Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.069374

ARTICLE

A Hybrid Deep Learning Approach Using Vision Transformer and U-Net for Flood Segmentation

Cyreneo Dofitas Jr[1], **Yong-Woon Kim**[2] **and Yung-Cheol Byun**[3,*]

[1]Department of Institute of Information Science and Technology, Jeju National University, Jeju-si, 63243, Republic of Korea
[2]Department of Computer Engineering, Jeju National University, Jeju, 63243, Republic of Korea
[3]Department of Computer Engineering, Major of Electronic Engineering, Jeju National University, Food Tech Center (FTC), Jeju National University, Jeju, 63243, Republic of Korea

*Corresponding Author: Yung-Cheol Byun. Email: ycb@jejunu.ac.kr

Received: 21 June 2025; Accepted: 27 August 2025; Published: 09 December 2025

ABSTRACT: Recent advances in deep learning have significantly improved flood detection and segmentation from aerial and satellite imagery. However, conventional convolutional neural networks (CNNs) often struggle in complex flood scenarios involving reflections, occlusions, or indistinct boundaries due to limited contextual modeling. To address these challenges, we propose a hybrid flood segmentation framework that integrates a Vision Transformer (ViT) encoder with a U-Net decoder, enhanced by a novel Flood-Aware Refinement Block (FARB). The FARB module improves boundary delineation and suppresses noise by combining residual smoothing with spatial-channel attention mechanisms. We evaluate our model on a UAV-acquired flood imagery dataset, demonstrating that the proposed ViT-UNet+FARB architecture outperforms existing CNN and Transformer-based models in terms of accuracy and mean Intersection over Union (mIoU). Detailed ablation studies further validate the contribution of each component, confirming that the FARB design significantly enhances segmentation quality. To its better performance and computational efficiency, the proposed framework is well-suited for flood monitoring and disaster response applications, particularly in resource-constrained environments.

KEYWORDS: Flood detection; vision transformer (ViT); U-Net segmentation; image processing; deep learning; artificial intelligence

1 Introduction

The increasing frequency and severity of floods driven by climate change demand advanced, accessible, and efficient flood detection solutions. Recent progress in deep learning, particularly with convolutional neural networks (CNNs) and Vision Transformers (ViTs), has significantly enhanced flood segmentation using aerial and satellite imagery. However, current methods face notable limitations. CNN-based models, such as U-Net variants, excel at capturing local spatial details but struggle to model global context, resulting in inaccurate delineation of complex flood boundaries in regions with water reflections, shadows, or low-contrast edges [1,2]. Furthermore, many existing approaches lack domain-specific post-processing to address flood-specific artifacts, such as noisy textures or ambiguous boundaries, which are prevalent in high-resolution aerial imagery [3,4]. These limitations hinder real-time deployment in disaster management, particularly in resource-constrained settings like developing countries, where affordable and robust flood notification systems are scarce.

Copyright © 2025 The Authors. Published by Tech Science Press.
This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

These systems can enhance the effectiveness of crisis response and resilience to floods, which have experienced a relatively high incidence in recent years [5]. Timely flood detection and monitoring are essential for minimizing these effects and implementing preventive measures to protect the affected population and mitigate financial losses. Despite the challenges and noise present in various image types, such as those from fixed surveillance cameras, drones, crowdsourced observations, and social media streams, automated image analysis techniques have rapidly evolved into flexible and robust tools for information extraction [6]. Here, we introduce a tailored post-processing refinement block that further improves segmentation boundaries, which is detailed in the Methods.

This study builds upon the established ViT-UNet hybrid backbone and centers its core contribution on the design and integration of the Flood-Aware Refinement Block (FARB), a novel post-processing module for improving flood boundary delineation. The key contributions of this work are as follows:

- Proposed a hybrid ViT-UNet framework that performs pixel-wise flood segmentation, integrating a novel Flood-Aware Refinement Block (FARB) that enhances boundary delineation through spatial-channel attention and residual refinement.
- To introduce a domain-specific flood-aware refinement block, a post-processing module that enhances segmentation quality by refining boundaries affected by reflections, shadows, and noisy textures. It integrates residual smoothing with spatial-channel attention to improve segmentation of complex flood contours.
- We develop a comprehensive and efficient training pipeline, incorporating a targeted data augmentation strategy that simulates diverse environmental conditions. This ensures model robustness and generalization to unseen flood scenarios.

2 Literature Review

Based on the literature, we selected widely adopted architectures for segmentation tasks, including U-Net variants and Swin Transformer, as strong CNN and transformer baselines. These models have been used extensively in both general and flood-specific segmentation studies.

The selection of U-Net variants (e.g., U-Net Conv2D, U-Net Residual, U-Net Attention Gate) was motivated by their established success in medical image segmentation—where precise boundary delineation is crucial—and their frequent adaptation to remote sensing applications, including flood mapping. Swin Transformer was included due to its strong performance in vision tasks, specifically its ability to model global spatial dependencies, which is critical for complex flood imagery. While other state-of-the-art architectures such as DeepLabv3+ and SegFormer were considered, they were excluded due to higher computational demands or limited suitability for UAV-based flood datasets at the target resolution (512 × 512 pixels). By focusing on U-Net and Transformer hybrids, our study aligns the comparative evaluation with the proposed ViT-UNet framework to directly highlight the added value of the Flood-Aware Refinement Block (FARB).

Vision Transformers (ViTs) have emerged as powerful architectures for image segmentation due to their ability to capture long-range dependencies and global contextual information. Recent studies have applied ViTs to flood detection tasks, demonstrating their potential in diverse imaging scenarios. We enhanced a Vision Transformer model for binary flood classification by transferring knowledge across multispectral imagery, and we reached CNN-based models for disaster management [7,8]. SemT-Former detects real floods in high-risk regions [9]. The latest methods leverage UAV images with probabilistic optimization for flood mapping [10]. FWSARNet leverages deformable convolutions rather than self-attention to better detect the boundaries in SAR imagery [11]. This presents Vision Transformer-based models with differential attention to improve flood mapping in SAR and camera images. Federated learning transformers for rain and flood classification in urban areas [12]. While previous models such as Swin Transformer and DeepSARFlood

lack a mechanism to refine segmented boundaries in flood-prone urban environments, we address this persistent limitation by introducing a domain-specific post-processing innovation. Recent advances demonstrate the effectiveness of combining Transformer-based encoders with CNN-based decoders for semantic segmentation, and our work adopts this established hybrid backbone [13,14]. We extend it with the Flood-Aware Refinement Block (FARB), specifically designed to correct boundary-level artifacts—such as water reflections, terrain occlusions, or low-contrast flood areas—that remain problematic in earlier designs.

Selective feature reinforcement and adaptive token extraction have proven effective for highlighting flood-affected areas and improving the delineation of water borders, debris, and infrastructure damage. Following this line, our work integrates Vision Transformer (ViT) with UNet-based segmentation, drawing inspiration from hybrid transformer-CNN strategies like AtTransUNet [8,15], to strike a balance between fine-grained spatial segmentation and global contextual understanding.

