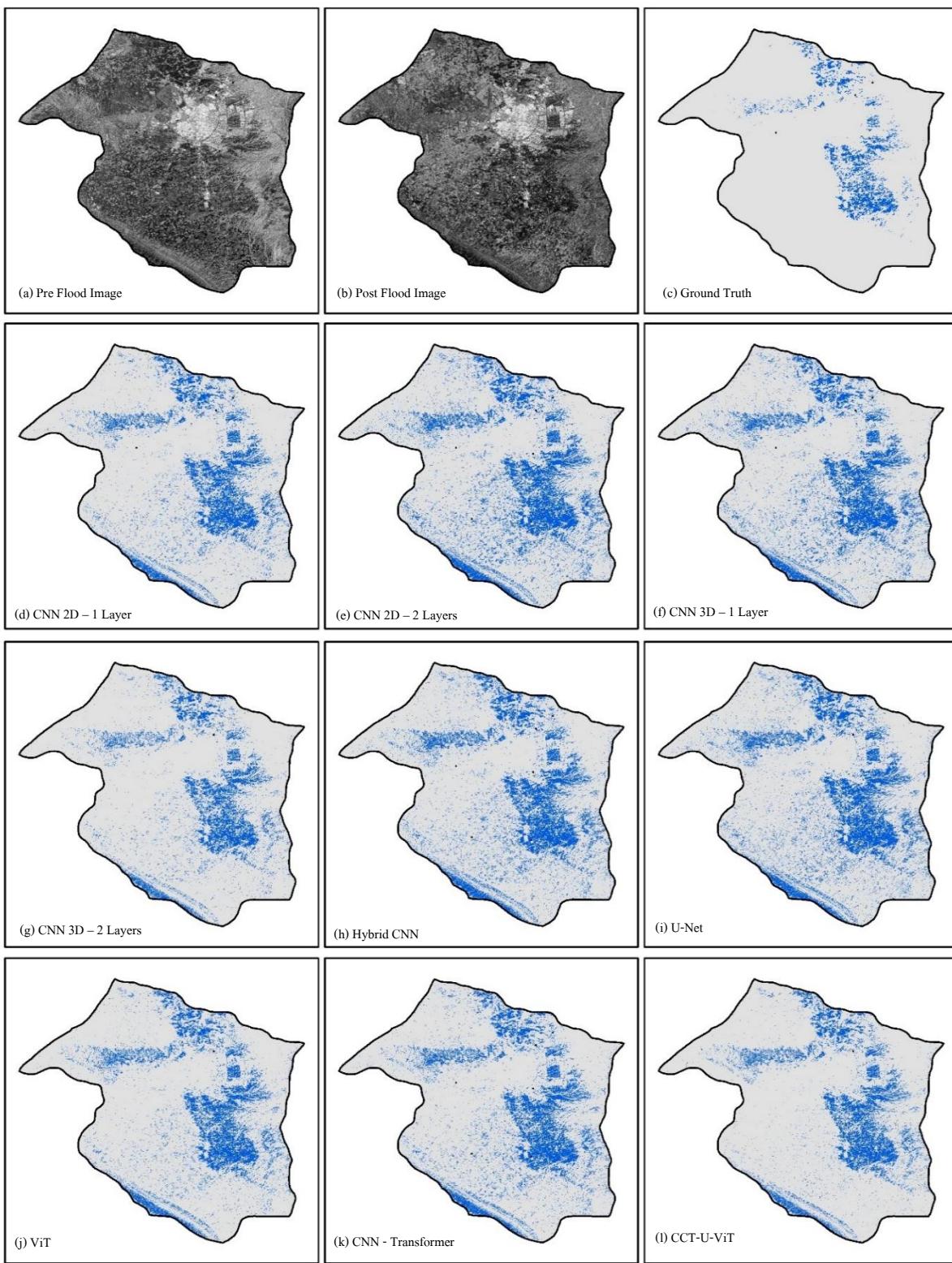


Figure 9. Maps of the flood detection of the Erbil dataset using different deep learning models.

