Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.086985

ARTICLE

Attention-Guided Cross-Modal Transformer for Multimodal SAR-Optical Image Fusion and Flood Change Detection

Bayan Alabdullah[1], Muhammad Waqas Ahmed[2], Mohammad Shorfuzzaman[3,*], Jasem Almotiri[4], Mohammed Alonazi[5] and Ahmad Jalal[6,7,*]

[1]Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia
[2]Department of Computer Games Development, Air University, E-9, Islamabad, Pakistan
[3]Department of Software Engineering, College of Engineering and Advanced Computing, Alfaisal University, Riyadh, Saudi Arabia
[4]Department of Computer Science, College of Computers and Information Technology, Taif University, Taif, Saudi Arabia
[5]Department of Information Systems, College of Computer Engineering and Sciences, Prince Sattam bin Abdulaziz University, Al-Kharj, Saudi Arabia
[6]Faculty of Computing and AI, Air University, E-9, Islamabad, Pakistan
[7]Department of Computer Science and Engineering, College of Informatics, Korea University, Seoul, Republic of Korea

*Corresponding Authors: Mohammad Shorfuzzaman. Email: mshorfuzzaman@alfaisal.edu;
Ahmad Jalal. Email: ahmadjalal@mail.au.edu.pk

Received: 08 June 2026; Accepted: 13 July 2026; Published: 13 August 2026

ABSTRACT: Multimodal data fusion and deep learning have opened new frontiers in the analysis of complex visual data acquired from heterogeneous sensing systems. Flood inundation mapping represents one of the most demanding applications in this domain, requiring robust interpretation of complementary but conflicting image modalities under severe real-world constraints. This paper presents CAG-Transformer, a novel multimodal AI architecture for bi-temporal flood change detection through intelligent fusion of Sentinel-1 SAR and Sentinel-2 multispectral imagery. Three tightly integrated contributions address the core challenges of heterogeneous multimodal image analysis. A Change Attention Gate (CAG) performs adaptive channel-wise representation learning, selectively amplifying flood-relevant spectral and backscatter variations while suppressing temporally static scene content. A Cross-Modal Transformer (CMT) bottleneck employs multi-head self-attention to model long-range spatial dependencies and enable context-aware reasoning across heterogeneous sensor representations capabilities fundamentally beyond convolutional fusion. A differentiable soft Dice loss ensures stable gradient flow under the severe class imbalance inherent to real-world flood datasets. Evaluated on the Ombria multimodal benchmark, CAG-Transformer achieves IoU = 0.7774, Dice = 0.8894, and AUC-ROC = 0.9768, outperforming single-modality and conventional fusion baselines. Cross-event validation on Albania (IoU = 0.7647) and Timor (IoU = 0.7313) confirms generalization across environmentally and spatially distinct flood. With 3.5 million parameters and sub-100 ms inference on freely available Copernicus data, the framework offers an efficient and deployable solution for near-real-time flood monitoring and multimodal image-based environmental assessment.

KEYWORDS: Multimodal image analysis; data fusion; heterogeneous sensing systems; deep learning; change detection; flood inundation mapping; Sentinel-1 SAR; Sentinel-2 multispectral; transformer architecture; sensor calibration

1 Introduction

Floods account for more than 40% of all weather-related disasters, causing the largest cumulative economic losses of any natural hazard category over the past two decades [1]. The accelerating availability of Earth observation data has opened substantial new avenues for disaster monitoring, emergency response coordination, and geospatial decision support. Within this landscape, the integration of multimodal remote sensing data with advanced computational methods has emerged as a particularly productive research direction, drawing together disciplines ranging from signal and image processing to large-scale data management and machine learning. Accurate delineation of flood inundation extents is central to this effort, underpinning applications in emergency relief logistics, hydrodynamic model calibration, infrastructure risk assessment, and long-term urban resilience planning. The European Space Agency's Copernicus programme, specifically the Sentinel-1 synthetic aperture radar (SAR) and Sentinel-2 multispectral missions, provides freely accessible, complementary geospatial observations that are well-suited to continuous flood monitoring at both regional and global scales [2].

The analysis of heterogeneous multimodal imagery for flood detection presents fundamental challenges that have motivated substantial recent research at the intersection of deep learning and remote sensing. Sentinel-2 optical imagery captures rich spectral information through three flood-sensitive bands Green (Band 3, 0.560 μm), Near Infrared (Band 8, 0.842 μm), and Short-Wave Infrared (Band 11, 1.610 μm) which directly encode the spectral basis of established water detection, enabling effective discrimination of water bodies and land-cover types through spectral signatures and water indices however, optical observations are inherently susceptible to cloud cover and adverse atmospheric conditions that routinely accompany flood events. Sentinel-1 SAR imagery circumvents this limitation through its all-weather, day-and-night monitoring capability, yet SAR data introduce their own complexities, including speckle noise, geometric distortions, and intricate backscatter interactions that can obscure accurate flood delineation [3]. These contrasting and complementary sensing characteristics have naturally motivated multimodal data fusion approaches that combine heterogeneous sources to achieve more robust geospatial interpretation. Conventional fusion strategies and purely convolutional architecture have demonstrated encouraging results in this domain, but they face well-recognized constraints in their capacity for cross-modal representation learning and context-aware spatial reasoning across heterogeneous sensor modalities. Developing frameworks capable of adaptive multimodal feature integration, long-range spatial dependency modeling, and robust flood representation learning therefore remains an open and significant research challenge.

A systematic analysis of existing architectures reveals three specific and interconnected limitations that constrain performance on multimodal flood change detection. First, convolutional encoders treat all feature channels uniformly throughout the network, even though a substantial proportion of channels respond to static, temporally invariant scene properties persistent vegetation texture, permanent water bodies, building edge signatures that do not encode flood-related temporal change [4]. The propagation of these uninformative activations through difference operations introduces structured noise that degrades change signal quality at the bottleneck. Second, feature fusion between SAR and optical encoder branches are typically implemented through channel-wise concatenation followed by convolution, a strategy that restricts cross-modal interactions to spatially local neighborhoods and is fundamentally unable to model the long-range, context-dependent relationships that characterize spatially heterogeneous flood propagation across diverse land-cover types particularly the hydraulic connectivity between upstream tributaries and downstream floodplains that may span hundreds of pixels. Third, a known implementation risk in Dice-based segmentation losses involves the binarization of either predicted probabilities or ground-truth masks prior to gradient computation [5]. Under the severe positive-class imbalance of flood datasets (where flooded pixels represent only 8%–12% of image area), this non-differentiable step effectively zeroes out gradients

across the majority of training pixels, rendering the Dice component functionally inoperative and reducing training to near-exclusive reliance on binary cross-entropy, which optimizes pixel-wise classification rather than spatial overlap.

To address these limitations in a unified framework, we propose CAG-Transformer, an end-to-end deep learning architecture explicitly designed for multimodal flood inundation change detection via joint processing of Sentinel-1 SAR and Sentinel-2 optical bi-temporal image pairs. The principal contributions of this work are:

- A novel Change Attention Gate (CAG) is introduced a lightweight spatially adaptive channel-wise attention mechanism embedded within the encoder, which selectively enhances flood-relevant spectral and backscatter variations while suppressing redundant feature responses at each spatial location independently, enabling the same channel to be amplified at flood boundaries and suppressed over stable background regions within a single feature map.
- A Cross-Modal Transformer (CMT) bottleneck is developed, where tokenized SAR and optical difference features are jointly processed as a single concatenated sequence of 2048 tokens via stacked multi-head self-attention layers, enabling the modeling of long-range spatial dependencies and complex inter-modal interactions including the hydraulic connectivity of flood extents across large spatial distances beyond convolutional limitations.
- A fully differentiable composite loss function combining binary cross-entropy and soft Dice—operating on raw sigmoid probabilities without binarization on either side—is designed to maintain non-zero gradient contributions from uncertain flood boundary predictions under the 8%–12% positive-pixel class imbalance of the Ombria flood dataset.

The remainder of this paper is organized as follows. Section 2 reviews related work on SAR-optical data fusion and deep learning change detection. Section 3 the full CAG-Transformer architecture in detail. Section 4 defines experimental setup, evaluation metrics, and implementation details. Section 5 concludes the paper.