In urban flood detection, visual transformers combined with hydrometeorological forecasting systems have shown strong performance in real-time rain and flood classification, supporting early warning and post-disaster assessment efforts. Additionally, incorporating attention gates into UNet architectures improves focus on critical features, significantly enhancing flood area segmentation for disaster management and response [16]. The adoption of deep learning, convolutional neural network (CNN) has enhanced flood detection precision. Models like Multi-Scale Deeplab, UNet, and Feature Pyramid Networks (FPN) segment flood-affected areas by leveraging neural network architectures [17,18].

To tackle challenges in SAR image processing and the lack of labeled data, a near real-time flood detection method for the Yangtze River Basin, was presented. In examining the performance of CNN models, a FCN-8, SegNet, UNet, and DeepResUNet, on 16 flood cases, it was found that UNet and DeepResUNet had the best pixelwise accuracy compared to traditional thresholding [19]. Existing solutions have included connected vision systems and deep learning pipelines to improve flood detection by leveraging various, real-time data sources like social media, traffic cameras and river feeds to enrich their training datasets [20,21].

Deep learning models, such as UNet, SegNet, and FCN32, have achieved high accuracy in flood zone delineation using aerial and satellite images while improved performance is observed in the case of UNet and DeepResUNet for SAR image analysis [19]. High-level architectures involve transformer encoders and geophysical information as well as U-Net frameworks for generating flood inundation maps, without forgetting EfficientNet-B7 encoders and multi-scale averaging, facilitating better performance on varying areas [22,23]. In addition, a CNN-based approach for classification of flood water depth as knee-high, waist-high, and dangerously high was also developed, where a strong real-time classification performance was obtained by using deep architectures, such as VGG16, ResNet, and Inception, with ResNet-50 architecture offering the best results. Their system, along with being connected to a mobile application, gleans the feasible application of deep learning for real-time disaster coping [24–27].

Recent research has explored hybrid architectures have been proposed to improve computational efficiency and segment quality. WVResU-Net Another recent method for flood mapping using the Sentinel-1 SAR source, which combines Vision Multi-Layer Perceptrons and a ResU-Net to cope with the computational complexity and to enhance the features extraction ResU-Net. DeepLabv3+ and ASPP have achieved good results on post-disaster elastic research of UAV images [28]. LSTM-SegNet-MSA and ES-DeepLab, attention-based schemes, improve flood risk estimation by integrating spatial and temporal information. U-Net is the classic model in medical image processing and ViT is widely used to process global contextual information, which are all important in flood segmentation [29].

Building on these advancements, our proposed study integrates Vision Transformer (ViT) with UNet-based segmentation models to exploit their complementary strengths. By focusing on multi-temporal analysis and post-processing refinement, we aim to deliver more precise flood delineation, advancing the state of the art in flood detection research.

3 Methods

This research presents a hybrid deep learning approach that integrates Vision Transformers (ViT) and U-Net to achieve high-precision flood segmentation. The methodological pipeline consists of dataset acquisition, pre-processing, feature extraction, segmentation, post-processing, and evaluation, shown in the Fig. 1. This workflow is designed to ensure accuracy, computational efficiency, and real-world applicability.

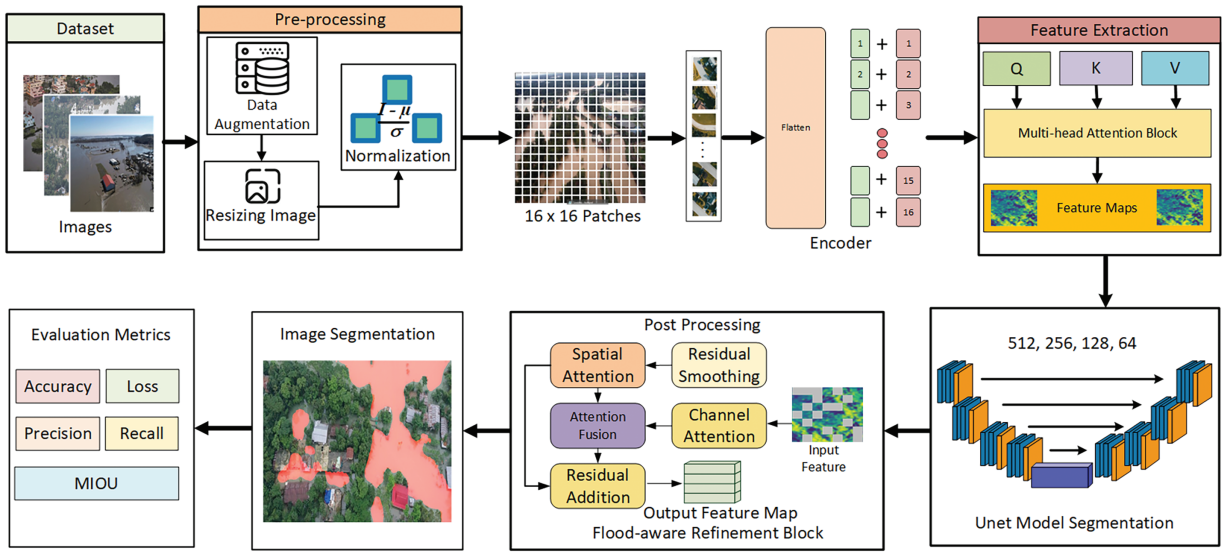

Figure 1: Workflow of the proposed model. This diagram illustrates the end-to-end pipeline, including pre-processing, Vision Transformer encoding, U-Net decoding, post-processing with the Flood-Aware Refinement Block, and output evaluation

3.1 Dataset Collection

Flood-specific datasets were developed by processing fine-resolution imagery obtained through autonomous aerial systems. This dataset comprised 290 original images, which were expanded to 1160 images through data augmentation, as shown in Fig. 2. Augmentation procedures included controlled brightness adjustment (±20%), contrast shifting (±25%), horizontal and vertical flipping (applied with a 50% probability), and salt-and-pepper noise injection (applied with a 5% probability). These transformations were carefully chosen to simulate diverse environmental conditions observed in real-world flood imagery, such as varying lighting, reflections, and occlusions. The augmentation strategy enhances data diversity, mitigates overfitting, and promotes better generalization across challenging flood scenarios. The dataset was split into 70% for training with 812 images, 20% for validation with 232 images, and 10% for testing with 116 images, ensuring a balanced evaluation.

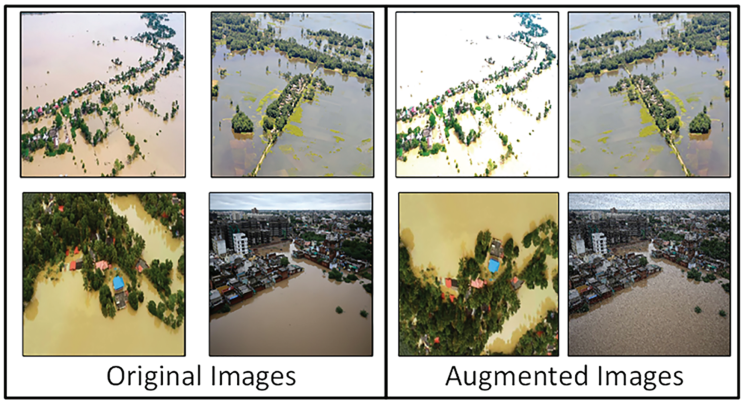

Figure 2: Sample original and augmented flood images. The left panel shows the original UAV-acquired images, while the right panel displays augmented images generated using brightness adjustment, contrast shift, flipping, and salt-and-pepper noise to simulate diverse environmental conditions

3.2 Pre-Processing

Prior to model training, a comprehensive series of pre-processing steps was applied to all images to ensure high-quality input for the segmentation models. First, each image was resized to a standard resolution of 512 × 512 pixels, aligning with the input specifications required by both the Vision Transformer (ViT) and U-Net architectures. The images were then converted to the YCbCr color space and normalized, which served to standardize pixel intensity distributions, reduce variability, and promote stable training behavior.