The detailed error analysis presented in Figure 10 provides essential insights into the modes of model failure and their root physical and methodological causes. The patterns of error distribution highlight consistent biases linked to key difficulties in satellite-based flood detection, such as spectral confusion, mixed pixel effects, and atmospheric interference. False positive detections, illustrated in yellow, mainly occur in regions with high soil moisture, shadows from topography, and permanent water bodies that closely resemble the spectral signatures of floodwaters. These errors are particularly evident in 3D CNN models, suggesting that including the temporal dimension, despite

its theoretical benefits, may introduce noise when training data lacks adequate temporal diversity or when atmospheric conditions differ significantly between acquisition dates.

False negative detections (shown in magenta) primarily occur in shallow flood zones and vegetated wetlands, where emergent vegetation or sediment load weakens the water signal. The lower false negative rates in transformer-based models are attributed to their superior capacity to capture contextual details and subtle spectral changes via global attention mechanisms. The CCT-U-ViT model exhibits the most balanced error distribution, featuring significantly reduced false positive rates (indicating enhanced precision) while maintaining high sensitivity for actual flood detection. This indicates that the hybrid architecture effectively integrates the spatial inductive biases of CNNs with the contextual modeling strengths of transformers.

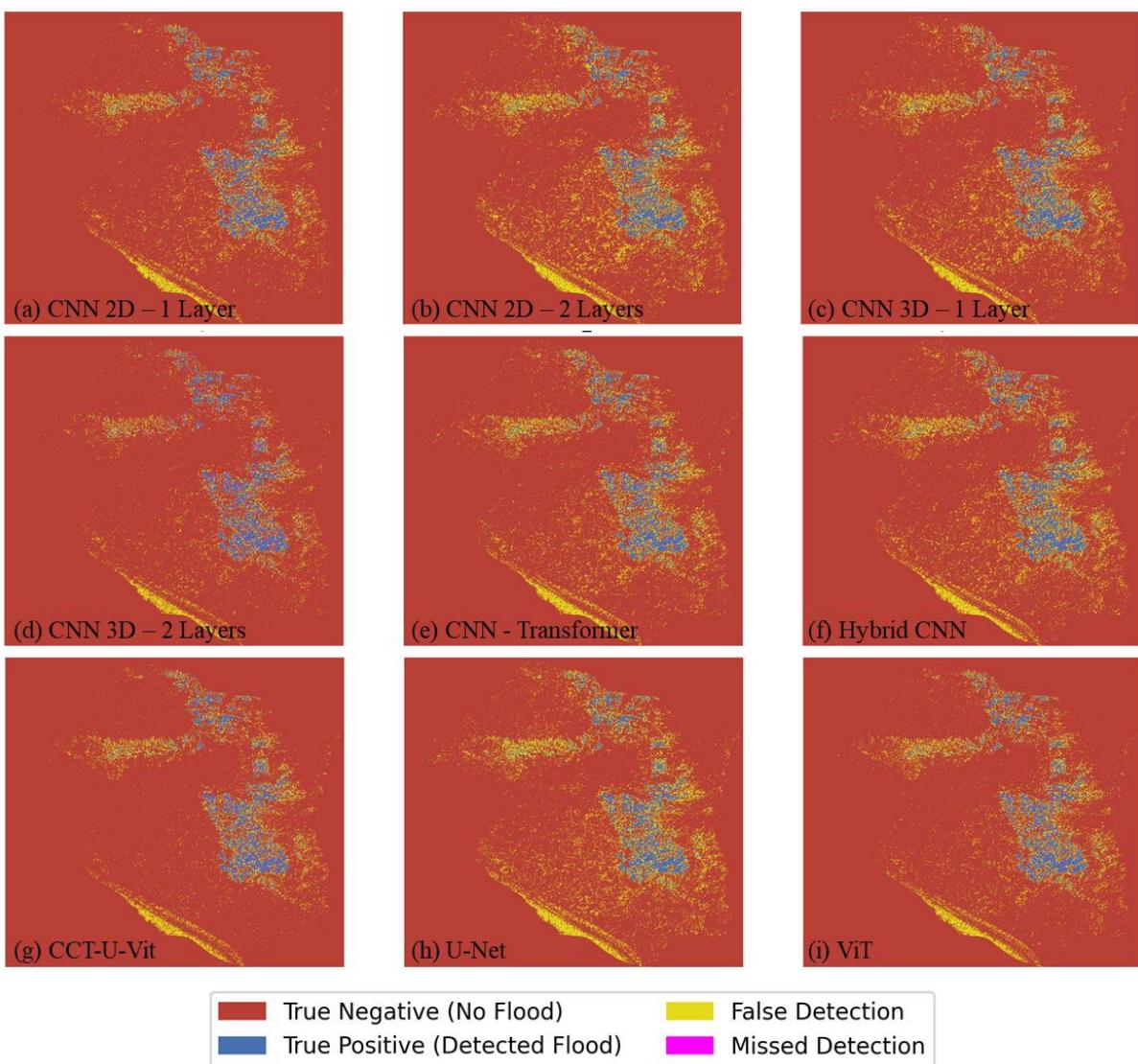


Figure 10. Error distributions of the flood detection for the proposed and baseline models.

5.3. Experimental Results on the S1GFloods Dataset

The thorough assessment of the S1GFloods benchmark dataset reveals the outstanding performance of the CCT-U-ViT architecture, achieving leading results in all evaluation metrics and setting new benchmarks for satellite-based flood detection (Table 3). This model reached a top overall accuracy of 97.9%, marking a noteworthy 1.1% improvement over the second-best method (CNN-Transformer at 96.8%), a considerable leap given the already impressive baseline performance on this challenging dataset. The Kappa coefficient 0.969 suggests near-perfect agreement beyond mere

chance, underscoring the model's exceptional reliability for operational flood monitoring applications.