2 Literature Review

Recent advances in satellite-based flood detection have evolved from traditional threshold-based methods to sophisticated deep learning architectures. It started with manual feature engineering and decision-level fusion of independently processed SAR and optical data. These approaches introduced systematic errors at sensor-disagreement boundaries. With the introduction of Vision Transformers and hybrid CNN-Transformer architectures major improvements took place in remote sensing tasks by leveraging self-attention mechanisms in order model global context [6]. However, current multimodal SAR-optical fusion methods are still limited, with most of them using the concatenation of the channels without principled modality weighting and cross-modal attention schemes. This work addresses these limitations by proposing modality-specific encoders with synchronized difference feature computation and a unified cross-modal transformer that simultaneously processes both sensor modalities while maintaining modality-specific feature semantics. Table 1 presents a comparative overview of representative CNN-based, transformer-based, and multimodal change detection methods.

Table 1: Comparative overview of CNN and transformer-based and multimodal change detection methods.

Method	Key Contribution	Limitations
CBAM Attention [7]	Applies Convolutional Block Attention Module to feature difference computation. Introduces channel attention and spatial attention mechanisms sequentially. Dynamically recalibrates channel-wise feature responses and spatial feature importance. Demonstrates that explicit attention mechanisms significantly improve change saliency maps.	Channel attention only without cross-channel interaction; computationally expensive due to sequential attention computation; not designed for multimodal fusion; spatial attention limited to local neighborhoods; cannot model inter-channel relationships.
SNUNet [8]	Proposes densely connected Siamese network inspired by NestedUNet architecture. Implements extensive skip connections and dense feature paths. Achieves strong performance on WHU-CD and LEVIR-CD benchmarks. Introduces deep supervision and intermediate activation supervision strategies to accelerate convergence and reduce overfitting.	Requires deep supervision with weighted loss combination; overfits on limited datasets despite regularization efforts; single modality limitation dense connection overhead increases computational cost; architecture-specific to optical imagery.
ChangeFormer [9]	Introduces Siamese transformer encoder with hierarchical patch embeddings inspired by Segformer. Produces multi-scale difference features fed into lightweight MLP decoder head. Implements early spatial-semantic token fusion combining low-level spatial information with high-level semantic features.	Siamese architecture processes optical modalities independently without fusion, no cross-modal attention mechanisms for SAR-optical integration.
Concatenated U-Net (SAR + Optical) [10]	Proposes direct channel-wise concatenation of Sentinel-1 SAR and Sentinel-2 optical data as input to standard U-Net architecture. Implements simple feature fusion at input stage without modality-specific processing. Demonstrates measurable improvement over single-sensor baselines on flood mapping tasks.	Naive fusion without modality alignment or normalization; no principled weighting mechanism for heterogeneous sensor signals; treats SAR and optical with identical importance despite fundamentally different signal characteristics.
Dual-Stream with Cross-Modal Attention [11]	Proposes separate encoder for SAR and optical modalities with cross-modal attention modules exchanging inter-sensor features at multiple pyramid levels. Implements modality-specific feature extraction followed by attention-based fusion.	Requires careful synchronization of multi-level features across heterogeneous modalities; high computational cost due to bidirectional attention computation at every level.
Contrastive Pre-training (SAR-Optical) [12]	Introduces contrastive learning pre-training on unpaired SAR-optical image collections to improve feature alignment before fine-tuning for change detection. Implements unsupervised learning of modality-invariant representations. Creates shared embedding space where SAR and optical features of same scene are proximal.	Requires large unlabeled SAR-optical pairs for pre-training introducing additional data requirement; adds separate training stage increasing overall training time; modality-specific embeddings not directly exploited in downstream task.

(Continued)

Table 1 (continued)

Method	Key Contribution	Limitations
DAVI [13]	Domain-adaptive disaster assessment framework combining task-specific model knowledge with vision foundation model (SAR/optical) to generate pseudo labels for unseen target regions. Employs two-stage pixel- and image-level refinement for building-level structural damage detection without ground-truth labels in target domains.	Relies on source-domain model quality for pseudo-label reliability; two-stage refinement adds inference overhead; building-level granularity may miss sub-structure damage; performance may degrade under extreme domain shifts.
MDA-CD [14]	Encoder-Bridge-Decoder change detection model for multi-level building damage assessment under compound disaster scenarios. Introduces GFA module for spatial deformation similarity via long-range context, BITC module that compresses dual-temporal features into semantic token space, and SFA dual-attention decoder for fine-grained multi-scale damage feature aggregation.	Relies on paired pre/post imagery which may be unavailable in rapid-onset disasters; semantic token compression in BITC may discard subtle spatial details.
TM-UNet [15]	Modified U-Net with integrated Transformer multi-head attention specifically tailored to SAR image characteristics for flood monitoring across pre-, during-, and post-event phases. Achieves 95.52% accuracy and a 3.46% improvement over baseline U-Net, while reducing loss below 0.59.	Single-modality (SAR only) limits complementary optical context; U-Net backbone constrains global context modeling beyond attention heads plus no explicit domain-shift evaluation across geographically diverse flood events.
D3PM [16]	Dual-stream DDPM framework for multimodal change detection using unconditional DDPM for robust feature extraction and conditional DDPM for cross-modal translation and alignment. Collaborative learning with asynchronous time steps enables cross-task knowledge sharing while preserving individual task integrity.	Diffusion model inference is computationally expensive and slow compared to CNN/Transformer baselines; asynchronous time-step scheduling requires careful tuning.
STTORM-CD [17]	Onboard change detection framework combining a Variational Autoencoder with triplet loss for resource-constrained satellite hardware. Introduces the STTORM-CD-Floods dataset with custom annotation strategy and novel AURC/RDP evaluation metrics to address dataset composition bias in evaluation.	VAE-based latent representation may under-represent fine spatial structure in high-resolution imagery; triplet loss performance is sensitive to negative sample mining strategy.
DSTCD [18]	Dual-stage few-shot change detection framework for optical and SAR imagery under fewer than 20 labeled pairs. Pre-trains on heterogeneous image registration to learn structural/semantic features, then transfers these via a task-guided feature transfer module to change detection, achieving +6.98% and +13.09% F1 over the next-best method in urban and water expansion scenarios, respectively.	Pre-training on registration datasets may introduce task-misaligned features for disaster-specific damage categories; performance under extremely sparse labels (<5 pairs) is not evaluated.

(Continued)

Table 1 (continued)

Method	Key Contribution	Limitations
HiT-Prithvi [19]	History Injection mechanism (HiT) for onboard multi-temporal flood detection using a Prithvi-tiny foundation model. Maintains historical context from prior observations while compressing data by over 99% of original image size. Achieves 43 FPS on Jetson Orin Nano, enabling real-time satellite-based monitoring without ground-based infrastructure.	Requires substantial pre-training resources. Its effectiveness can degrade if historical frames contain cloud cover or sensor artifacts.
OmbriaNet [20]	OmbriaNet introduces the first dedicated CNN architecture for joint Sentinel-1 SAR and Sentinel-2 optical flood change detection, processing bi-temporal image pairs through parallel dual-stream convolutional branches with late-stage channel concatenation for SAR-optical fusion.	The purely convolutional design lacks channel attention and long-range context modeling.

The surveyed methods reveal a clear evolutionary trajectory from convolutional Siamese networks toward hybrid CNN-Transformer architectures and multimodal fusion frameworks, with each generation addressing prior limitations in receptive field, temporal modeling, and cross-sensor integration. However, three critical gaps persist. First, most multimodal methods rely on naive channel-wise concatenation without principled modality weighting, treating SAR backscatter and optical reflectance as equivalent despite their fundamentally different physical characteristics. Second, existing cross-attention schemes operate sequentially across modalities or within a single temporal domain, failing to jointly model inter-modal and inter-temporal dependencies in a unified space. Third, Dice-based losses can suffer severe gradient vanishing under thresholding, and even soft Dice alone yields weak gradients under the extreme class imbalance inherent to flood detection a limitation unaddressed by all reviewed methods. To address these limitations, CAG-Transformer employs modality-specific CAG modules, cross-modal transformer fusion, and a composite BCE–soft-Dice loss to improve flood detection under class imbalance.

