DYNAMIC RELIABILITY ROUTED CROSS-ATTENTION NETWORK FOR MULTIMODAL FLOOD MAPPING

Aniket Zope, Ushasi Chaudhuri, Shreyansh Aswal, Swagata Biswas and Shailesh Deshpande
TCS Research, Pune, India
{aniket.zope, ushasi.c, shreyansh.aswal, swagata.biswas, shailesh.deshpande}@tcs.com

***Abstract*—Accurate flood extent mapping is essential for rapid disaster response, yet reliable segmentation remains challenging due to the heterogeneous characteristics of optical and Synthetic Aperture Radar (SAR) observations. Pre-flood optical imagery provides valuable semantic information describing land-cover conditions, whereas post-flood SAR observations capture inundation induced changes under all weather conditions. Existing multimodal approaches often rely on fixed fusion strategies that may not effectively exploit the complementary information available from both modalities. To address this limitation, we propose *Dynamic Reliability-Routed Cross-Attention Network (DR-RCA)*, a multimodal framework for flood segmentation using pre-flood Sentinel-2 and post-flood Gaofen-3 imagery. The proposed framework combines modality-specific feature projection, multi-scale optical-SAR cross-attention, difference-correlation feature modeling and dynamic feature routing to effectively integrate modality-specific and cross-modal information. By jointly exploiting complementary, consistent and conflicting evidence from optical and SAR observations, the proposed approach enables more effective multimodal feature fusion for flood extent mapping. Experimental results on the Sen2GF3Floods benchmark demonstrate that the proposed framework consistently outperforms established segmentation baselines, achieving an F1-score of 92.23% and an mIoU of 92.21%.**

***Index Terms*—Flood extent mapping, multimodal remote sensing, optical–SAR fusion, cross-attention, adaptive routing.**

I. INTRODUCTION

Floods are among the most damaging natural hazards affecting human settlements, agriculture, transportation systems and critical infrastructure[1]. Accurate flood extent mapping is therefore essential for disaster response, damage assessment and risk mitigation. Remote sensing has become a fundamental component of operational flood monitoring because it provides extensive spatial coverage and frequent data acquisition enabling timely assessment of flood dynamics. Unlike image level flood detection, flood segmentation generates pixel-wise inundation maps that enable detailed characterization of affected regions.

From a remote sensing perspective, flood extent mapping can be formulated as a change detection problem in which newly inundated areas are identified from differences between pre-event and post-event observations. Reliable characterization of flood-induced change requires both contextual information describing the baseline landcover and flood observations that directly capture flood impacts, motivating the use of multi-temporal and multimodal remote sensing data.

Optical sensors such as Sentinel-2 provide rich spectral and spatial information describing landcover like vegetation, urban structures, agricultural land and permanent water bodies. In contrast, Synthetic Aperture Radar(SAR) provides all weather, day and night imaging and captures inundation induced changes in surface scattering behaviour. Consequently, combining pre-flood optical imagery with SAR observations acquired during or immediate after flood enables the joint exploitation of baseline land-cover context and flood-induced physical surface changes for more robust flood extent mapping.

Recent studies have demonstrated the effectiveness of deep learning for flood extent extraction from optical and SAR imagery. CNN-based approaches have successfully exploited spatial contextual information for flood segmentation, while multimodal frameworks have shown that integrating SAR and Optical observations improves flood delineation robustness and accuracy [2–4]. Attention mechanisms and feature-fusion networks have further improved flood mapping performance by enabling adaptive integration of complementary input features and contextual information [5]. Recent developments have also explored disaster intelligence derived from textual reports like social media posts, flood related news articles and large language models to improve flood extent mapping during extreme events [6]. Despite these advances, effective fusion of optical and SAR representations remains a challenging problem [7]. The contribution of each modality can vary considerably across a scene due to cloud contamination, land-cover heterogeneity, speckle noise and flood-induced backscatter ambiguity. Nevertheless, most existing approaches rely on feature concatenation, fixed fusion schemes or global attention mechanisms that do not explicitly exploit agreement and disagreement characteristics between modalities.