For processing through the Vision Transformer encoder, images were divided into 16 × 16 patches with a 4-pixel overlap, maintaining spatial continuity across patch boundaries and facilitating efficient patch-wise embedding. Finally, batch normalization was integrated into the training pipeline, stabilizing the learning process by reducing covariate shift in the feature maps and accelerating convergence. Collectively, these pre-processing and augmentation steps optimized the dataset for deep learning, improved model generalization, and contributed to the reliability of segmentation results across diverse and challenging flood scenarios. Cross-entropy loss was used for training, and evaluation metrics included precision, recall, accuracy, F1-score, and mean IoU. Model comparisons were statistically validated using the Kruskal-Wallis test.

The images were normalized, resized and converted to YCbCr color space before embedding into the ViT Encoder using patch-wise embedding. The uniform resolution matches the ViT and U-Net segmentation model for computational efficiency. The dataset is separated into three subsets for learning, tuning, and evaluation. These pre-processing steps prepare data for the ViT-UNet model to delineate flood-affected regions, ensuring reliable segmentation results as shown in Eq. (1). Additionally, batch normalization was applied during training to stabilize learning and improve convergence.

$$I_{\text{norm}} = \frac{I - \mu}{\sigma} \tag{1}$$

Here I, μ, and σ represent the input image, mean, and standard deviation. This reduces input variability.

During normalization, images are divided into 16 × 16 pixel patches with 4-pixel overlap between adjacent patches for efficient Vision Transformer processing. The overlap maintains spatial continuity across boundaries. A 512 × 512 pixel input image generates 1024 patches.

3.3 ViT-UNet with Flood-Aware Refinement

Shown in Fig. 3, the ViT-UNet architecture starts by leveraging a Vision Transformer backbone to extract both global and local contextual features from input imagery. This is followed by a U-Net decoder, which reconstructs segmentation masks through a four-stage upsampling strategy. Each upsampling stage applies a transposed convolution succeeded by a 3 × 3 convolution, batch normalization, and ReLU activation to progressively restore spatial resolution and refine feature representations. Skip connections are employed to link corresponding encoder and decoder layers, effectively preserving fine-grained details and compensating for information lost during downsampling. Throughout these stages, the number of feature channels is systematically decreased (512, 256, 128, 64) to balance computational efficiency with representational power.

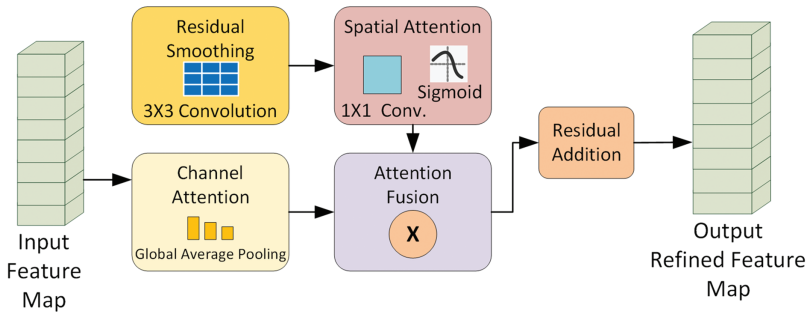

Figure 3: The input feature map from the U-Net decoder is processed in three parallel pathways, residual smoothing via a 3 × 3 convolution, spatial attention computed with a 1 × 1 convolution followed by sigmoid activation, and channel attention derived through global average pooling and fully connected layers. Attention Fusion combines spatial and channel attention via element-wise multiplication, after which the result is added to the smoothed features through a residual addition operation. The output is a refined feature map with improved boundary delineation and noise suppression for flood segmentation

To address the persistent challenges of complex flood boundaries, such as reflections and blurred edges the final feature map produced by the decoder is processed by the proposed Flood-Aware Refinement Block (FARB). This specialized module integrates residual smoothing with spatial and channel attention to enhance boundary delineation and suppress noise. The network outputs a binary, pixel-wise segmentation mask, in which each pixel is classified as either flooded or non-flooded, enabling fine-grained flood extent delineation.

To address the ambiguity of flood boundaries in aerial imagery, we propose a flood-aware refinement block (FARB), a learnable module integrating residual refinement for preserving essential spatial features, spatial-channel attention for highlighting flood regions and suppressing noise, and a convolution-based smoothing layer to ensure boundary continuity. This yields sharper and more reliable flood masks.

Unlike standard smoothing operations, we propose a refinement block that combines spatial-channel attention and a residual learning path. The residual connection preserves essential spatial features by combining smoothed inputs while filtering out irrelevant regions, improving accuracy. This process can be mathematically formalized, as shown in Eq. (6).

Let X be the input feature map:

$$X_{\text{smooth}} = \text{ReLU}(\text{BN}(\text{Conv}_{3\times3}(X))) \tag{2}$$

$$M_s = \sigma(\text{Conv}_{1\times1}(X)) \tag{3}$$

where σ is the sigmoid function.

$$M_c = \sigma(\mathrm{FC}_2(\mathrm{ReLU}(\mathrm{FC}_1(\mathrm{GAP}(X))))) \tag{4}$$

where GAP is global average pooling, FC_1 reduces channels to C/r, FC_2 restores to C, and σ is sigmoid.

$$X_{\mathrm{attn}} = X \times M_s \times M_c \tag{5}$$

$$X_{\mathrm{refined}} = X_{\mathrm{smooth}} + X_{\mathrm{attn}} \tag{6}$$

The decoder output is further refined by the proposed Flood-Aware Refinement Block (FARB). A detailed, step-by-step schematic of the FARB architecture, including all operations and their mathematical formulations, is provided in Table 1 and Eqs. (2) to (6). As illustrated in Fig. 3, FARB integrates residual smoothing, spatial attention, and channel attention with fusion and residual addition, sharpening segmentation boundaries and suppressing irrelevant noise. Complete hyperparameter details are provided to ensure full reproducibility. The complete architecture of the Flood-Aware Refinement Block (FARB) is summarized in which lists each operation, its function, and the resulting output shape. This table provides a clear step-by-step view of how residual smoothing, spatial attention, and channel attention are integrated and fused, ultimately producing the refined feature map used by the segmentation head.