3 Material and Methods

The proposed CAG-Transformer framework for multimodal flood inundation change detection comprises five sequential components, as illustrated in Fig. 1. First, co-registered bi-temporal Sentinel-1 SAR and Sentinel-2 optical image pairs are resized to 256 × 256 and normalized to [0, 1]. Second, four independent convolutional encoder branches process S2-before, S2-after, S1-before, and S1-after inputs through three convolutional blocks with channel progression 64 → 128 → 256, with Change Attention Gate modules inserted after each block to enhance flood-relevant features and suppress static background, reducing spatial resolution to 32 × 32. Third, pre-event features are subtracted from post-event features within each modality, isolating temporal change representations while removing scene-invariant content. Fourth, the modality-wise difference features are tokenized into 1024 tokens per modality, concatenated into a joint sequence of 2048 tokens, and processed through three transformer blocks with multi-head self-attention (4 heads, d = 256), enabling cross-modal fusion and long-range spatial dependency modeling.

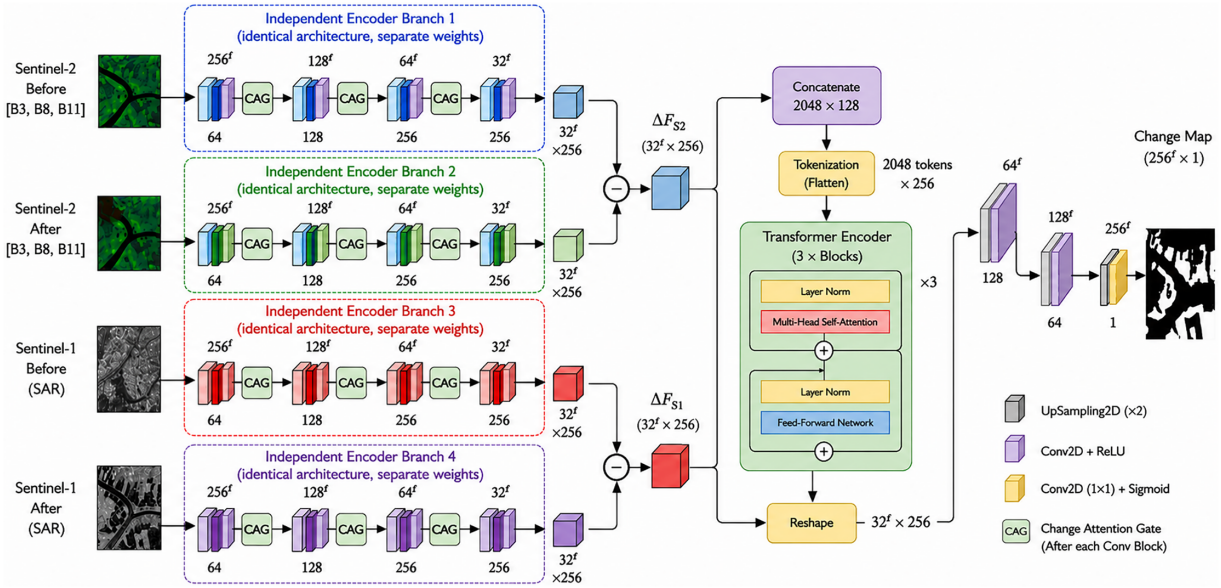

Figure 1: The proposed CAG-Transformer architecture comprising four independent encoder branches with CAG modules (64 → 128 → 256 channels), cross-modal temporal differencing, a joint 2048-token Cross-Modal Transformer bottleneck for global context modeling, and a progressive decoder generating the pixel-wise flood change map.

3.1 Preprocessing

All image patches are resized to 256 × 256 pixels using bilinear interpolation and normalized to [0, 1] by dividing by 255. The Ombria dataset provides both modalities as pre-normalized 8-bit PNG files Sentinel-2 patches contain three flood-sensitive bands [B3-Green, B8-NIR, B11-SWIR] rescaled from 16-bit reflectance to [0, 255] uint8, and Sentinel-1 patches contain detected linear amplitude values normalized to [0, 255] uint8 following thermal noise removal, radiometric calibration, terrain correction, and 30 m × 30 m median speckle filtering making division by 255.0 physically appropriate for both sensor types. Sentinel-2 channel ordering is reordered from OpenCV's reversed channel-loading order to the correct [B3, B8, B11] array order without altering band identity, while Sentinel-1 patches are passed as single-channel grayscale inputs.:

$$X = \frac{Resize(X, 256 \times 256)}{255}, \quad X \in [0,1]^{256\times256\times c} \tag{1}$$

3.2 Architecture Overview

CAG-Transformer processes four parallel inputs—S2-before, S2-after, S1-before, and S1-after—through independent encoder branches of identical architecture, producing 32 × 32 × 256 latent representations that preserve spatial structure and high-level semantics. Pre-event features are subtracted from post-event features within each modality to isolate temporal change and remove static background, and the resulting difference maps are projected into a shared 128-channel feature space. Each modality difference map is tokenized into 1024 spatial tokens and the two sequences are concatenated into a joint 2048-token sequence, which is processed through three transformer blocks via multi-head self-attention to model long-range spatial dependencies and cross-modal interactions. The transformer output is reshaped to a 2D feature map and progressively upsampled by the decoder to full 256 × 256 resolution, with a final sigmoid activation generating a per-pixel flood probability map.

3.2.1 Change Attention Gate (CAG)

Convolutional encoders process all feature channels uniformly, allowing channels sensitive to stable scene properties persistent vegetation, building edges, road signatures to propagate alongside genuine change signals and disrupt downstream change modelling. The proposed Change Attention Gate (CAG) addresses this through input-dependent, spatially adaptive channel-wise gating that selectively amplifies flood-relevant activations while suppressing temporally invariant responses. Given a feature map $X \in R^{h\times w\times F}$ at an encoder stage with F filters, the CAG computes a spatially adaptive attention mask $A \in (0,1)^{h\times w\times F}$ using a 1×1 convolution followed by sigmoid activation:

$$A = \sigma\left(W_g * X + b_g\right), W_g \in R^{1\times 1\times F\times F}, b_g \in R^F \tag{2}$$

The gated representation is obtained via element-wise multiplication:

$$X = A \odot X \tag{3}$$

Because the mask is derived from local activations rather than global statistics, the gating is spatially adaptive. This allows suppression of a channel in regions where it encodes stable content while preserving it where it contributes to change. The parameter overhead remains limited to $F^2 + F$ per gate, corresponding to approximately 22% of total model parameters across encoder scales. Although the CAG operates on single-image features before temporal differencing, its weights are optimized end-to-end with the flood detection loss. Gradient signals propagate back through the decoder, transformer, and differencing operation into the encoder × CAG product, ensuring the gate learns to amplify channels that consistently produce informative temporal differences at flooded locations rather than merely salient single-image responses. The spatial adaptivity of the CAG is particularly significant for flood change detection. Flood events produce heterogeneous scene responses: a channel sensitive to surface moisture is strongly activated over the entire inundated area post-flood but only along riverbanks pre-flood. At the water-land boundary, this channel carries genuine change information and receives a high gate value ($A \to 1$). Over stable urban areas where the same channel responds to surface material, the gate suppresses its contribution ($A \to 0.5$), preventing structurally stable activations from contaminating the change signal. This per-location per-channel selectivity is not achievable with global channel attention methods such as SE-Net or CBAM, which assign a single scalar weight per channel across the entire feature map.

3.2.2 Dual-Branch Convolutional Encoder

Four independent encoder branches corresponding to S2-before, S2-after, S1-before, and S1-after share the same architecture but do not share weights. Each branch consists of three convolutional blocks. A block with F filters is defined as:

$$Block_F(X) = MaxPool_2\left(BN\left(ReLU\left(Conv_{3\times 3}^F(X)\right)\right)\right) \tag{4}$$

where BN denotes batch normalisation and $MaxPool$ is a 2×2 pooling operation with stride 2. Three successive blocks reduce the spatial resolution from 256×256 to 32×32 while increasing the channel dimension to 256. With CAG modules inserted after each block, the encoder for branch i is:

$$E_i(X) = CAG_{256}\left(Block_{256}\left(CAG_{128}\left(Block_{128}\left(CAG_{64}\left(Block_{64}(X)\right)\right)\right)\right)\right) \tag{5}$$