To address these limitations, we propose ***Dynamic Reliability-Routed Cross-Attention Network (DR-RCA)***, a multimodal flood segmentation framework that combines modality-aware feature projection, multi-scale optical–SAR cross-attention, difference-correlation feature modeling and dynamic feature routing. By explicitly modeling complementary, consistent and conflicting information between modalities, the proposed framework enables more effective multimodal fusion for flood extent mapping. Experimental evaluation on the Sen2GF3Floods benchmark [8] demonstrates accurate and robust flood segmentation across diverse flood

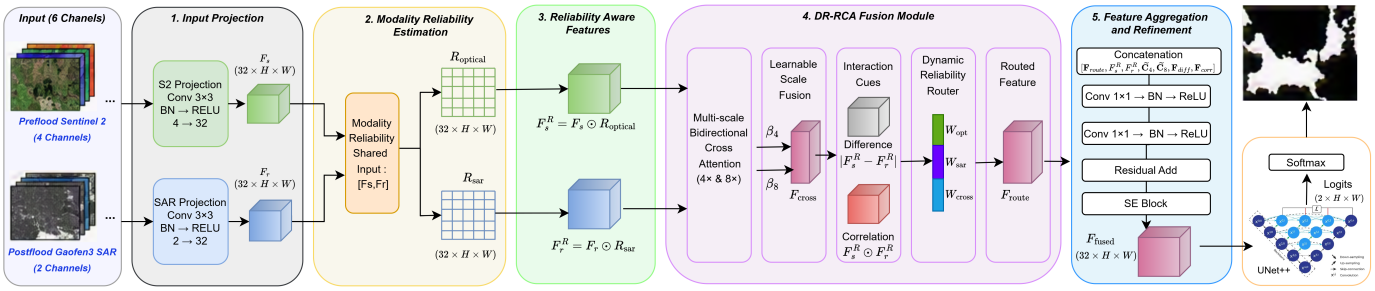


Fig. 1. Overview of the proposed Dynamic Reliability-Routed Cross-Attention UNet++ (DR-RCA) framework. Reliability-aware optical and SAR features are generated through modality-specific reliability estimation and fused using the proposed Dynamic Reliability Router (DRR) and multi-scale cross-attention. The fused representation is refined and decoded by a UNet++ with SCSE attention to produce the final flood segmentation map.

scenarios and land-cover conditions.

II. Problem Formulation

Let

$$\mathbf{X}^s \in \mathbb{R}^{4\times H\times W}, \qquad \mathbf{X}^r \in \mathbb{R}^{2\times H\times W}$$

denote the pre-flood Sentinel-2 optical image (R, G, B and NIR bands) and during or immediate after flood Gaofen-3 dual-polarization (HH and HV) SAR image respectively. The multimodal input is represented as,

$$\mathbf{X} = [\mathbf{X}^s, \mathbf{X}^r] \in \mathbb{R}^{6\times H\times W}$$

The objective is to learn a mapping function

$$\mathcal{F}(\mathbf{X};\Theta) : \mathbb{R}^{6\times H\times W} \rightarrow \mathbb{R}^{2\times H\times W}$$

where, Θ denotes the trainable parameters of DR-RCA Unet++. The network predicts

$$\mathbf{P} = \mathcal{F}(\mathbf{X};\Theta), \qquad \mathbf{P}(i,j) = [p_{bg}(i,j), p_f(i,j)]$$

where p_{bg} and p_f denote the probabilities of the background and flooded classes respectively. The final segmentation mask is obtained as,

$$\hat{\mathbf{Y}} = \arg\max(\mathbf{P}), \qquad \hat{\mathbf{Y}} \in \{0,1\}^{H\times W}$$

A major challenge in multimodal flood mapping is the spatially varying reliability of optical and SAR observations. Optical imagery provides rich spectral information but is susceptible to clouds and illumination variations, whereas SAR imagery is weather independent but affected by speckle noise and backscatter ambiguity in urban settings. To address this challenge, the proposed DR-RCA estimates modality-specific reliability, models optical–SAR interactions through multi-scale cross-attention and dynamically routes modality-specific and cross-modal information for accurate flood segmentation.

III. Proposed DR-RCA Framework

The proposed DR-RCA framework comprises following key components: modality-specific feature projection, modality reliability estimation, multi-scale cross-attention interaction modeling, dynamic reliability routing and feature aggregation and flood segmentation using a ResNet50 backbone-based UNet++ decoder. The overall architecture is illustrated in Fig. 1.