Table 1: Detailed architecture of the attention refinement block

Layer/Operation	Description	Output shape
Input	Feature map from U-Net decoder	(B, C, H, W)
Residual smoothing	3×3 convolution (padding = 1), batch normalization, ReLU	(B, C, H, W)
Spatial attention	1×1 convolution, sigmoid (attention map)	$(B, 1, H, W)$
Channel attention	$\mathrm{GAP} \to \mathrm{FC}(C \to C/r) \to \mathrm{ReLU} \to \mathrm{FC}(C/r \to C) \to$ sigmoid	$(B, C, 1, 1)$
Attention fusion	Multiply by spatial and channel attention maps	(B, C, H, W)
Residual addition	Add smoothed features to attention-weighted features	(B, C, H, W)
Output	Refined feature map for segmentation head	(B, C, H, W)

The post-processing block also contributes to real-time prediction capability, enabling accurate and actionable outputs in resource-constrained flood response environment.

Computational Complexity Analysis

The Vision Transformer encoder introduces $O(n^2 d)$ complexity due to its self-attention mechanism, where n is the number of image patches (e.g., 1024 for a 512×512 input with 16×16 patches) and d is the embedding dimension (768). Although effective at modeling global dependencies, the quadratic scaling with sequence length can be computationally intensive. The U-Net decoder's complexity is mainly from convolutional operations, approximated as $O(hwc^2k^2)$ per layer, where h and w are spatial dimensions, c is the channel count, and k is the kernel size (typically 3). For greater efficiency, our model reduces channel dimensions progressively from 512 to 256 to 128 during upsampling. The Flood-Aware Refinement Block (FARB), a post-processing module, improves boundary delineation and suppresses noise by combining residual smoothing with spatial-channel attention mechanisms. Our model achieves a balanced trade-off, requiring only 1.29 h for training and 36.17 min for testing. It yields a segmentation accuracy of 93.09% and an mIoU of 0.8640.

We also report model complexity in terms of trainable parameters and FLOP ViT-UNet+FARB has 45.2 M parameters and 34.8 GFLOPs per 512 × 512 image during inference. For comparison, baseline U-Net Conv2D has 7.8 M parameters and 8.9 GFLOPs, while the Swin Transformer baseline has 63.1 M parameters and 42.5 GFLOPs. This demonstrates that ViT-UNet+FARB achieves a favorable trade-off between accuracy and computational cost, being more efficient than Swin and significantly outperforming traditional U-Net models.

$$Y_{\text{out}} = W_2 \cdot \text{ReLU}\left(\frac{W_1 \cdot X + b_1 - \mu}{\sqrt{\sigma^2 + \varepsilon}} + \beta\right) + b_2 \tag{7}$$

As shown in Table 2, our method achieves a favorable trade-off between accuracy and computational cost, being more efficient than Swin Transformer and significantly outperforming traditional U-Net models.

Table 2: Computational complexity comparison of segmentation models

Model	Trainable parameters (M)	FLOPs (G)	Test time (min)
U-Net Conv2D	7.8	23.5	35.52
U-Net residual	9.3	26.7	38.33
U-Net attention	10.1	28.9	37.19
Swin transformer	27.4	44.6	36.40
ViT-UNet (Ours)	88.2	62.1	36.17

To address the challenges of accurately segmenting ambiguous flood boundaries, we propose a Flood-Aware Refinement Block (FARB) as a learnable post-processing stage. This block integrates residual fine-tuning, spatial-channel attention, and convolutional smoothing to enhance mask predictions. The residual pathway preserves essential spatial features by learning an additive correction over the smoothed input, effectively reducing noise while maintaining edge sharpness. The spatial-channel attention mechanism emphasizes flood-relevant regions and suppresses background interference, allowing precise delineation of water extents. The convolution-based relative smoothing and attention module ensures boundary continuity and smooth transitions. These combined operations significantly sharpen the raw segmentation outputs while adding negligible computational overhead.

The model was trained using a low learning rate of 0.00001 to prevent oscillations in the loss function and ensure stable convergence. A batch size of 8 was used to balance GPU memory constraints with gradient accuracy. Training was conducted over 50 epochs, with validation loss stabilizing around epoch 45. To prevent overfitting on the small dataset, a dropout rate of 0.5 was applied, encouraging robust feature learning. The Adam optimizer was selected for its ability to dynamically adapt learning rates across parameters, proving more effective than SGD for our hybrid architecture. Furthermore, a smoothing module was incorporated to refine edge clarity and align the predicted features with the marginal distributions of flood and non-flood regions, thereby reducing noise and weak artifacts. Learnable parameters γ_c and β_c further allow adaptive refinement of feature maps, correcting inconsistencies from earlier layers. Overall, the post-processed masks are both visually precise and quantitatively reliable, supporting practical deployment in real-world flood detection systems.

4 Results

4.1 Training Configuration and Evaluation Metrics

To assess the segmentation performance of the proposed ViT-UNet with Flood-Aware Refinement Block (FARB), a comprehensive set of metrics was employed, accounting for the class imbalance inherent in flood datasets where non-flooded areas often dominate. The primary focus is on metrics that provide a balanced evaluation of segmentation quality. Mean Intersection over Union (mIoU) serves as a key indicator, measuring the overlap between predicted and ground truth segmentation masks across both flooded and non-flooded classes. Precision and recall are included to evaluate positive predictions and coverage, respectively, while the Dice Score (also known as F1-Score) is introduced as a critical metric, balancing precision and recall to better reflect performance on imbalanced datasets. Accuracy, while reported, is de-emphasized due to its potential to mislead in the presence of class imbalance, as a model predicting the majority (non-flooded) class could achieve high scores without effective segmentation.

4.2 Comparative Analysis with Other Models

The proposed ViT-UNet was benchmarked against recent state-of-the-art segmentation models, including U-Net variants and the Swin Transformer. Our model achieved the highest performance with an mIoU of 0.8640 and test accuracy of 93.09%, demonstrating a significant improvement over conventional CNN-based approaches. This can be attributed to the hybrid architecture, which leverages the ViT encoder for global context modeling and the U-Net decoder for precise spatial recovery.

All baseline models U-Net variants and Swin Transformer were retrained and evaluated using our UAV dataset with identical data splits, augmentation strategies, and preprocessing steps to ensure fair comparison. Our proposed ViT-UNet leverages the complementary strengths of Vision Transformers for global context modeling and U-Net decoders for fine spatial recovery. This combination, along with the Flood-Aware Refinement Block (FARB), enables superior boundary delineation and noise suppression in complex flood scenarios, as supported by both quantitative and visual results.

While other state-of-the-art segmentation architectures such as DeepLabv3+ and SegFormer offer strong performance in general-purpose vision tasks, they were not included in this study due to their significantly higher computational demands when applied to 512 × 512 UAV imagery. These models often require larger GPU memory and longer inference times, making them less practical for real-time flood monitoring in resource-constrained environments. Our benchmark selection thus focused on models that are widely adopted, flood-specific, and aligned with real-world deployment requirements.

While the total training duration for 50 epochs was approximately 1.29 h on an NVIDIA RTX 4080 GPU, this was achievable due to the relatively small dataset size (1160 images) and the moderate input resolution (512 × 512). A batch size of 8 and early convergence around epoch 45 further reduced training time. Although the measured inference speed of 0.15 s per image on the same GPU suggests real-time capability in high-performance environments, this does not directly imply feasibility on UAVs or other edge platforms. Real-time performance in such contexts must account for hardware limitations, power constraints, and thermal efficiency.