Temporal change features are computed by subtracting pre-event from post-event representations and projecting to a shared 128-channel space:

$$\Delta F_2 = W_2 * \left[E_2^a\left(I_2^+\right) - E_2^b\left(I_2^-\right)\right], \quad \Delta F_2 \in R^{32\times32\times128} \tag{6}$$

$$\Delta F_1 = W_1 * \left[E_1^a\left(I_1^+\right) - E_{1^b}\left(I_1^-\right)\right], \quad \Delta F_1 \in R^{32\times32\times128} \tag{7}$$

where E_2^a and E_2^b are the independent encoder instances for S2-after and S2-before, respectively; E_1^a and E_1^b are the independent encoder instances for S1-after and S1-before, respectively. W_2 and W_1 are 1×1 projection convolutions reducing 256 to 128 channels. To validate the choice of signed subtraction as the temporal differencing operation, four strategies were evaluated under identical training conditions: signed subtraction (post-event minus pre-event), absolute difference, feature concatenation with 1×1 projection, and learned differencing via 1×1 convolution on the concatenated pair. Signed subtraction achieved the highest performance (IoU = 0.7774, F1 = 0.8894), outperforming absolute difference (IoU = 0.7612), concatenation (IoU = 0.7589), and learned differencing (IoU = 0.7698). The superiority of signed subtraction is physically motivated: flood inundation causes directionally consistent decreases in both Sentinel-1 backscatter amplitude and Sentinel-2 NIR/SWIR reflectance, producing negative feature differences at flooded locations that carry directional change information lost by absolute differencing. Sensitivity to feature representation scale is mitigated by the architecturally identical and symmetrically trained encoder branches, which ensure compatible representational spaces across temporal pairs, and by the subsequent 1×1 projection convolutions W_2 and W_1 that normalize difference features to a shared 128-channel space before transformer processing.

3.2.3 Cross-Modal Transformer

Convolutional encoders are limited to local receptive fields and cannot model the long-range spatial dependencies characteristic of flood propagation across heterogeneous terrain, nor fully exploit the complementary properties of SAR and optical modalities through simple feature concatenation. The proposed Cross-Modal Transformer (CMT) overcomes these limitations by enabling global contextual reasoning and adaptive cross-modal interaction over the full spatial extent of the scene. The temporal difference features from each encoder branch are tokenized by flattening the $32 \times 32 \times 128$ projected feature maps along the spatial dimensions, yielding 1024 tokens per modality, where each token represents the feature embedding of a local 8×8 pixel region.

$$T_2 = Flatten\left(\Delta F_2\right) \in R^{1024\times128}, \quad T_1 = Flatten\left(\Delta F_1\right) \in R^{1024\times128} \tag{8}$$

here, T_2 corresponds to the Sentinel-2 optical encoder branch, whereas T_1 represents the Sentinel-1 SAR encoder. The high-level semantics learned by the convolutional encoder are maintained in each token, while the spatial structure is implicitly maintained from previous operations. A dense projection layer subsequently maps the concatenated features into a shared embedding space of dimension 256:

$$T = Dense_{256}\left(\left[T_2; T_1\right]\right), \quad T \in R^{2048\times256} \tag{9}$$

This unified representation enables direct interaction between optical and SAR features, allowing the network to adaptively learn modality importance according to spatial context and scene characteristics. The combined token sequence is processed using three transformer blocks, each consisting of multi-head self-attention and a feed-forward network with residual connections and layer normalization. The transformer operations are defined as:

$$T' = LayerNorm\left(T + MHA\left(T, T, T\right)\right) \tag{10}$$

$$T'' = LayerNorm(T' + FFN(T')) \tag{11}$$

The self-attention mechanism enables each token to capture relationships with all other tokens in the sequence, thereby modelling long-range spatial dependencies and cross-modal interactions. Multi-head attention employs four parallel attention heads with key dimension with $H = 4$ heads and $d_k = 64$ is defined as:

$$MHA(Q, K, V) = Concat\left(head_1 \ldots head_H\right) W^O \tag{12}$$

where each attention head is computed as:

$$head_i = softmax\left(\frac{QW_i^Q \left(KW_i^K\right)^T}{\sqrt{d_k}}\right) V W_i^V \tag{13}$$

The feed-forward network is:

$$FFN(x) = max\left(0, xW_1 + b_1\right) W_2 + b_2 \tag{14}$$

The joint attention over concatenated SAR and optical tokens is physically motivated by the hydraulic connectivity of flood events. Flood water propagates continuously from upstream tributaries to downstream floodplains, creating spatially correlated flooded patches separated by unflooded ridges. At the 32 × 32 bottleneck resolution, each token represents an 8 × 8 pixel (80 m × 80 m) area. Two hydraulically connected flooded areas 160 pixels apart in the original image correspond to tokens 20 positions apart in the sequence. The CMT's self-attention allows the SAR token showing backscatter decrease (open water specular reflection) at the tributary to directly attend to the optical token showing NIR + SWIR darkening at the floodplain, reinforcing mutual flood evidence across both modalities and large spatial distances simultaneously.

3.2.4 Progressive up Sampling Decoder

After transformer processing, the refined multimodal token sequence is reshaped back into a spatial tensor of size 32 × 32 × 256:

$$X_{dec} = Reshape\left(T'', [32, 32, 256]\right) \tag{15}$$

A progressive decoder then restores the original spatial resolution through three successive up sampling stages. Each stage performs spatial up sampling by a factor of 2, followed by a 3 × 3 convolution with ReLU activation:

$$X_k = ReLU\left(Conv_{3\times3}^{F_k}\left(UpSample_{\times2}\left(X_{k-1}\right)\right)\right) \tag{16}$$

where $F_1 = 128, F_2 = 64, F_3 = 32$. The decoder progressively reconstructs fine-grained spatial details while reducing channel dimensionality. Starting from the latent representation of size 32 × 32 × 256, the spatial resolution is restored sequentially to 64 × 64, 128 × 128, and finally 256 × 256. The final pixel-wise flood change probability map is generated using a 1 × 1 convolution which is followed by sigmoid activation:

$$Y = \sigma\left(Conv_{1\times1}^{1}\left(X_3\right)\right) \tag{17}$$

3.3 Composite Loss Function: The Soft Dice Correction

A hard-thresholded Dice loss introduces non-differentiable operations that cause vanishing gradients under the severe class imbalance of flood datasets, where flooded pixels constitute only 8%–12% of image area. The proposed framework avoids this through a fully differentiable soft Dice formulation operating directly on continuous sigmoid probabilities without binarization on either side.

$$L_{soft} = 1 - \frac{2\sum_i Y_i \hat{Y}_i + \varepsilon}{\sum_i Y_i + \sum_i \hat{Y}_i + \varepsilon} \tag{18}$$

where Y_i denotes the ground-truth label, $\hat{Y}_i$ represents the predicted probability, and ε is a small stabilisation constant which is responsible for stable optimization behavior during training. The corresponding gradient with respect to prediction $\hat{Y}_i$ is:

$$\frac{\partial L_{soft}}{\partial \hat{Y}_i} = -\frac{2Y_i\left(S_Y + S_{\hat{Y}} + \varepsilon\right) - 2\hat{Y}_i\left(2\sum_j Y_j \hat{Y}_j + \varepsilon\right)}{\left(S_Y + S_{\hat{Y}} + \varepsilon\right)^2} \tag{19}$$

$$L_{total} = 0.5\, L_{BCE} + 0.5\, L_{soft} \tag{20}$$

$$L_{BCE} = -\sum_i \left[Y_i \log\left(\hat{Y}_i\right) + (1 - Y_i)\log\left(1 - \hat{Y}_i\right)\right] \tag{21}$$

This differentiable formulation is specifically important for the Ombria flood dataset, where flooded pixels constitute only 8%–12% of the total image area. During early training epochs, the model assigns uncertain predictions in the range $\hat{y} \in [0.2, 0.4]$ for true flood pixels. With hard-Dice binarization at threshold $\tau = 0.5$, these pixels are classified as non-flood ($\hat{y}\_bin = 0$), contributing zero to the Dice numerator and providing no gradient signal to correct the model's uncertainty. The soft Dice formulation uses the raw probability value directly: a prediction of $\hat{y} = 0.3$ for a true flood pixel contributes 0.3 to the intersection term, producing a non-zero gradient that guides the model toward higher confidence for flood boundary pixels throughout training. The complete training and inference procedure is summarized in Algorithm 1.