A. Modality-Specific Feature Projection

Given pre-flood Sentinel-2 optical imagery X^s and post-flood Gaofen-3 SAR imagery X^r, modality-specific projection layers transform the inputs into a common feature space:

$$F_s = \phi_s(X^s), \qquad F_r = \phi_r(X^r), \quad F_s, F_r \in \mathbb{R}^{32\times H\times W}$$

Each projection block consists of a 3×3 convolution, batch normalization and ReLU activation:

$$\phi(\cdot) = ReLU(BN(Conv_{3\times 3}(\cdot)))$$

B. Modality Reliability Estimation

The projected optical and SAR features are concatenated:

$$F_{cat} = [F_s, F_r] \in \mathbb{R}^{64\times H\times W}$$

Two independent reliability branches estimate channel-wise reliability maps for optical and SAR modalities:

$$R_{\text{opt}} = \sigma(Conv_{1\times 1}(ReLU(BN(Conv_{3\times 3}(F_{cat}))))),$$

$$R_{\text{sar}} = \sigma(Conv_{1\times 1}(ReLU(BN(Conv_{3\times 3}(F_{cat})))))$$

where, the optical and SAR reliability branches use independent parameters and do not share weights.

$$R_{\text{opt}}, R_{\text{sar}} \in \mathbb{R}^{32\times H\times W}$$

The reliability-aware feature representations are obtained as,

$$F_s^R = F_s \odot R_{\text{opt}}, \qquad F_r^R = F_r \odot R_{\text{sar}}$$

In this work, reliability refers to the estimated confidence of a modality specific feature representation in contributing useful flood-related information at a given spatial location and channel. The learned reliability maps act as adaptive calibration weights that emphasize informative features while suppressing noisy or ambiguous responses before cross-modal interaction and fusion.

C. Multi-Scale Cross-Attention Interaction Modeling

To capture complementary optical–SAR information at different contextual scales, reliability-aware features are down-sampled using average pooling:

$$F_s^{R,i}, F_r^{R,i} \in \mathbb{R}^{32\times \frac{H}{i}\times \frac{W}{i}}$$

where, $i \in \{4, 8\}$ denotes the feature scale.

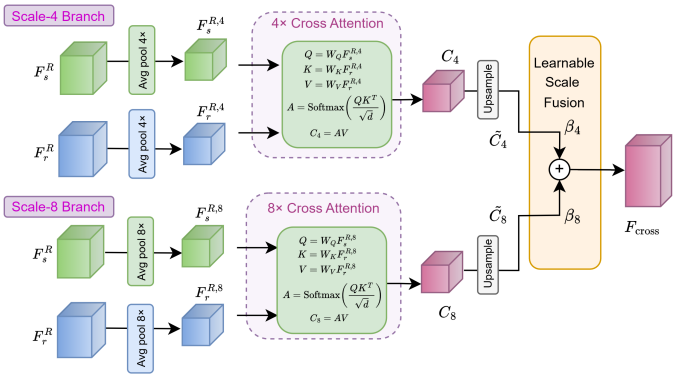


Fig. 2. Multi-scale optical-SAR cross-attention with adaptive scale fusion.

For each scale, query, key and value tensors are generated through learnable 1×1 convolutions:

$$Q = W_Q F_s^{R,i}, \qquad K = W_K F_r^{R,i}, \qquad V = W_V F_r^{R,i}$$

Cross-attention is computed as

$$A = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right), \qquad C = AV, \qquad d = 32,$$

where Q is derived from optical features while K and V are obtained from SAR features. The resulting multi-scale attention features are denoted by C_4 and C_8. After bilinear upsampling to the original resolution, the features are represented as $\tilde{C}_4$ and $\tilde{C}_8$, respectively. Learnable scale fusion is then performed as,

$$[\beta_4, \beta_8] = \text{Softmax}(\mathbf{w}_{scale}),$$

$$F_{\text{cross}} = \beta_4 \tilde{C}_4 + \beta_8 \tilde{C}_8$$

This module enables the network to jointly exploit local contextual interactions from the $4\times$ scale and global semantic relationships from the $8\times$ scale.