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i \left(\bar{R}_i - \bar{R}\right)^2 \tag{8}$$

where N denotes the total sample size across all groups, k represents the number of comparison groups, and n_i refers to the sample size within group i, $\bar{R}$ is the group i average rank, and $\bar{R}$ is average rank across all the groups. Applying the Kruskal-Wallis H-test to the model results yielded a test statistic of 16.73 and an

associated p-value of 0.002. Since this p-value falls below the conventional threshold of 0.05 and we confirm that the differences in performance among models are statistically significant. This finding suggests that the improvement of the proposed ViT-UNet model is not due to a random data-dependent fluctuation, but a systematic methodological advancement over the state-of-the-art techniques.

To assess whether the observed performance differences were statistically significant, we conducted a Kruskal-Wallis H-test using mIoU scores from all models under identical data splits. The test yielded a statistically significant result (p = 0.0003), indicating meaningful performance differences. Pairwise comparisons confirmed that the proposed ViT-UNet significantly outperformed both U-Net (p = 0.0032) and Swin Transformer (p = 0.0081), while the full three-way test also highlighted the contribution of FARB (p = 0.0017), as shown in Table 3.

Table 3: Statistical significance of segmentation performance differences among models using the Kruskal-Wallis H-test

Compared models	Metric	H-statistic	*p*-value
U-Net vs ViT-UNet	mIoU	8.72	0.0032
Swin vs ViT-UNet	mIoU	7.01	0.0081
ViT vs ViT-UNet vs FARB	mIoU	9.85	0.0017

To ensure a fair and consistent comparison, we re-implemented and trained all baseline models, including U-Net Conv2D, Residual U-Net, Attention U-Net, Multi-Residual U-Net, and Swin Transformer, on our augmented UAV dataset using the same experimental settings. These models were selected based on their frequent usage in recent flood segmentation and general semantic segmentation literature. Table 4 presents the implementation of various segmentation models with data augmentation. Results show that data augmentation increases accuracy, recall and precision across models. Classification error rates dropped significantly for models with poor coverage of the original dataset. UNet variants and Transformer architectures showed improved feature extraction and segmentation with augmented data. The Swin Transformer outperformed other approaches, demonstrating improvements in class imbalance as found previously and highlighting needs in data methodology and mining techniques. This confirms augmentation's importance for segmentation accuracy and model robustness.

Table 4: Performance comparison of segmentation models. All models in this table were trained and tested under the same experimental settings and data splits for consistent evaluation

	Evaluation metrics original dataset				Evaluation metrics data augmentation			
Model	Accuracy %	Recall %	Precision %	F1-score	Accuracy %	Recall %	Precision %	F1-score
Unet Conv 2D	75.82	74.28	73.53	73.90	79.24	75.53	76.22	75.87
UNet residual	79.27	76.88	76.49	76.68	81.28	79.57	73.61	76.47
Unet attention gate	81.16	80.88	80.86	80.87	82.04	83.59	81.77	82.67
Multi residual Unet	79.23	78.63	79.98	79.30	80.31	82.78	81.54	82.16
Swin transformer	87.30	85.36	86.14	85.75	89.37	83.85	84.97	84.41
Based model	84.72	81.86	82.27	84.78	86.15	84.26	85.43	84.82
Our proposed model	89.43	86.92	88.32	87.61	93.09	84.93	86.41	85.72

Table 5 presents a comparative summary of segmentation performance and computational cost. Traditional CNN models like U-Net Conv2D achieved faster training (01:03 h) but yielded lower mIoU (0.6952), while Swin Transformer demonstrated better context modeling (mIoU = 0.7962) but still struggled with boundary delineation, especially in reflective or ambiguous regions.

Table 5: Comparison of different model using the MIoU, training and test time. All models in this table were trained and tested under the same experimental settings and data splits for consistent evaluation

Model	Test mIoU	Training time hh:mm	Test time minutes	Key strengths
Unet convolutional 2D	0.6952	01:03	35.52	It struggles with long-range dependencies
UNet residual	0.7299	02:48	38.33	Residual connections improve feature flow and accuracy
Unet attention gate	0.7510	02:46	37.19	Attention mechanisms enhance focus on relevant features
Multi residual UNet	0.7247	03:26	38.64	Multi residual connections improve robustness but slow
Swin transformer	0.7962	02:15	36.40	Global context with self-attention but lacks boundary precision
Our proposed model	0.8640	01:29	36.17	This ViT-UNet with boundary refinement and efficient computation

The total training time for 50 epochs was approximately 1.29 h on an NVIDIA RTX 4080 GPU. This efficiency stems from the input resolution of 512 × 512 pixels. With a batch size of 8, each epoch involved approximately 145 steps, resulting in 7250 steps in total. Given this configuration, the average processing time per image per step was approximately 0.62 s, which is reasonable given the GPU capabilities and the streamlined hybrid architecture. Additionally, early convergence was observed around the 45th epoch, further reducing unnecessary iterations. While this training speed is feasible under these controlled settings, we acknowledge that training times would increase substantially for larger datasets or higher-resolution inputs due to the quadratic complexity of self-attention in ViT. Therefore, this timing should be interpreted within the context of the current experimental scale.

The multi-residual U-Net, despite improved architecture, required 4.06 h for training and 38.64 min for testing, while yielding an mIoU of only 0.7247—indicating diminishing returns with added complexity. Our model's architecture, which integrates the ViT's long-range contextual modeling with the U-Net's fine-grained spatial recovery, results in a superior segmentation framework suitable for real-time or near-real-time flood monitoring applications. We compare flood segmentation effects from various models including UNet-based architectures, Swin Transformer and our proposed ViT-UNet model, as shown in Fig. 4. The ViT-UNet model yields the highest MIoU of 0.8640, indicating best segmentation performance. UNet Convolutional 2D and ResNet architectures show lower MIoU scores and misclassify larger areas along complex flood boundaries. The Swin Transformer performs competitively due to self-attention mechanisms, though inferior to our proposed model. These results confirm the ViT-UNet hybrid approach better captures local and global flood features and improves segmentation compared to ViT and UNet. The model achieved strong metrics, reaching of 0.1448 and 0.1906 in training and validation loss, demonstrating good generalization.