The performance hierarchy identified in the S1GFloods dataset provides important insights into architectural design principles for flood detection tasks. Traditional CNN methods display a distinct performance gradient, with deeper architectures (CNN 2D - 2 Layers: 95.6% OA) surpassing their shallower versions (CNN 2D - 1 Layer: 93.7% OA) due to better feature abstraction capabilities and broader receptive field coverage. 3D CNN architectures demonstrate competitive performance (95.9% OA for a single layer, 95.5% OA for two layers), suggesting that the temporal dimension offers valuable information when ample training data is available, as evident with the extensive S1GFloods benchmark. The Vision Transformer achieved excellent results (96.1% OA, 0.945 Kappa) by utilizing global attention mechanisms, although its computational complexity poses a practical challenge for large-scale use.

The CNN-Transformer hybrid approach (96.8% OA, 0.958 Kappa) demonstrates the effectiveness of combining convolutional inductive biases with transformer attention mechanisms, achieving superior performance compared to pure CNN or transformer architectures. However, the proposed CCT-U-ViT surpasses all baseline methods with an F1-score of 0.966 and mIoU of 0.933, indicating exceptional precision-recall balance and spatial overlap accuracy. The consistent performance gains across all metrics (2.1% improvement in Kappa, 1.1% in OA, 1.4% in F1-score, and 1.8% in mIoU compared to the second-best method) provide strong empirical evidence for the architectural innovations incorporated in the CCT-U-ViT design, particularly the effective integration of multi-scale feature fusion with hierarchical attention mechanisms optimized for satellite imagery analysis.

Table 3. Accuracy assessment of the proposed and benchmark models for flood detection on the S1GFloods dataset.

Method	Kappa	OA	F1 Score	mIoU
CNN 2D - 1 Layer	0.872	0.937	0.911	0.843
CNN 2D - 2 Layers	0.908	0.956	0.933	0.877
CNN 3D - 1 Layer	0.911	0.959	0.937	0.882
U-Net	0.927	0.955	0.932	0.877
CNN 3D - 2 Layers	0.930	0.955	0.932	0.878
ViT	0.945	0.961	0.941	0.896
Hybrid CNN	0.936	0.962	0.942	0.894
CNN-Transformer	0.958	0.968	0.952	0.915
CCT-U-ViT (ours)	0.969	0.979	0.966	0.933

5.4. Comparison of Computational Efficiency

The assessment of computational efficiency reveals trade-offs between model complexity and performance, affecting deployment in resource-limited settings (Table 4). Conventional CNN architectures show excellent efficiency, with the 2D single-layer model needing only 2,386 parameters (0.03 MB) and achieving an inference time of 0.0181 seconds per batch. The 3D CNN variants have a slight increase in parameters (4,306-9,618) while maintaining competitive inference speeds, demonstrating effective use of the temporal dimension. However, this efficiency correlates with lower accuracy compared to attention-based methods.