D. *Difference-Correlation Feature Modeling*

To explicitly model modality disagreement and agreement, difference and correlation descriptors are computed from the reliability-aware features as,

$$F_{\text{diff}} = |F_s^R - F_r^R|, \qquad F_{\text{corr}} = F_s^R \odot F_r^R$$

Both descriptors belong to, $F_{\text{diff}}, F_{\text{corr}} \in \mathbb{R}^{32 \times H \times W}$

E. *Dynamic Reliability Router*

The Dynamic Reliability Router receives,

$$F_s^R,\ F_r^R,\ F_{\text{cross}},\ F_{\text{diff}},\ F_{\text{corr}}$$

These feature groups are concatenated into,

$$F_{\text{router}} = \text{Concat}(F_s^R, F_r^R, F_{\text{cross}}, F_{\text{diff}}, F_{\text{corr}}) \in \mathbb{R}^{160 \times H \times W}$$

The f_{router} consists of two successive 1×1 convolutions with channel dimensions $(160 \to 64 \to 3)$. Adaptive routing weights are generated through a spatial softmax operation,

$$[W_{\text{opt}}, W_{\text{sar}}, W_{\text{cross}}] = \text{Softmax}(f_{\text{router}}(F_{\text{router}}))$$

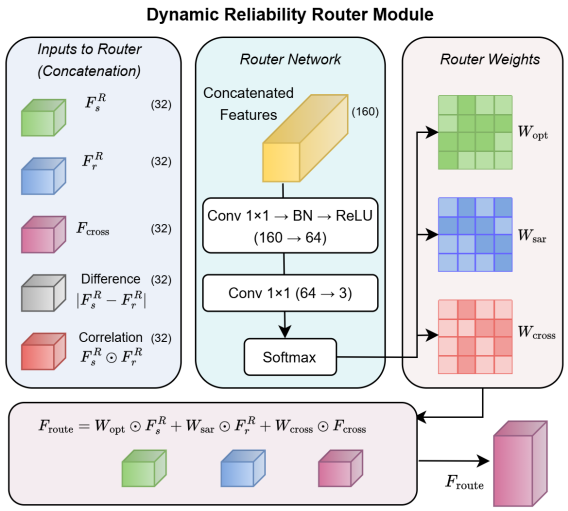


Fig. 3. Dynamic Reliability Router (DRR) for adaptive optical, SAR and cross-modal feature weighting.

where,

$$W_{\text{opt}}, W_{\text{sar}}, W_{\text{cross}} \in \mathbb{R}^{1 \times H \times W}$$

The routing weights satisfy, $W_{\text{opt}} + W_{\text{sar}} + W_{\text{cross}} = 1$

The dynamically routed representation is computed as

$$F_{\text{route}} = W_{\text{opt}} \odot F_s^R + W_{\text{sar}} \odot F_r^R + W_{\text{cross}} \odot F_{\text{cross}}$$

The router therefore adaptively determines whether optical, SAR or cross-modal information should dominate each spatial location, thereby overcoming the limitations of fixed fusion strategies and facilitating more effective exploitation of complementary multimodal information.

F. *Feature Aggregation and Refinement*

The final fusion tensor is constructed as,

$$F_{\text{concat}} = [F_{\text{route}}, F_s^R, F_r^R, \tilde{C}_4, \tilde{C}_8, F_{\text{diff}}, F_{\text{corr}}] \in \mathbb{R}^{224 \times H \times W}$$

The concatenated representation is compressed through two successive 1×1 convolutions $(224 \to 64 \to 32)$

Let the compressed representation be denoted by F_{agg}. A residual refinement step is then applied,

$$F_{\text{res}} = F_{\text{agg}} + F_{\text{route}}$$

Channel-wise feature recalibration is subsequently performed using a squeeze-and-excitation (SE) block. The SE module adaptively reweights feature channels, emphasizing salient flood-related representations and suppressing noisy or redundant activations.

$$F_{\text{fused}} = SE(F_{\text{res}}), \qquad F_{\text{fused}} \in \mathbb{R}^{32 \times H \times W}$$

G. *Flood Segmentation Using UNet++*

The refined fusion representation F_{fused} is forwarded to a ResNet50-based UNet++ segmentation[9] backbone with SCSE decoder attention[10]. The decoder produces

$$L \in \mathbb{R}^{2 \times H \times W}$$

The final flood segmentation map is obtained using

$$P = \text{Softmax}(L), \qquad \hat{Y} = \arg\max(P)$$

IV. Experimental Setup

A. Dataset Description

1) Dataset Overview

The proposed DR-RCA framework is evaluated on the ***Sen2GF3Floods*** dataset [8], a recently introduced multi-source flood mapping benchmark that integrates pre-flood Sentinel-2 optical imagery with during or post-flood Gaofen-3 (GF-3) SAR observations. The dataset covers nine major flood events occurring between 2020 and 2024 across diverse geographic regions and flood scenarios, including riverine and urban flooding. A total of 21,483 image patches of size 256×256 pixels are provided, capturing a wide range of terrain, land-cover conditions and inundation patterns.