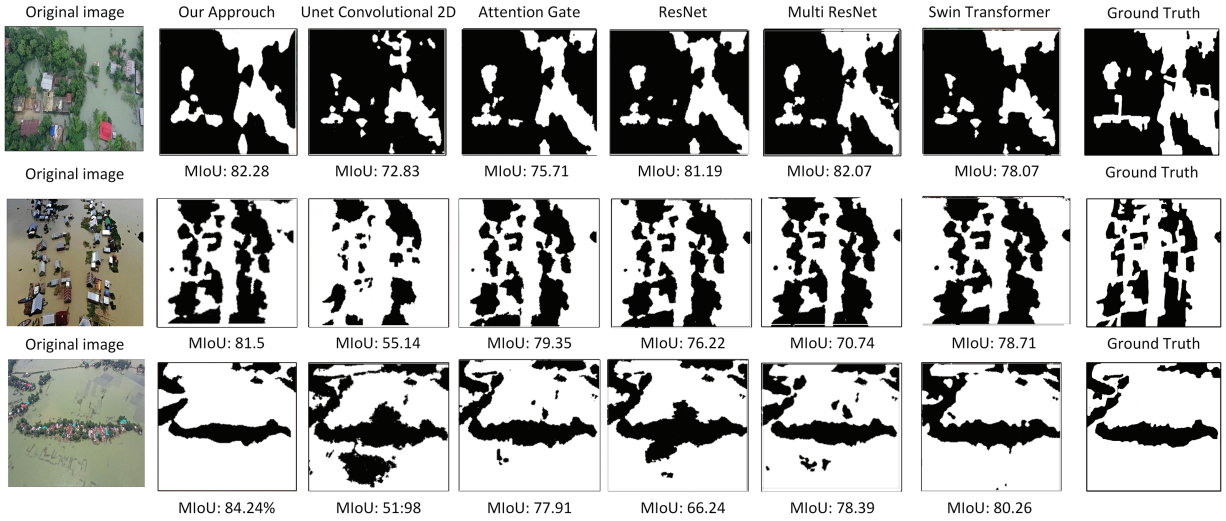

Figure 4: Visual comparison of segmentation results for the proposed ViT-UNet and various baseline models. The first and second rows highlight successful predictions with accurate flood boundary delineation. The third row illustrates failure or challenging cases, such as scenes with heavy reflections or complex textures, where the models perform suboptimally. Ground truth masks are included for each row to support visual comparison. This figure demonstrates both the strengths and limitations of the models across diverse flood

Table 6: Comparison of recent state-of-the-art flood segmentation approaches from reputable literature using public datasets

Reference	Year	Model	Dataset used	Accuracy (%)	mIoU
[30]	2023	Res-inception	Flood area segmentation	–	0.8472
[10]	2024	LAB color space method	UAV imagery (custom)	87.4	0.7400
[31]	2025	U-Net, SegNet, FCN32	Flood area segmentation	91.3	–
[32]	2024	SegNet	Flood area segmentation	89.46	–
[33]	2023	Graph attention convolutional, U-NET	Flood area segmentation	91	0.7600
Our proposed study	2025	ViT + Unet	Flood area segmentation	93.09	0.8649

As shown in Table 6, several recent models have reported strong flood segmentation performance on similar public datasets. For example, the Graph Attention Convolutional U-Net achieved an accuracy of 91% and mIoU of 0.7600, while our proposed ViT-UNet attained a higher accuracy of 93.09% and mIoU of 0.8640. Unlike [33], which focuses on graph-based spatial modeling, our approach explicitly combines global attention from Vision Transformers with the spatial precision of U-Net and further enhances boundary detection using the refinement block. Moreover, compared to [10], which used a color-space segmentation approach with 87.4% accuracy, and [32], which used SegNet with 89.46% accuracy, our model consistently

outperformed both in all metrics. These comparisons confirm that our method not only exceeds recent SOTA results in segmentation accuracy but also provides a tailored architecture for flood-specific challenges, validated using a publicly available benchmark.

In Fig. 5, specifically includes challenging and failure cases (bottom row), such as strong reflections or complex terrain, where all models, including our proposed ViT-UNet, demonstrate degraded boundary segmentation. Common failure patterns include missed small flood patches and confusion with urban surfaces or debris. These highlight the need for further data enrichment and targeted attention mechanisms.

Figure 5: Qualitative comparison of segmentation accuracy on external flood images using a baseline model and the proposed ViT-UNet. Panel **(a)** shows original aerial flood imagery; panel **(b)** displays predictions from the baseline model with notable misclassifications (highlighted in red circles); and panel **(c)** shows results from the proposed model, which more accurately captures flood extents, particularly in ambiguous regions, demonstrating superior generalization to unseen flood scenarios

The flooded areas are also presented on the segmented regions. Column (b) both the baseline and our model fail to segment regions with intense surface reflections (Image 2), while debris occlusion causes false positives in Image 3, while column (c) is our proposed method. For example in Image 1 the baseline is missing the water shown by the red circle, but ours is able to detect the water. The baseline does not recognize water

bodies in the red-circled locations in Image 2, while our method successfully detects them. In Image 3, the baseline predicts water, while ours establishes water from non-water correctly. These outcomes show the improvement of our ViT-UNet model in capturing flood boundaries.

We present a case study comparing segmentation outputs between the baseline models and our proposed ViT-UNet+FARB using test images. The top panel highlights successful cases where the proposed model captures flood boundaries more accurately and with better visual clarity. The bottom panel showcases difficult scenes with water reflections, debris, and occlusions where baseline models tend to fail, either missing flood regions or producing blurred boundaries. These visual comparisons demonstrate that the ViT-UNet+FARB model provides more consistent and refined segmentation across a variety of challenging conditions.

The ViT encoder uses multi-head self-attention (MHSA) to extract global context. The encoder consists of 12 attention layers, a configuration chosen based on extensive experimentation to balance performance and computational efficiency. A comparative analysis of different configurations (6, 8, and 12 attention layers) was conducted, and 12 layers provided the best trade-off between accuracy and training stability. The decision to use 12 layers was based on empirical results, as shown in Table 7, where this configuration demonstrated the highest accuracy of segmentation and Mean Intersection over Union (MIoU). This structure allows the model to effectively capture long-range dependences in flood imagery while maintaining feasible training times.

Table 7: Comparative analysis of attention layers

Attention layers	Parameter	Test accuracy %	MIoU
6	45,676,682	84.12	0.7721
8	59,852,426	87.45	0.8173
12	88,203,914	93.09	0.8640
14	102,379,658	88.29	0.8451

5 Discussion

Flood segmentation in imagery acquired by unmanned aerial vehicles (UAVs) or satellites remains a persistent challenge due to factors such as blurred boundaries, water reflections, low contrast, and debris-induced noise. Conventional convolutional neural network (CNN) models, including various U-Net architectures, exhibit strong capabilities in spatial detail recovery; however, they often lack the capacity to capture global context. Conversely, transformer-based approaches, such as Swin Transformer, are effective in modeling global dependencies but generally fall short in achieving precise spatial delineation, particularly around complex flood contours.

The proposed hybrid ViT-UNet architecture effectively addresses these limitations by integrating the long-range attention mechanisms inherent in Vision Transformers with the hierarchical spatial reconstruction abilities of U-Net decoders. A core innovation of this model is the Flood-Aware Refinement Block (FARB), a post-processing module designed to improve boundary clarity through the use of residual smoothing and spatial-channel attention mechanisms.

Ablation studies and comparative analyses substantiate that the FARB module markedly enhances segmentation outcomes, especially in scenarios where traditional models struggle. As detailed in Tables 5 and 8, incorporating FARB leads to significant improvements in mean Intersection over Union (mIoU) and boundary consistency, as demonstrated both quantitatively and through visual analysis, as shown in Fig. 4.

Unlike many prior models that demand substantial computational resources or fail to provide pixel-level refinement, the proposed approach achieves an optimal balance between segmentation accuracy and computational efficiency. This balance is particularly advantageous for operational flood monitoring in resource-constrained environments, where deployment on UAVs and local servers is more practical than reliance on high-performance computing infrastructure.