Transformer-based architectures require more computation, with ViT needing 372,770 parameters (1.57 MB) and the slowest inference time at 0.0383 seconds per batch, highlighting the quadratic complexity of self-attention. In contrast, the CNN-Transformer hybrid balances performance, having 407,683 parameters (1.64 MB) and faster inference (0.0231 seconds) than pure ViT, illustrating the benefits of convolutional inductive biases. U-Net shows moderate complexity with 237,946 parameters (1.12 MB) and competitive speed (0.0233 seconds), making it suitable for

accuracy and efficiency. The CCT-U-ViT is the most complex architecture, with 669,482 parameters (2.82 MB) and an inference time of 0.0258 seconds per batch, positioned between pure transformers and hybrids in computational cost. Despite its size, CCT-U-ViT offers an excellent accuracy-efficiency balance, with only an 11% slower time than the fastest transformer method (CNN-Transformer), achieving greater accuracy. The parameter-to-performance ratio indicates CCT-U-ViT gains a 0.145% accuracy enhancement per additional 1,000 parameters compared to CNN-Transformer, demonstrating effective model capacity use for flood detection. Operationally, the model's 2.82 MB size is feasible for modern systems, and its sub-second inference allows real-time processing of satellite imagery for emergency response.

Table 4. Comparison of computational efficiency of the proposed and baseline models for flood detection based on the Erbil dataset.

Model	# Layers	# Parameters	Trainable Parameters	Model Size (MB)	Inference Time (Batch 64) (sec)	Batch Time (Batch 64) (sec)
CNN 2D - 1 Layer	4	2386	2386	0.03	0.0232	634.62
CNN 2D - 2 Layers	5	4754	4754	0.04	0.0181	495.86
CNN 3D - 1 Layer	4	4306	4306	0.03	0.0218	595.25
CNN 3D - 2 Layers	5	9618	9618	0.06	0.0183	501.54
CNN - Transformer	29	407683	407683	1.64	0.0231	632.13
Hybrid CNN	9	4130	4130	0.04	0.02	545.8
U-Net	69	237946	236474	1.12	0.0233	638.21
CCT-U-ViT	95	669482	668010	2.82	0.0258	705.91
ViT	47	372770	372770	1.57	0.0383	1047.76

6. Conclusions

This research investigated how deep learning techniques can detect floods using synthetic aperture radar (SAR) imagery, concentrating on U-Net and Transformer models for segmenting flood-affected areas. The primary aim was to evaluate the accuracy and appropriateness of these models in recognizing flood zones under varied conditions.

The results suggest that U-Net and Transformer models can achieve an excellent performance on flood detection, and Transformer better than U-Net in terms of precision and recall would come to the aid when the flood situation gets more complicated. Moreover, the findings suggest that the proposed deep learning techniques generalize well to other floods and locations and thus have potential in a wide range of flood monitoring systems. Compared with traditional methods, such as segmentation by thresholding and rule-based models, deep learning model significantly improves the quality and efficiency of segmentation. This enhancement is particularly striking in the reduction of human annotation that has been a bottleneck in many remote sensing tasks.

One important implication of our results is that recent advanced deep learning methodologies will also improve automated flood monitoring systems when applied to high-resolution SAR images. Such systems facilitate near real-time flood mapping, which support disaster response efforts, especially when on-ground information is scarce. The results of this study demonstrate the importance of SAR data for advanced machine learning-based flood response and preparedness, particularly with respect to such flood-prone areas where traditional infrastructure for flood detection is limited.