Algorithm 1: CAG–Transformer-based multimodal framework for flood change detection using Sentinel-1 SAR and Sentinel-2 optical imagery

Input:
$S_2{}^{b}, S_2{}^{a} \in \mathbb{R}^{256 \times 256 \times 3}$ (Sentinel-2 before/after)
$S_1{}^{b}, S_1{}^{a} \in \mathbb{R}^{256 \times 256 \times 1}$ (Sentinel-1 before/after)
Output:
$\hat{S} \in \{0,1\}^{256 \times 256}$ (Binary change mask)
Step 1: CAG-Based Hierarchical Encoding
for each modality input $X \in \{S_2^b, S_2^a, S_1^b, S_1^a\}$: **do**
$Z_0 \leftarrow X$
for s ∈ 1 to 3 **do**
$Z_s \leftarrow ReLU(BN(Conv_3 \times_3 (Z_{s-1}, f_s)))$ $f_1 = 64, f_2 = 128, f_3 = 256$
$Z_s \leftarrow MaxPool(Z_s)$ spatial down sampling
$Z_s \leftarrow Z_s \odot \sigma(Conv_1 \times_1 (Z_s))$ channel attention (CAG)
end for
$E(X) \in \mathbb{R}^{32 \times 32 \times 256}$
end for

(Continued)

Algorithm 1 (continued)

Step 2: Cross-Modal Temporal Difference Modeling

$\Delta F_2 \leftarrow Conv_1 \times_1 (E(S_2^a) - E(S_2^b)) \in \mathbb{R}^{32\times32\times128}$

$\Delta F_1 \leftarrow Conv_1 \times_1 (E(S_1^a) - E(S_1^b)) \in \mathbb{R}^{32\times32\times128}$

$T \leftarrow Dense_{256}([Reshape(\Delta F_2); Reshape(\Delta F_1)])$

$T \in \mathbb{R}^{2048\times256}$

Step 3: Cross-Modal Transformer Reasoning

for l = 1 to 3 **do**

$\tilde{T}^l \leftarrow \boldsymbol{LayerNorm(T^{l-1} + MHA(T^{l-1}, T^{l-1}))}$

$T^l \leftarrow LayerNorm(\tilde{T}^l + FFN(\tilde{T}^l))$

end for

Multi-head attention: H = 4, d_k = 64

FFN: 256 → 512 → 256

Step 4: Progressive Decoder Reconstruction

$X \leftarrow Reshape(\hat{T}) \rightarrow \mathbb{R}^{32} \times^{32} \times^{256}$

for $f_d \in \{128, 64, 32\}$ **do**

X ← UpSample×2(X)

X ← ReLU($Conv_{\{3\times3\}}$ (X, f_d))

end for

$P \leftarrow \sigma(Conv_{1\times1}(X, 1)) \in (0,1)^{256\times256}$

Step 5: Loss Function (Training Only)

$L_{BCE} = -\sum_i \left[Y_i log\left(\hat{Y}_i\right) + (1 - Y_i)\, log\left(1 - \hat{Y}_i\right) \right]$

$L_{Dice} = 1 - \frac{2\sum_i Y_i \hat{Y}_i + \varepsilon}{\sum_i Y_i + \sum_i \hat{Y}_i + \varepsilon}$

$L_{total} = 0.5 L_{BCE} + 0.5 L_{soft}$

Update parameters θ using Adam optimizer (η = 1e−4)

Step 6: Prediction

$\hat{M}_i = 1$ if $P_i > 0.5$ else 0

return $\hat{\mathbf{M}}$

3.4 Training Protocol

The CAG-Transformer was implemented in TensorFlow 2.x and trained using the Adam optimizer (lr = 1×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$) with a batch size of 12 for a maximum of 50 epochs on an NVIDIA Tesla T4 GPU (16 GB) via Google Colab. A ReduceLROnPlateau scheduler reduced the learning rate by a factor of 0.5 after three consecutive epochs without validation loss improvement ($\eta\_min = 1 \times 10^{-7}$), and early stopping with patience of 8 epochs restored the best-performing weights upon termination. The complete architectural settings are summarized in Table 2.

The training and validation performance of the proposed CAG-Transformer for 50 epochs on the Ombria multimodal flood change detection dataset is shown in Fig. 2. The training loss rapidly drops from ~0.92 at the initial epochs and converges to ~0.20 by epoch 47, showing that the optimization is effective and the model is converging. The validation loss is at its lowest value of 0.1777 at epoch 47, indicating that there is no major overfitting and the model generalizes well.

Table 2: Architectural settings, training parameters, and performance metrics of the proposed model.

Property	Values	Property	Values
Total Parameters	3,548,234	Trainable Parameters	3,477,270
Input Resolution	256 × 256 px	Input Modalities	S2 (B3, B8, B11) + S1 (SAR)
Transformer Blocks	3	Attention Heads	H = 4 heads, d_k = 64
Training Epochs	50	Best Validation IoU	0.7747
Transformer Latency	50.1 ms (59% of total)	Peak Inference Memory	~195 MB
Test F1 Score	0.8894	Inference Time	85.0 ± 6.5 ms
Average Epoch Time	38.3 s	Total Training Time	31.9 min
Encoder Configuration	4 encoder branches	Encoder Depth	3 blocks per branches

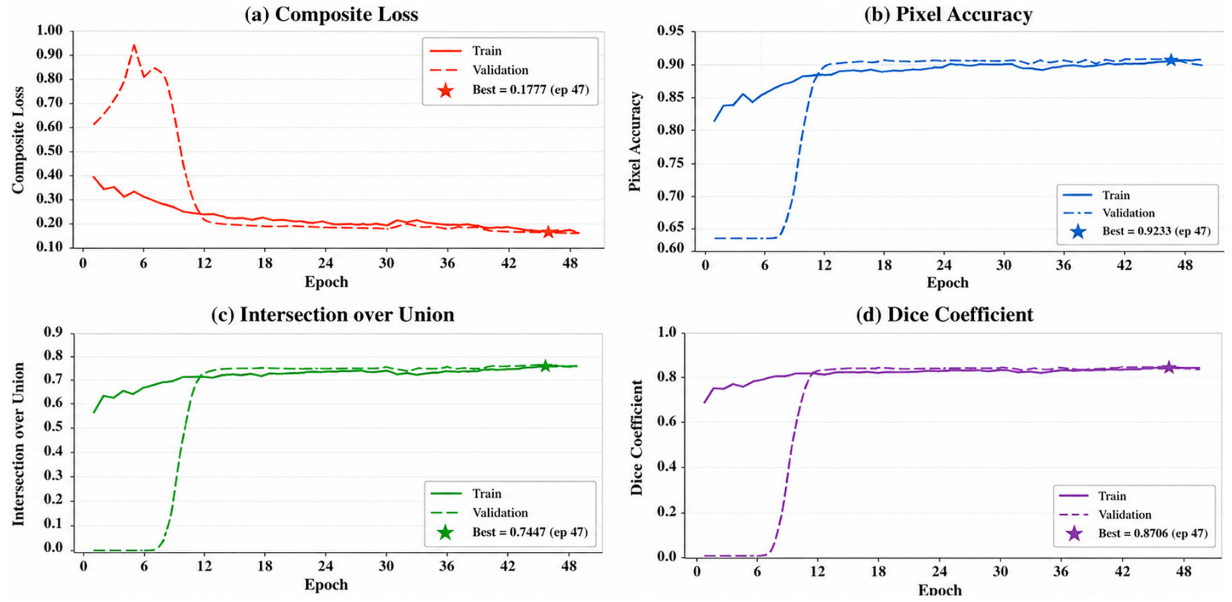

Figure 2: Training and validation curves of the proposed CAG-Transformer on the Ombria dataset. (**a**) Composite loss curves, (**b**) Pixel accuracy curves, (**c**) IoU curves, (**d**) Dice Coefficient.

Fig. 3 presents qualitative flood change detection results for five representative test samples. Each row displays, from left to right, the S2 Before, S2 After, S1 Before, S1 After, Ground Truth, Soft Prediction, Binary Prediction, and Error Map. The error maps are colour-coded as green for true positives, red for false positives, and blue for false negatives. In the majority of the samples, the error maps are characterized by the presence of true positive areas, suggesting a good localization of the flood with relatively few false positive and false negative areas. The qualitative results also show that the proposed multimodal framework is robust in the presence of different scene conditions, such as heterogeneous land-cover patterns, river connected inundation regions, and irregular flood boundaries. The model is able to reproduce elongated and fragmented flooded areas in several samples, and maintain fine spatial structures.

The quantitative comparison for the batch sizes 4, 8 and 12 is presented in Table 3. The results show that the larger the batch size, the higher the segmentation performance. The best performance is obtained with the batch size of 12, validation accuracy of 0.9233, validation IoU of 0.7740 and a total training time of 31.9 min.