Table 8: Ablation study showing the impact of architectural components on segmentation performance

Model variant	Description	Accuracy (%)	mIoU
U-Net	Baseline U-Net decoder	81.24	0.7247
Vision transformer	ViT encoder with linear head	85.45	0.8140
ViT + U-Net	Hybrid model without FARB	88.12	0.8321
ViT + U-Net + Residual only	FARB with only residual smoothing	89.05	0.8379
ViT + U-Net + Spatial attention only	FARB with only spatial attention	88.67	0.8345
ViT + U-Net + Channel attention only	FARB with only channel attention	89.93	0.8364
ViT + U-Net + FARB (Ours)	**Full model**	**93.09**	**0.8640**

Comparative evaluations against state-of-the-art methods, including Swin Transformer and U-Net Attention Gate, demonstrate that ViT-UNet consistently delivers superior segmentation structure and enhanced resistance to noise. Qualitative assessments further reveal that ViT-UNet produces sharper boundaries and more accurately detects flood-affected regions, particularly in ambiguous areas characterized by shadows or intense reflectance.

Despite strong benchmark performance, some limitations persist, the detection of small, fragmented water patches or scenarios involving considerable debris. These challenges, discussed in Fig. 5, underscore the need for broader environmental datasets and the development of targeted augmentation strategies to further enhance model robustness.

We performed an ablation study to assess the contribution of key architectural components in our model, as shown in Table 8. The baseline U-Net decoder achieved an accuracy of 81.24% and a mean Intersection over Union (mIoU) of 0.7247. Incorporating the Vision Transformer (ViT) encoder alone improved the performance to 85.45% accuracy and 0.8140 mIoU. Further enhancement using the Flood-Aware Refinement Block (FARB) as a learnable post-processing module significantly boosted the accuracy to 93.09% and mIoU to 0.8640. This reflects a 7.64% absolute improvement in accuracy and a 5.00% increase in mIoU over the ViT-only configuration, highlighting the effectiveness of our proposed architectural design in improving segmentation precision.

The model was trained using a low learning rate of 0.00001 to prevent oscillations in the loss function and ensure stable convergence. A batch size of 8 was used to balance GPU memory constraints with gradient accuracy. Training was conducted over 50 epochs, with validation loss stabilizing around epoch 45. To prevent overfitting on the small dataset, a dropout rate of 0.5 was applied, encouraging robust feature learning. The Adam optimizer was selected for its ability to dynamically adapt learning rates across parameters, proving more effective than SGD for our hybrid architecture. Furthermore, a smoothing module was incorporated to refine edge clarity and align the predicted features with the marginal distributions of flood and non-flood regions, thereby reducing noise and weak artifacts. Learnable parameters γ_α and β_α further allow adaptive refinement of feature maps, correcting inconsistencies.

Hyperparameter Justification: The learning rate of 0.00001 was determined through empirical testing, where higher rates led to unstable convergence due to the complex feature distributions in flood imagery,

while lower rates ensured gradual optimization of the ViT-UNet hybrid model. The batch size of 8 was chosen to balance computational efficiency with gradient stability, as larger batch sizes 16 exceeded GPU memory limits, and smaller sizes 4 increased training time without significant performance gains. The dropout rate of 0.5 was selected based on common practices in deep learning literature for segmentation tasks, as it effectively mitigated overfitting on the augmented dataset of 1160 images, validated through cross-validation experiments showing improved generalization compared to lower rates of 0.3.

A principal limitation of this study is the reliance on a single UAV-acquired dataset, which, even after augmentation, may not capture the full spectrum of flood events encountered in different regions, under varying imaging conditions, or across diverse sensor types such as SAR or multispectral satellites. As a result, the generalizability of the model to new domains remains to be established.

In addition to segmentation quality, our model is designed for real-time processing, with optimized training and inference times. The competitive advantage of our approach lies in this balance between high accuracy and computational efficiency, making it suitable for deployment in real-world flood detection systems with limited computational resources.

6 Conclusion and Future Work

Our study shows a hybrid ViT-UNet segmentation model specifically designed for better flood detection using aerial imagery. By integrating a Vision Transformer encoder with a U-Net decoder and a specialized post-processing module, our approach enhances the delineation of flood boundaries. The extensive evaluation indicates that our suggested model performs a test accuracy of 93.09% and an mIoU of 0.8640, outperforming state-of-the-art models such as Swin Transformer, DeepSARFlood, and various U-Net variants in both accuracy and computational efficiency. Although we utilize a ViT-UNet hybrid backbone, our core contribution is the design and integration of the Flood-Aware Refinement Block (FARB), which is tailored to enhance segmentation quality in flood-prone environments. Overall, this framework offers substantial practical value for flood response and disaster management, particularly in disaster-prone or resource-limited regions utilizing UAV or satellite data. The enhanced ablation study also confirms the importance of each internal component of the FARB module, showing that combining residual refinement with spatial and channel attention yields the best segmentation performance.

For future work, we will enhance model generalization by incorporating a broader range of flood types, environmental dynamics, and geographic regions. Architectural improvements, such as relative positional encoding and adaptive token extraction, will be explored to further refine spatial understanding in complex flood environments. We also aim to support real-time flood tracking through temporal analysis and predictive modeling, optimize deployment for edge devices, and integrate multi-modal data sources from aerial imagery, ground sensors, and social platforms for a comprehensive flood detection solution.

Recognizing that our current validation is limited to a single UAV dataset, future work will rigorously evaluate the model on additional benchmarks such as FloodNet, Sen1Floods11, and SAR/satellite datasets to assess robustness across different sensors, geographic contexts (urban and rural), and environmental conditions.

Moreover, error cases, especially those involving reflections and occlusions, highlight limitations in training data diversity and model expressiveness. Addressing these will require additional data collection, targeted augmentation strategies, and improved attention mechanisms designed for challenging flood boundaries.

Acknowledgement: Not applicable.

Funding Statement: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00405278). This research was also partially supported by the Jeju Industry-University Convergence District Project for Promoting Industry-Campus Cooperation, funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea) [Project Name: Jeju Industry-University Convergence District Project for Promoting Industry-Campus Cooperation/Project Number: P0029950].

Author Contributions: The authors confirm contribution to the paper as follows: writing, visualization, methodology, formal analysis, conceptualization: Cyreneo Dofitas Jr; conceptualization, validation, review: Yong-Woon Kim; supervision, software, project administration, investigation, funding acquisition, resources: Yung-Cheol Byun. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The dataset used in this study is publicly available at: https://www.kaggle.com/datasets/faizalkarim/flood-area-segmentation (accessed on 26 August 2025).