2) Input Modalities and Annotation Protocol

Each sample consists of six channels derived from complementary optical and SAR sources. The optical component includes four Sentinel-2 bands namely Red, Green, Blue and Near-Infrared (NIR), while the radar component comprises dual-polarization GF-3 SAR measurements (HH and HV). All modalities are co-registered and provided at a spatial resolution of 10 m. Pixel-wise binary flood masks are generated through a dual-temporal annotation framework that combines semi-automatic labeling, manual refinement and active learning to ensure high annotation quality.

3) Training and Evaluation Split

Following the official benchmark protocol, the dataset is divided into training, validation and test subsets using an 8:1:1 ratio. The diversity of flood types, geographic regions and sensing modalities makes Sen2GF3Floods a challenging benchmark for evaluating the robustness and effectiveness of different fusion and segmentation strategies.

B. Implementation Details

The proposed DR-RCA framework was implemented in PyTorch and trained on a Linux workstation equipped with an NVIDIA RTX 6000 Ada Generation GPU with 48 GB GDDR6 memory. The multimodal DR-RCA module generates a 32-channel reliability-aware representation that is subsequently processed by a ResNet50 based UNet++ decoder with SCSE attention. To reduce effect of class imbalance between flooded and non-flooded pixels, a hybrid loss combining weighted cross-entropy and Dice loss was employed:

$$L = 0.5L_{\mathrm{WCE}} + 0.5L_{\mathrm{Dice}}$$

Model parameters were optimized using the Adam optimizer with an initial learning rate of 10^{-3}. Learning-rate adaptation was performed using a ReduceLROnPlateau scheduler with a reduction factor of 0.5 and a patience of three epochs. Training was conducted for 100 epochs using a batch size of 8. The model achieving the highest validation mIoU was retained for testing.

Performance was evaluated using Overall Accuracy (OA), Precision, Recall, F1-score and mean Intersection over Union (mIoU).

TABLE I
PERFORMANCE COMPARISON OF DIFFERENT SEGMENTATION MODELS.

Model	Acc.	Prec.	Rec.	F1	mIoU
U-Net	0.9874	0.9316	0.8828	0.9065	0.9078
U-Net++	0.9891	**0.9423**	0.8978	0.9195	0.9197
DeepLabV3+	0.9796	0.8613	0.8402	0.8506	0.8592
SegFormer	0.9816	0.8366	0.9111	0.8723	0.8769
Proposed DR-RCA	**0.9892**	0.9172	**0.9274**	**0.9223**	**0.9221**

V. Results and Discussion

The proposed DR-RCA framework was evaluated on the Sen2GF3Floods dataset, which integrates pre-flood Sentinel-2 optical imagery with GF-3 SAR observations acquired during flood events. Table I presents a comparison between the reproduced baseline models (U-Net, U-Net++, SegFormer and DeepLabV3+)[9] [11–13] and the proposed DR-RCA framework. The baseline models were reproduced following their respective implementations and evaluated under the same experimental setting to ensure a fair comparison [8]. As shown, DR-RCA achieved the highest Accuracy (98.92%), Recall (92.74%), F1-score (92.23%) and mIoU (92.21%). Although U-Net++ attained the highest Precision (94.23%), its lower Recall indicates a greater omission of flooded pixels. The superior Recall and mIoU achieved by DR-RCA demonstrate its ability to capture inundated regions more completely while maintaining strong segmentation accuracy.

The results demonstrate the benefit of combining complementary optical and SAR observations for flood mapping. Pre-flood Sentinel-2 imagery provides information on baseline land-cover conditions and permanent water bodies, whereas GF-3 SAR captures flood-induced variations in surface backscatter under all-weather conditions. The integration of these data sources improves the discrimination of flooded and non-flooded areas, particularly in heterogeneous floodplain environments where single-sensor approaches often face ambiguities.

Visual inspection of the predicted flood masks further shows that DR-RCA better preserves narrow inundation corridors and fragmented flood patches compared with the benchmark models. This suggests an improved capability to capture spatially discontinuous flooding, which is critical for operational flood extent monitoring.