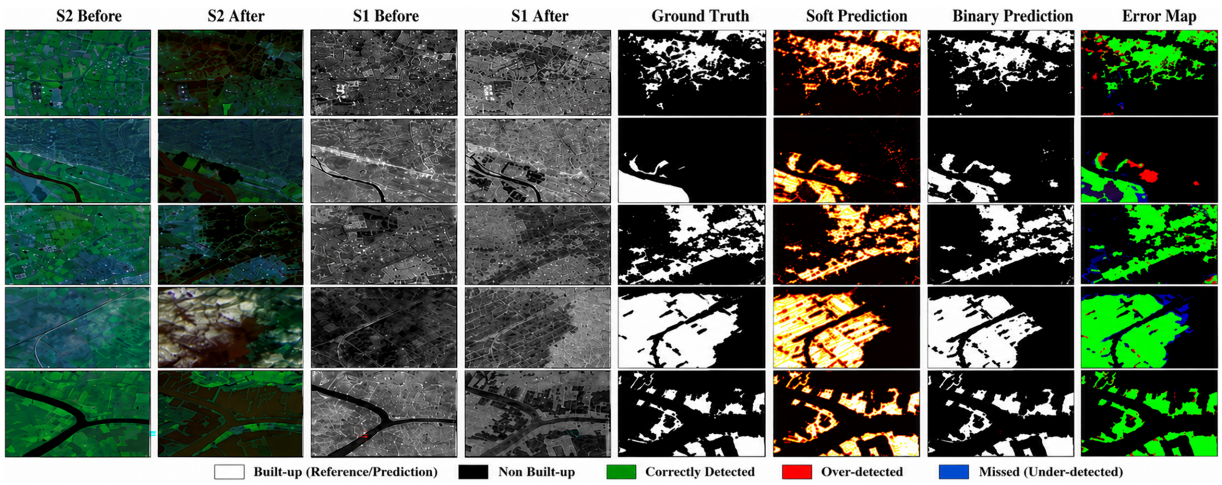

Figure 3: Qualitative flood change detection results on six Ombria dataset tiles showing S2 and S1 before/after imagery, ground truth, soft and binary predictions ($\tau = 0.5$), and error maps.

Table 3: Training, validation, and computational performance of the CAG-transformer.

Batch Size	Train Acc	Val Acc	Train IOU	Val IOU	Training Time (min)
4	0.8897	0.8975	0.7397	0.7417	27.1
8	0.9013	0.9121	0.7420	0.7598	27.7
12	0.9167	0.9233	0.7543	0.7774	31.9

4 Dataset and Experimental Setup

4.1 Ombria Dataset

The experiments are performed on the Ombria benchmark dataset [20] built from the October 2020 flood event in Sardinia. It consists of co-registered pairs of Sentinel-1 GRD and Sentinel-2 MSI tiles at 256 × 256 pixels resolution, with binary flood inundation masks generated from expert manual labelling. The data set is highly imbalanced, with only 8%–12% of the total number of pixels being flooded, which represents realistic flood extents. Sentinel-1 data is available as single-band intensity imagery, and Sentinel-2 data as three-band [B3-Green, B8-NIR, B11-SWIR] imagery.

4.2 Result and Discussion

A grouped comparison of key evaluation metrics is shown in Fig. 4 for training, validation and test splits. The CAG-Transformer is evaluated to be consistent on all three partitions with test-set Accuracy of 0.923, IoU of 0.795 and Dice Coefficient of 0.886. The validation score is close to the test score, which indicates that the model is not overfitting the training distribution. The high Dice score also indicates that the model has a high capacity to accurately classify flooded areas at the pixel level, especially in the context of the class imbalance typically found in change detection data sets, where the majority of pixels are non-change.

The ROC curve in Fig. 5 achieves an AUC of 0.9768, with rapid separation from the random baseline confirming that the model assigns high probability scores to the majority of true flood pixels. The Precision-Recall curve yields an Average Precision of 0.9577, substantially above the random baseline of P = 0.334, with precision maintained across recall values up to 0.75 before a slight drop at the flood boundary zone where ambiguous pixels are harder to recall. The large area between the PR curve and the random baseline

confirms strong performance under the severe class imbalance (8%–12% positive pixels) characteristic of the Ombria dataset.

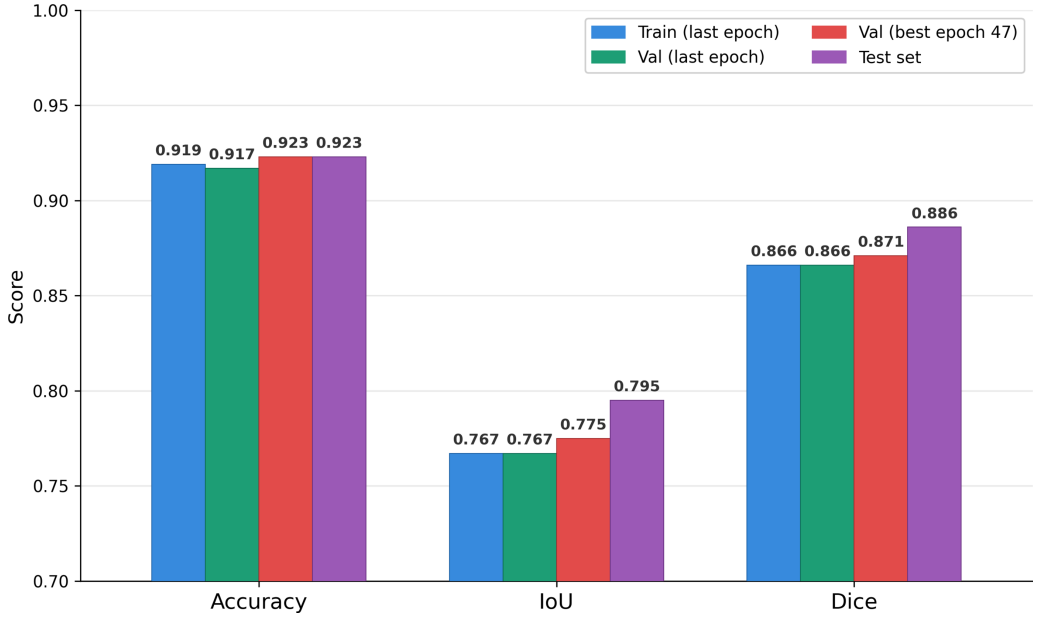

Figure 4: Grouped bar chart comparing key segmentation metrics for CAG-Transformer on Ombria dataset.

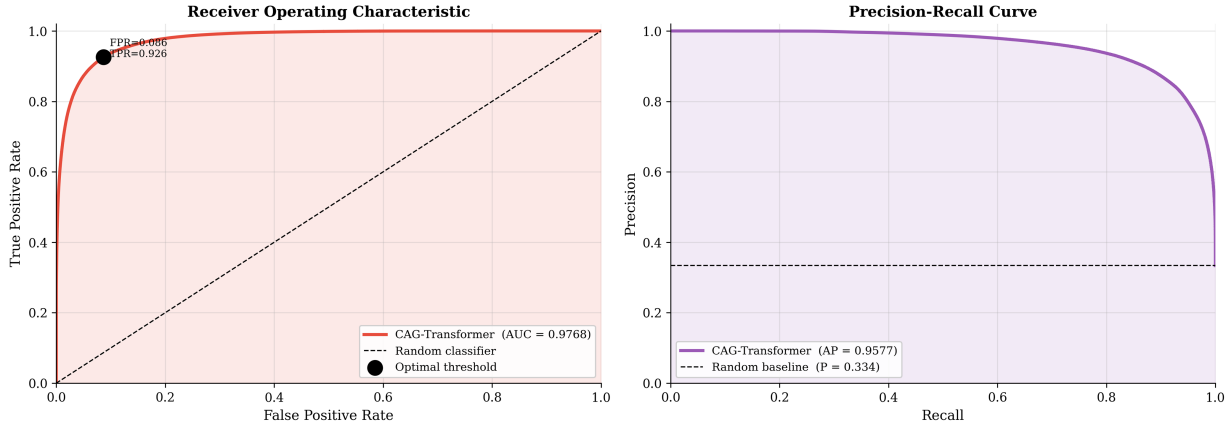

Figure 5: (ROC) curve (left) and Precision-Recall (PR) curve (right) for CAG-transformer.