The result shows that the U-Net and Transformer models could successfully detect the flood in the SAR image. Both models perform well in identifying inundated regions, and they generalize well over diverse conditions, which illustrates their capacity for wide scope applications. Integration of these models in operational flood monitoring systems would substantially advance the detection of floods in terms of its speed accuracy by providing instant and timely information for disaster management.

Nevertheless, several issues remain to be resolved. To further improve model performance, hyperparameter fine-tuning and architectural refinement for real time flood detection can be potential directions for future exploration. Further, integrating SAR data with other remote sensing techniques can produce improved flood detection power, especially for poor environmental conditions. The integration of such models into operational flood monitoring systems would be key for the evaluation of their performance in dynamic, real time applications. Finally, the development of transfer learning techniques that allow models trained in one region to generalize to others may reduce the reliance on large annotated datasets in under-sampled regions, thereby making these models more accessible and viable for global flood monitoring.

References

1. Amitrano, D., Martino, G., Simone, A., & Imperatore, P. (2024). Flood Detection with SAR: A Review of Techniques and Datasets. *Remote. Sens.*, 16, 656. <https://doi.org/10.3390/rs16040656>.
2. Bai, Y., Wu, W., Yang, Z., Yu, J., Zhao, B., Liu, X., ... & Koshimura, S. (2021). Enhancement of detecting permanent water and temporary water in flood disasters by fusing sentinel-1 and sentinel-2 imagery using deep learning algorithms: Demonstration of sen1floods11 benchmark datasets. *Remote Sensing*, 13(11), 2220.
3. Bereczky, M., Wieland, M., Krullkowski, C., Martinis, S., & Plank, S. (2022). Sentinel-1-based water and flood mapping: Benchmarking convolutional neural networks against an operational rule-based processing chain. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15, 2023-2036.
4. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., & Wang, M. (2022, October). Swin-unet: Unet-like pure transformer for medical image segmentation. In European conference on computer vision (pp. 205-218). Cham: Springer Nature Switzerland.
5. Cao, H., Zhang, H., Wang, C., & Zhang, B. (2019). Operational Flood Detection Using Sentinel-1 SAR Data over Large Areas. *Water*. <https://doi.org/10.3390/W11040786>.
6. Chamatidis, I., Istrati, D., & Lagaros, N. D. (2024). Vision Transformer for Flood Detection Using Satellite Images from Sentinel-1 and Sentinel-2. *Water*, 16(12), 1670.
7. Chini, M., Pelich, R., Li, Y., Hostache, R., Zhao, J., Mauro, C., & Matgen, P. (2021). Sar-Based Flood Mapping, Where We Are and Future Challenges. 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, 884-886. <https://doi.org/10.1109/IGARSS47720.2021.9554975>.
8. Doan, T. N., & Le-Thi, D. N. (2025). A Novel Deep Learning Model for Flood Detection from Synthetic Aperture Radar Images. *Journal of Advances in Information Technology*, 16(1).
9. Doan, T. N., & Le-Thi, D. N. (2025). A Novel Deep Learning Model for Flood Detection from Synthetic Aperture Radar Images. *Journal of Advances in Information Technology*, 16(1).
10. Eftekhari, A., Samadzadegan, F., & Javan, F. D. (2023). Building change detection using the parallel spatial-channel attention block and edge-guided deep network. *International Journal of Applied Earth Observation and Geoinformation*, 117, 103180. <https://doi.org/10.1016/j.jag.2023.103180>.
11. Ghosh, B., Garg, S., Motagh, M., & Martinis, S. (2024). Automatic flood detection from Sentinel-1 data using a nested UNet model and a NASA benchmark dataset. *PFG-Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 92(1), 1-18.
12. Gu, Z., Cheng, J., Fu, H., Zhou, K., Hao, H., Zhao, Y., ... & Liu, J. (2019). Ce-net: Context encoder network for 2d medical image segmentation. *IEEE transactions on medical imaging*, 38(10), 2281-2292.
13. Guo, C., Szemenyei, M., Yi, Y., Wang, W., Chen, B., & Fan, C. (2021, January). Sa-unet: Spatial attention U-Net for retinal vessel segmentation. In 2020 25th international conference on pattern recognition (ICPR) (pp. 1236-1242). IEEE.