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Bahrami B, Arbabkhah H. Enhanced flood detection through precise water segmentation using advanced deep learning models. J Civil Eng Res. 2024;6(1):1–8. doi:10.61186/jcer.6.1.1.
2. Song W, Guan M, Yu D. SwinFlood: a hybrid CNN-swin transformer model for rapid spatiotemporal flood simulation. J Hydrol. 2025;660(16):133280. doi:10.1016/j.jhydrol.2025.133280.
3. Fernandes FE Jr, Nonato LG, Ueyama J. A river flooding detection system based on deep learning and computer vision. Multimed Tools Appl. 2022;81(28):40231–51. doi:10.1007/s11042-022-12813-3.
4. Zhang Y, Crawford P. Automated extraction of visible floodwater in dense urban areas from RGB aerial photos. Remote Sens. 2020;12(14):2198. doi:10.3390/rs12142198.
5. Vineeth V, Neeba E. Flood detection using deep learning. In: 2021 International Conference on Advances in Computing and Communications (ICACC); 2021 Oct 21–23; Kochi, India. p. 1–5.
6. Zaffaroni M, Rossi C. Water segmentation with deep learning models for flood detection and monitoring. In: Proceedings of the 17th ISCRAM Conference; 2020 May 24–27; Blacksburg, VA, USA. p. 66–74.
7. Chamatidis I, Istrati D, Lagaros ND. Vision transformer for flood detection using satellite images from Sentinel-1 and Sentinel-2. Water. 2024;16(12):1670. doi:10.3390/w16121670.
8. Debnath A, Kim YW, Byun YC. LightSTATE: a generalized framework for real-time human activity detection using edge-based video processing and vision language models. IEEE Access. 2025;13(14):97609–27. doi:10.1109/access.2025.3574659.
9. Saleh T, Holail S, Xiao X, Xia GS. High-precision flood detection and mapping via multi-temporal SAR change analysis with semantic token-based transformer. Int J Appl Earth Obs Geoinf. 2024;131(1):103991. doi:10.1016/j.jag.2024.103991.
10. Simantiris G, Panagiotakis C. Unsupervised color-based flood segmentation in UAV imagery. Remote Sens. 2024;16(12):2126. doi:10.3390/rs16122126.
11. Yu H, Wang R, Li P, Zhang P. Flood detection in polarimetric SAR data using deformable convolutional vision model. Water. 2023;15(24):4202. doi:10.3390/w15244202.
12. Le QC, Le MQ, Tran MK, Le NQ, Tran MT. FL-Former: flood level estimation with vision transformer for images from cameras in urban areas. In: International Conference on Multimedia Modeling. Cham, Switzerland: Springer; 2023. p. 447–59.

13. Sharma NK, Saharia M. DeepSARFlood: rapid and Automated SAR-based flood inundation mapping using vision transformer-based deep ensembles with uncertainty estimates. Sci Remote Sens. 2025;11:100203. doi:10.1016/j.srs.2025.100203.
14. Nazir K, Byun YC. AETUnet: enhancing retinal segmentation with parameter-efficient unet architecture and lightweight attention mechanism. IEEE Access. 2025;13(3):33471–84. doi:10.1109/access.2025.3539372.
15. Li X, Pang S, Zhang R, Zhu J, Fu X, Tian Y, et al. Attransunet: an enhanced hybrid transformer architecture for ultrasound and histopathology image segmentation. Comput Biol Med. 2023;152(2):106365. doi:10.1016/j.compbiomed.2022.106365.
16. Madake J, Mali R, Shahapure A, More P, Bhatlawande S. FloodDetectionNet: U-Net attention based flooded area segmentation. In: International Conference on Innovations and Advances in Cognitive Systems. Cham, Switzerland: Springer; 2024. p. 319–33.
17. Wu H, Song H, Huang J, Zhong H, Zhan R, Teng X, et al. Flood detection in dual-polarization SAR images based on multi-scale deeplab model. Remote Sens. 2022;14(20):5181. doi:10.3390/rs14205181.
18. Islam MS, Sun X, Wang Z, Cheng I. FAPNET: feature fusion with adaptive patch for flood-water detection and monitoring. Sensors. 2022;22(21):8245. doi:10.3390/s22218245.
19. Wu X, Zhang Z, Xiong S, Zhang W, Tang J, Li Z, et al. A near-real-time flood detection method based on deep learning and SAR images. Remote Sens. 2023;15(8):2046. doi:10.3390/rs15082046.
20. Pally R, Samadi S. Application of image processing and convolutional neural networks for flood image classification and semantic segmentation. Environ Model Softw. 2022;148:105285.
21. Waqar M, Kim YW, Byun YC. A Stacking ensemble framework leveraging synthetic data for accurate and stable crop yield forecasting. IEEE Access. 2025;13(1):136909–26. doi:10.1109/access.2025.3591802.
22. Yin Z, Leon AS. Riverine flood hazard map prediction by neural networks. HydroResearch. 2025;8(16):139–51. doi:10.1016/j.hydres.2024.10.003.
23. Mesvari M, Shah-Hosseini R. Flood detection based on UNet++ segmentation method using Sentinel-1 satellite imagery. Earth Observ Geomat Eng. 2023;7(1):1–7.
24. Gandhi J, Gawde S, Ghorai A, Dholay S. Flood water depth classification using convolutional neural networks. In: 2021 International Conference on Emerging Smart Computing and Informatics (ESCI); 2021 Mar 5–7; Pune, India. p. 284–9.
25. Hidayat MT, Adiba F, Yahya M, Andayani DD, Kaswar AB. Enhanced flood detection on highways: a comparative study of MobileNet and VGG16 CNN models based on CCTV images. In: 2024 4th International Conference of Science and Information Technology in Smart Administration (ICSINTESA); 2024 Jul 12–13; Balikpapan, Indonesia. p. 125–30.
26. Du W, Qian M, He S, Xu L, Zhang X, Huang M, et al. An improved ResNet method for urban flooding water depth estimation from social media images. Measurement. 2025;242(2022):116114. doi:10.1016/j.measurement.2024.116114.
27. Lin H, Luo J, Zhou C, Tao Z. Research on model material selection based on inception similarity in impact analysis of flood overtopping on tailings dam. Environ Earth Sci. 2023;82(11):276. doi:10.1007/s12665-023-10965-5.
28. Wan J, Xue F, Shen Y, Song H, Shi P, Qin Y, et al. Automatic segmentation of urban flood extent in video image with DSS-YOLOv8n. J Hydrol. 2025;655(128545):132974. doi:10.1016/j.jhydrol.2025.132974.
29. Situ Z, Zhong Q, Zhang J, Teng S, Ge X, Zhou Q, et al. Attention-based deep learning framework for urban flood damage and risk assessment with improved flood prediction and land use segmentation. Int J Disaster Risk Reduct. 2025;116(2):105165. doi:10.1016/j.ijdrr.2024.105165.
30. Khan SD, Basalamah S. Multi-scale and context-aware framework for flood segmentation in post-disaster high resolution aerial images. Remote Sens. 2023;15(8):2208. doi:10.3390/rs15082208.
31. Ashikuzzaman M, Emran MA, Rahul MKA, Ornab SA, Rahman M, Abedin S. Flood area detection and segmentation using satellite image in Bangladesh by U-Net: a deep learning approach. In: 2025 International Conference on Electrical, Computer and Communication Engineering (ECCE); 2025 Feb 13–15; Chittagong, Bangladesh. p. 1–6.

32. Sunil P, Sinha S, Roy R. SegNet-ATT: cross-channel and spatial attention-enhanced U-Net for semantic segmentation of flood affected areas. In: International Conference on Pattern Recognition. Cham, Switzerland: Springer; 2024. p. 287–300.
33. Danish MU, Buwaneswaran M, Fonseka T, Grolinger K. Graph attention convolutional U-NET: a semantic segmentation model for identifying flooded areas. In: IECON 2024-50th Annual Conference of the IEEE Industrial Electronics Society; 2024 Nov 3–6; Chicago, IL, USA. p. 1–6.