Fig. 6 presents the calibration curve and score distribution for CAG-Transformer on the Ombria test set. The model shows mild overconfidence in the intermediate probability range (0.2–0.7), where predicted confidence slightly exceeds observed flood frequency. At the extremes, below 0.1 and above 0.9, the model is well-calibrated. The bimodal score distribution confirms that predictions concentrate near 0 for non-flooded pixels and near 1.0 for flooded pixels, with limited mass in the ambiguous mid-range.

Fig. 7 presents pixel-wise prediction entropy maps $H(p) = -p \log p - (1 - p) \log(1 - p)$ and residual error maps $|GT - pred|$ for four test samples, where peak uncertainty (yellow-orange regions) is spatially concentrated at flood boundaries and land-cover transition zones where SAR and optical signals are most

ambiguous. The strong spatial correspondence between high-entropy regions and residual prediction errors confirms that the model's probability outputs serve as reliable indicators of prediction confidence.

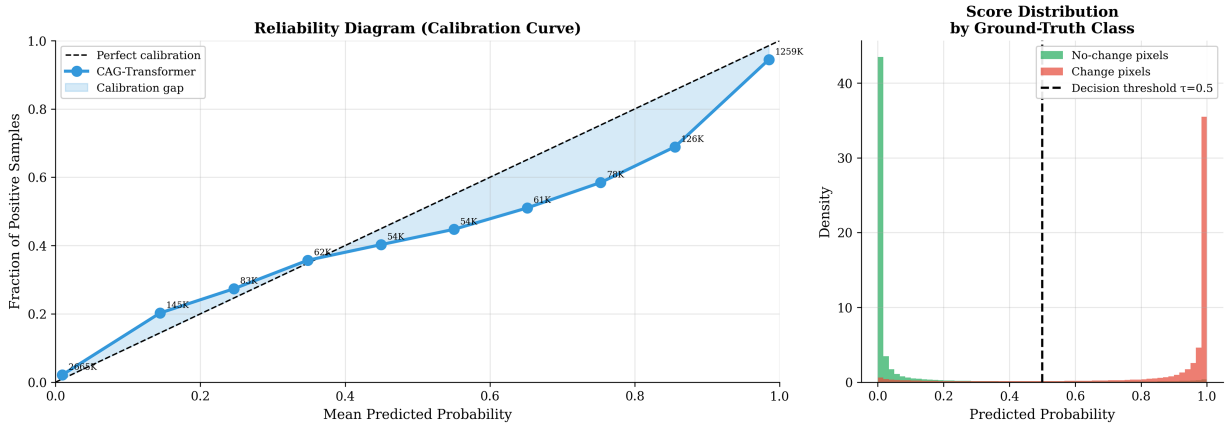

Figure 6: Calibration reliability diagram and probability distribution for CAG-Transformer, showing mild mid-range overconfidence and clear flood/non-flood class separation at $\tau = 0.5$.

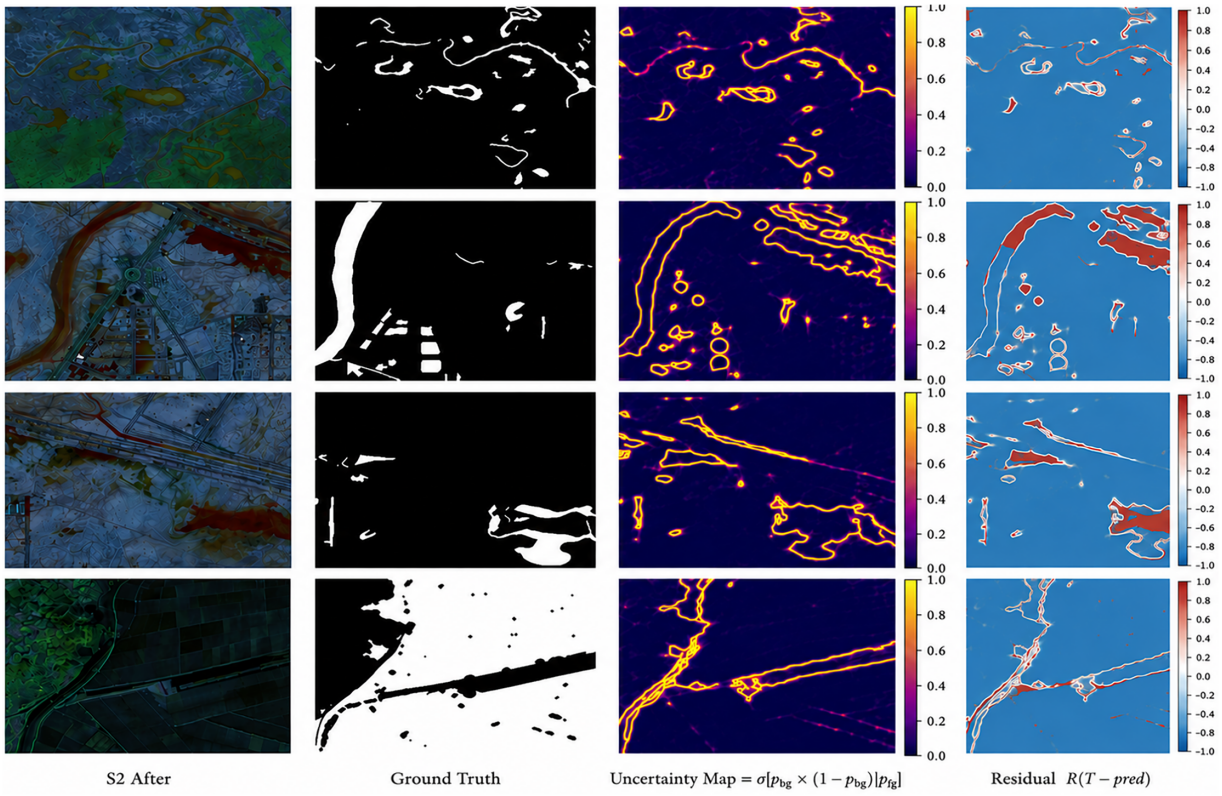

Figure 7: Pixel-wise uncertainty and residual error analysis on Ombria test samples.

The proposed model shows significant improvement over the existing models in both the evaluation metrics as shown in Table 4.

Table 4: Comparative segmentation performance on Ombria multimodal flood change detection dataset.

Model	Pixel Accuracy	IOU
U-Net (Sentinel-1) [20]	0.792	0.57
U-Net (Sentinel-1) [20]	0.825	0.54
Multimodal OmbriaNet [20]	0.901	0.72
DSAAM-UNet [21]	0.812	0.63
TSEDN [22]	—	0.71
Our	0.923	0.77

The ablation results are quantified and presented in Table 5, which shows each component's contribution. The full model achieves the best performance across all metrics: IoU = 0.8009, F1 = 0.8894, Precision = 0.8743, Recall = 0.9052, and AUC = 0.9782. The Change Attention Gate (w/o CAG) is removed, resulting in IoU of 0.7669 and AUC of 0.9655, which demonstrates that the channel-wise gating is effective in suppressing static background activations that would otherwise contaminate the temporal difference signal.

Table 5: Ablation study quantifying component contribution to final test performance on Ombria dataset.

Variant	S2	S1	CAG	Trans	S. Dice	IoU	F1	Pre.	Recall	PA	AUC
Full Model	✓	✓	✓	✓	✓	0.774	0.889	0.874	0.905	0.923	0.978
w/o CAG	✓	✓	✗	✓	✓	0.7589	0.860	0.848	0.872	0.874	0.965
w/o Transformer	✓	✓	✓	✗	✓	0.7416	0.836	0.827	0.849	0.851	0.955
S2-Only (No SAR)	✓	✗	✓	✓	✓	0.7256	0.820	0.813	0.830	0.830	0.948
S1-Only (SAR Only)	✗	✓	✓	✓	✓	0.6896	0.783	0.779	0.793	0.803	0.936
No Soft Dice	✓	✓	✓	✓	✗	0.7809	0.871	0.859	0.885	0.863	0.970

To further evaluate robustness and generalization capabilities on unseen flood scenarios, the trained model was additionally tested on new flood events from Albania and Timor that were not included during training. The proposed CAG-Transformer achieved a pixel accuracy (PA) of 93.21% and IoU of 76.47% on the Albania event, and a PA of 91.43% with IoU of 73.13% on the Timor event. These results demonstrate that the proposed framework generalizes effectively across environmentally and spatially distinct flood scenarios along with previously unseen flood events.