14. Hu, Y., Guo, Y., Wang, Y., Yu, J., Li, J., Zhou, S., & Chang, C. (2019). Automatic tumor segmentation in breast ultrasound images using a dilated fully convolutional network combined with an active contour model. *Medical physics*, 46(1), 215-228.
15. Huang, B., Li, P., Lu, H., Yin, J., Li, Z., & Wang, H. (2024). WaterDetectionNet: a new deep learning method for flood mapping with SAR image convolutional neural network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. 10.1109/JSTARS.2024.3440995.
16. Huang, M., & Jin, S. (2022). Backscatter Characteristics Analysis for Flood Mapping Using Multi-Temporal Sentinel-1 Images. *Remote. Sens.*, 14, 3838. <https://doi.org/10.3390/rs14153838>.
17. Lahsaini, M., Albano, F., Albano, R., Mazzariello, A., & Lacava, T. (2024). A Synthetic Aperture Radar-Based Robust Satellite Technique (RST) for Timely Mapping of Floods. *Remote. Sens.*, 16, 2193. <https://doi.org/10.3390/rs16122193>.
18. Li, Y., Martinis, S., Wieland, M., Schlaffer, S., & Natsuaki, R. (2019). Urban Flood Mapping Using SAR Intensity and Interferometric Coherence via Bayesian Network Fusion. *Remote. Sens.*, 11, 2231. <https://doi.org/10.3390/rs11192231>.
19. Liu, X., Gao, P., Yu, T., Wang, F., & Yuan, R. Y. (2025). CSWin-UNet: Transformer UNet with cross-shaped windows for medical image segmentation. *Information Fusion*, 113, 102634.
20. Liu, Z., Li, J., Wang, L., & Plaza, A. (2024). Integration of Remote Sensing and Crowdsourced Data for Fine-Grained Urban Flood Detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17, 13523-13532. <https://doi.org/10.1109/JSTARS.2024.3433010>.
21. Mason, D., Speck, R., Devereux, B., Schumann, G., Neal, J., & Bates, P. (2010). Flood Detection in Urban Areas Using TerraSAR-X. *IEEE Transactions on Geoscience and Remote Sensing*, 48, 882-894. <https://doi.org/10.1109/TGRS.2009.209236>.
22. Merz, B., Blöschl, G., Vorogushyn, S., Dottori, F., Aerts, J., Bates, P., Bertola, M., Kemter, M., Kreibich, H., Lall, U., & Macdonald, E. (2021). Causes, impacts and patterns of disastrous river floods. *Nature Reviews Earth & Environment*, 2, 592 - 609. <https://doi.org/10.1038/s43017-021-00195-3>.
23. Pech-May, F., Aquino-Santos, R., Álvarez-Cárdenas, O., Arandia, J. L., & Rios-Toledo, G. (2024). Segmentation and visualization of flooded areas through sentinel-1 images and u-net. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.
24. Puttinaovarat, S., & Horkaew, P. (2020). Internetworking flood disaster mitigation system based on remote sensing and mobile GIS. *Geomatics, Natural Hazards and Risk*, 11, 1886 - 1911. <https://doi.org/10.1080/19475705.2020.1815869>.
25. Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O. R., & Jagersand, M. (2020). U2-Net: Going deeper with nested U-structure for salient object detection. *Pattern recognition*, 106, 107404.
26. Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M., & Jagersand, M. (2019). Basnet: Boundary-aware salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7479-7489).
27. Qiu, P., Yang, J., Kumar, S., Ghosh, S. S., & Sotiras, A. (2024). AgileFormer: spatially agile transformer UNet for medical image segmentation. *arXiv preprint arXiv:2404.00122*.
28. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III* 18 (pp. 234-241)—Springer international publishing.
29. Saleh, T., Holail, S., Xiao, X., & Xia, G. (2024). High-precision flood detection and mapping via multi-temporal SAR change analysis with semantic token-based transformer. *Int. J. Appl. Earth Obs. Geoinformation*, 131, 103991. <https://doi.org/10.1016/j.jag.2024.103991>.
30. Saleh, T., Weng, X., Holail, S., Hao, C., & Xia, G. (2024). DAM-Net: Flood detection from SAR imagery using differential attention metric-based vision transformers. *ISPRS Journal of Photogrammetry and Remote Sensing*. <https://doi.org/10.1016/j.isprsjprs.2024.05.018>.
31. Saleh, T., Weng, X., Holail, S., Hao, C., & Xia, G. S. (2024). DAM-Net: Flood detection from SAR imagery using differential attention metric-based vision transformers. *ISPRS Journal of Photogrammetry and Remote Sensing*, 212, 440-453.