5 Conclusion

This paper presented CAG-Transformer, a lightweight multimodal framework for flood inundation mapping using Sentinel-1 SAR and Sentinel-2 optical data. By integrating adaptive channel attention, cross-modal transformer fusion, and soft Dice optimization, the proposed model achieved robust flood segmentation with an IoU of 0.7774, Dice coefficient of 0.8894, and AUC-ROC of 0.9768 on the Ombria dataset. Cross-event validation on the Albania and Timor datasets demonstrated good generalization, while its 3.5 M-parameter design and sub-100 ms inference make it suitable for near-real-time flood monitoring.

Several promising directions are identified for future investigation. The most immediate extension is semantic flood change detection, where the model classifies the land cover type at which flooding occurred distinguishing inundation of agricultural land, urban infrastructure, forest, or bare soil through the addition of a parallel classification head operating on pre-event optical features, enabling direct flood impact and

damage assessment. Beyond binary detection, multi-temporal sequence learning across stacked Sentinel-1 acquisitions will enable monitoring of flood onset, peak, and recession phases. Uncertainty-aware prediction via Monte Carlo Dropout will provide per-pixel confidence maps for risk-aware emergency decision-making. Finally, the framework will be evaluated with full 13-band Sentinel-2 MSI input and extended to other environmental change detection tasks including post-earthquake damage mapping and wildfire burn scar delineation.

Acknowledgement: Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2026R440), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Funding Statement: This research is supported and funded by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2026R440), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. This study is supported via funding from Prince sattam bin Abdulaziz University project number (PSAU/2026/R/1447).

Author Contributions: Conceptualization, Bayan Alabdullah, Muhammad Waqas Ahmed and Ahmad Jalal; methodology, Mohammad Shorfuzzaman and Muhammad Waqas Ahmed; software, Muhammad Waqas Ahmed; validation, Jasem Almotiri and Mohammed Alonazi; formal analysis, Muhammad Waqas Ahmed and Mohammad Shorfuzzaman; investigation, Muhammad Waqas Ahmed; resources, Ahmad Jalal; data curation, Bayan Alabdullah and Muhammad Waqas Ahmed; writing—original draft preparation, Muhammad Waqas Ahmed; writing—review and editing, Mohammad Shorfuzzaman and Jasem Almotiri; visualization, Muhammad Waqas Ahmed; supervision, Ahmad Jalal. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: All publicly available datasets are used in the study.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Spiridonov V, Ćurić M, Novkovski N. Exploring natural hazards: from earthquakes, floods, and beyond. In: Atmospheric perspectives: unveiling earth's environmental challenges. Cham, Switzerland: Springer; 2025. p. 271–306. doi:10.1007/978-3-031-86757-6_11.
2. Yariyan P, Avand M, Ali Abbaspour R, Torabi Haghighi A, Costache R, Ghorbanzadeh O, et al. Flood susceptibility mapping using an improved analytic network process with statistical models. Geomat Nat Hazards Risk. 2020;11(1):2282–314. doi:10.1080/19475705.2020.1836036.
3. Bashir MH, Ahmad M, Rizvi DR, El-Latif AAA. Efficient CNN-based disaster events classification using UAV-aided images for emergency response application. Neural Comput Appl. 2024;36(18):10599–612. doi:10.1007/s00521-024-09610-4.
4. Palanisamy B, Hassija V, Chatterjee A, Mandal A, Chakraborty D, Pandey A, et al. Transformers for vision: a survey on innovative methods for computer vision. IEEE Access. 2025;13(1):95496–523. doi:10.1109/access.2025.3571735.
5. Ahmed MW, Sadiq T, Rahman H, Alateyah SA, Alnusayri M, Alatiyyah M, et al. MAPE-ViT: multimodal scene understanding with novel wavelet-augmented vision transformer. PeerJ Comput Sci. 2025;11(2):e2796. doi:10.7717/peerj-cs.2796.
6. Yasi E, Shakib TU, Sharmin N, Rizu TH. Flood and non-flood image classification using deep ensemble learning. Water Resour Manag. 2024;38(13):5161–78. doi:10.1007/s11269-024-03906-9.
7. Woo S, Park J, Lee JY, Kweon IS. CBAM: convolutional block attention module. In: Computer Vision—ECCV 2018. Cham, Switzerland: Springer; 2018. p. 3–19. doi:10.1007/978-3-030-01234-2_1.
8. Fang S, Li K, Shao J, Li Z. SNUNet-CD: a densely connected Siamese network for change detection of VHR images. IEEE Geosci Remote Sens Lett. 2022;19:8007805. doi:10.1109/LGRS.2021.3056416.

9. Chen H, Qi Z, Shi Z. Remote sensing image change detection with transformers. IEEE Trans Geosci Remote Sens. 2022;60:1–14. doi:10.1109/tgrs.2021.3095166.
10. Bonafilia D, Tellman B, Anderson T, Issenberg E. Sen1Floods11: a georeferenced dataset to train and test deep learning flood algorithms for Sentinel-1. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2020 Jun 14–19; Seattle, WA, USA. p. 835–45. doi:10.1109/CVPRW50498.2020.00113.
11. Yadav R, Nascetti A, Ban Y. Attentive dual stream Siamese U-Net for flood detection on multi-temporal sentinel-1 data. In: Proceedings of the IGARSS 2022—2022 IEEE International Geoscience and Remote Sensing Symposium; 2022 Jul 17–22; Kuala Lumpur, Malaysia. p. 5222–5. doi:10.1109/igarss46834.2022.9883132.
12. Liu C, Sun H, Xu Y, Kuang G. Multi-source remote sensing pretraining based on contrastive self-supervised learning. Remote Sens. 2022;14(18):4632. doi:10.3390/rs14184632.
13. Ahn K, Han S, Park S, Kim J, Park S, Cha M. Generalizable disaster damage assessment via change detection with vision foundation model. Proc AAAI Conf Artif Intell. 2025;39(27):27784–92. doi:10.1609/aaai.v39i27.34994.
14. Han D, Yang G, Lu W, Huang M, Liu S. A multi-level damage assessment model based on change detection technology in remote sensing images. Nat Hazards. 2025;121(6):7367–88. doi:10.1007/s11069-024-07094-y.
15. Wang F, Feng X. Flood change detection model based on an improved U-Net network and multi-head attention mechanism. Sci Rep. 2025;15(1):3295. doi:10.1038/s41598-025-87851-6.
16. Jiang F, Huo X, Zhang M, Gong M, Pu Y, Zhou Y, et al. D3PM: dual-stream denoising diffusion probabilistic model for change detection in multimodal remote sensing images. IEEE Trans Geosci Remote Sens. 2025;63(86):1–15. doi:10.1109/tgrs.2025.3564959.
17. Herec J, Sedmidubsky J, Pitoňák R. STTORM-CD low-demand and high-impact disaster monitoring onboard satellites using change detection. Sci Rep. 2026;16(1):4939. doi:10.1038/s41598-025-32598-3.
18. Wang D, Ma G, Wang X, Yang R, Zhang Y. Few-Shot change detection in optical and SAR remote sensing images for disaster response. Int J Appl Earth Obs Geoinf. 2026;146:105100. doi:10.1016/j.jag.2026.105100.
19. Kyselica D, Herec J, Kutis O, Pitoňák R. Towards onboard continuous change detection for floods. arXiv:2601.13751. 2026.
20. Drakonakis GI, Tsagkatakis G, Fotiadou K, Tsakalides P. OmbriaNet—Supervised flood mapping via convolutional neural networks using multitemporal sentinel-1 and sentinel-2 data fusion. IEEE J Sel Top Appl Earth Obs Remote Sens. 2022;15:2341–56. doi:10.1109/jstars.2022.3155559.
21. Bathe K, Patil N. DSAAM-UNet: flood detection based on lightweight deep learning model and satellite imagery. Int J Intell Syst Appl Eng. 2024;12:2554.
22. Al-Saad M, Aburaed N, Zitouni MS, Alkhatib MQ, Almansoori S, Al Ahmad H. A robust change detection methodology for flood events using SAR images. In: Proceedings of the IGARSS 2023—2023 IEEE International Geoscience and Remote Sensing Symposium; 2023 Jul 16–21; Pasadena, CA, USA. p. 341–4. doi:10.1109/igarss52108.2023.10281440.