32. Schlaffer, S., Matgen, P., Hollaus, M., & Wagner, W. (2015). Flood detection from multi-temporal SAR data using harmonic analysis and change detection. *Int. J. Appl. Earth Obs. Geoinformation*, 38, 15-24. <https://doi.org/10.1016/j.jag.2014.12.001>.

33. Sha, Y., Zhang, Y., Ji, X., & Hu, L. (2021). Transformer-unet: Raw image processing with unet. arXiv preprint arXiv:2109.08417.

34. Sudiana, D., Riyanto, I., Rizkinia, M., Arief, R., Prabuwono, A. S., Sumantyo, J. T. S., & Wikantika, K. (2024). Performance Evaluation of 3-Dimensional Convolutional Neural Network for Multi-Temporal Flood Classification Framework with Synthetic Aperture Radar Image Data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. 10.1109/JSTARS.2024.3519523.

35. Tahermanesh, S., Mohammadzadeha, A., Mohsenifar, A., & Moghimi, A. (2025). SISCNet: A Novel Siamese Inception-Based Network with Spatial and Channel Attention for Flood Detection in Sentinel-1 Imagery. *Remote Sensing Applications: Society and Environment*, 101571. <https://doi.org/10.1016/j.rsa.2025.101571>.

36. Tanguy, M., Chokmani, K., Bernier, M., Poulin, J., & Raymond, S. (2017). River flood mapping in urban areas combining Radarsat-2 data and flood return period data. *Remote Sensing of Environment*, 198, 442-459. <https://doi.org/10.1016/J.RSE.2017.06.042>.

37. Tanoue, M., Taguchi, R., Nakata, S., Watanabe, S., Fujimori, S., & Hirabayashi, Y. (2020). Estimation of Direct and Indirect Economic Losses Caused by a Flood With Long-Lasting Inundation: Application to the 2011 Thailand Flood. *Water Resources Research*, 56. <https://doi.org/10.1029/2019WR026092>.

38. 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.

39. Tsokas, A., Rysz, M., Pardalos, P., & Dipple, K. (2022). SAR data applications in earth observation: An overview. *Expert Syst. Appl.*, 205, 117342. <https://doi.org/10.1016/j.eswa.2022.117342>.

40. Wang, D., Chen, X., Jiang, M., Du, S., Xu, B., & Wang, J. (2021). ADS-Net: An Attention-Based deeply supervised network for remote sensing image change detection. *International Journal of Applied Earth Observation and Geoinformation*, 101, 102348. <https://doi.org/10.1016/j.jag.2021.102348>.

41. Wang, F., & Feng, X. (2025). Flood change detection model based on an improved U-net network and multi-head attention mechanism. *Scientific Reports*, 15(1), 3295.

42. Wang, H., Xie, S., Lin, L., Iwamoto, Y., Han, X. H., Chen, Y. W., & Tong, R. (2022, May). Mixed transformer u-net for medical image segmentation. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 2390-2394). IEEE.

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

44. Wang, L., Li, R., Zhang, C., Fang, S., Duan, C., Meng, X., & Atkinson, P. M. (2022). UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 190, 196-214.

45. Wang, R., Cai, M., Xia, Z., & Zhou, Z. (2023). Remote sensing image road segmentation method integrating CNN-Transformer and UNet. *IEEE Access*, 11, 144446-144455.

46. Wang, Z., Zhang, C., & Atkinson, P. M. (2022). Combining SAR images with land cover products for rapid urban flood mapping. *Frontiers in Environmental Science*, 10, 973192. <https://doi.org/10.3389/fenvs.2022.973192>.

47. Yang, L., Wang, H., Meng, W., & Pan, H. (2024). CvT-UNet: A weld pool segmentation method integrating a CNN and a transformer. *Heliyon*, 10(15).

48. Zagoruyko, S., & Komodakis, N. (2015). Learning to compare image patches via convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4353-4361).

49. Zhang, H., Lian, J., & Ma, Y. (2025). FET-UNet: Merging CNN and transformer architectures for superior breast ultrasound image segmentation. *Physica Medica*, 133, 104969.

50. Zhang, Y., & Zhang, L. (2024). Detection of pavement cracks by deep learning models of transformer and UNet. *IEEE Transactions on Intelligent Transportation Systems*.

51. Zhou, L., Zhou, X., Feng, H., Liu, W., & Liu, H. (2025). Transformer-Based Semantic Segmentation for Flood Region Recognition in SAR Images. *IEEE Journal on Miniaturization for Air and Space Systems*.
52. Zhou, N., Xu, M., Shen, B., Hou, K., Liu, S., Sheng, H., ... & Wan, J. (2024). ViT-UNet: A Vision Transformer Based UNet Model for Coastal Wetland Classification Based on High Spatial Resolution Imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.
53. Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., & Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation. In Deep learning in medical image analysis and multimodal learning for clinical decision support: 4th international workshop, DLMIA 2018, and 8th international workshop, ML-CDS 2018, held in conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, proceedings 4 (pp. 3-11). Springer International Publishing.
54. Ziboon, Abdul Razzak T. 2019. "Monitoring of Agricultural Drought in the Middle Euphrates Area, Iraq Using Landsat Dataset." *Engineering and Technology Journal* 37(7 Part A).
55. Ziboon, Abdul Razzak T., Imzahim Abdulkareem Alwan, and Alaa Ghadban Khalaf. 2013. "Utilization of Remote Sensing Data and GIS Applications for Determination of the Land Cover Change in Karbala Governorate." *Engineering and Technology Journal* 31(15 Part (A) Engineering).
56. Noori, Abbas Mohammed, Abdul Razzak T. Ziboon, and Amjed N. AL-Hameedawi. 2024c. "Assessment of Flash Flood Detection in Erbil City Using Change Detection Indices for SAR Images." *Engineering and Technology Journal* 42(11):1378-86.
57. Noori, Abbas Mohammed, Abdul Razzak T. Ziboon, and Amjed Naser Al-Hameedawi. 2024a. "An Overview and Trends of Flood Detection, Hazard, Vulnerability and Risk Assessment." in AIP Conference Proceedings. Vol. 3219. AIP Publishing.
58. Noori, Abbas Mohammed, Abdul Razzak T. Ziboon, and Amjed Naser Al-Hameedawi. 2024b. "Flash Flood Susceptibility Mapping via Morphometric Analysis of Erbil City Basins, Iraq." in AIP Conference Proceedings. Vol. 3249. AIP Publishing.
59. Shihab, Tay H., and Amjed Naser Al-hameedawi. 2020. "Desertification Hazard Zonation in Central Iraq Using Multi-Criteria Evaluation and GIS." *Journal of the Indian Society of Remote Sensing* 48(3):397-409.
60. Aparna, A., & Sudha, N. (2022, May). SAR-FloodNet: a patch-based convolutional neural network for flood detection on SAR images. In 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC) (pp. 195-200). IEEE. 10.1109/ICAAIC53929.2022.9792770.
61. Fan, C. M., Liu, T. J., & Liu, K. H. (2022, May). SUNet: Swin transformer UNet for image denoising. In 2022 IEEE International Symposium on Circuits and Systems (ISCAS) (pp. 2333-2337). IEEE.
62. Sharma, N., & Saharia, M. (2024, December). Mapping floods from SAR data using CNNs and Vision transformers. In AGU Fall Meeting Abstracts (Vol. 2024, No. 1271, pp. H53M-1271).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.