Despite these promising results, some limitations remain. Temporal gaps between the pre-flood optical acquisitions and flood events may introduce land-cover changes unrelated to flooding. In addition, SAR-specific effects including shadowing and complex scattering over urban areas can still contribute to classification uncertainty. Future work will investigate more diverse flood events and geographic regions to further assess model robustness and transferability.

Overall, the results highlight the effectiveness of integrating pre-event optical observations with flood-period SAR imagery for accurate flood extent mapping, demonstrating the potential of multi-sensor approaches for operational flood monitoring.

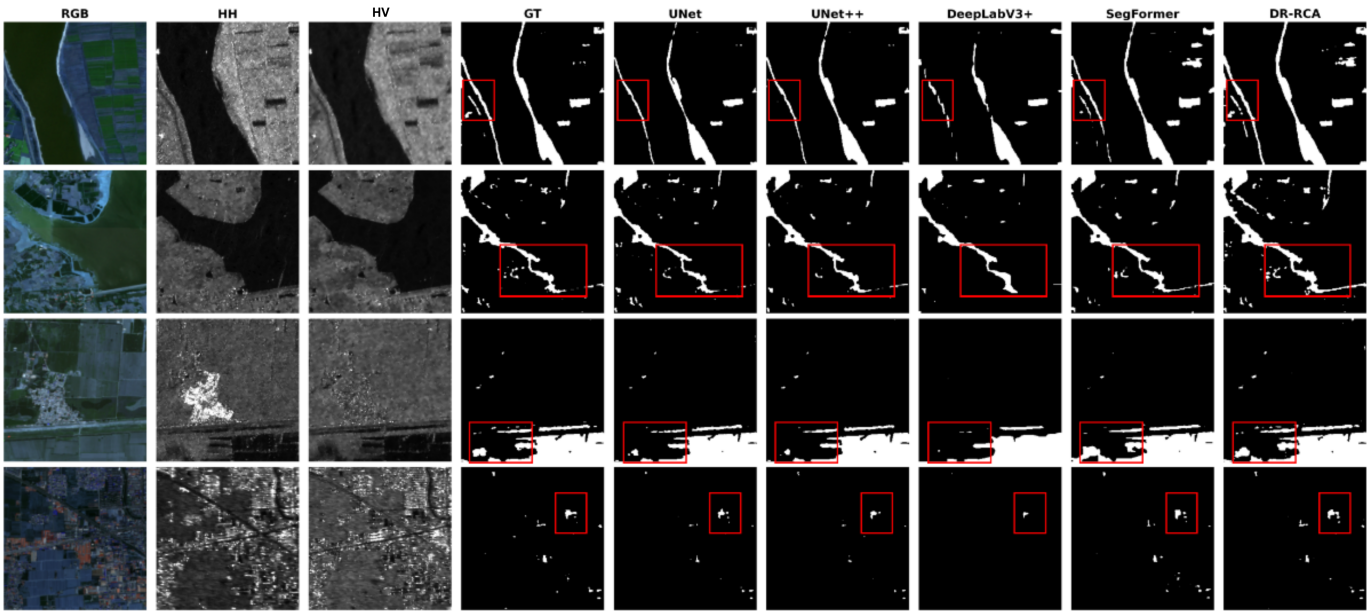


Fig. 4. Qualitative comparison of flood segmentation results using pre-flood Sentinel-2 RGB imagery and post-flood GF-3 SAR (HH,HV) observations. Columns show RGB, HH, HV, ground truth, U-Net, U-Net++, DeepLabV3+, SegFormer and DR-RCA predictions.

VI. Conclusion

This paper presented DR-RCA, a multimodal flood segmentation framework that combines modality-adaptive feature projection, multi-scale cross-attention and dynamic feature routing for integrating pre-flood Sentinel-2 optical imagery with post-flood Gaofen-3 SAR observations. By explicitly exploiting complementary optical–SAR information through agreement-disagreement feature modeling and adaptive routing, the proposed framework improves the effectiveness of multimodal fusion and produces more accurate flood delineation under diverse scene conditions. Experimental results on the Sen2GF3Floods benchmark demonstrate the potential of dynamic routing strategies for robust multimodal flood mapping. Future work will focus on multi-temporal flood monitoring, transformer-based cross-modal interaction mechanisms and improving model generalization across diverse flood events and geographic regions.

References

[1] M. Fawakherji, J. Blay, M. Anokye, L. Hashemi Beni, and J. Dorton, "Deepflood for inundated vegetation high-resolution dataset for accurate flood mapping and segmentation," *Scientific Data*, vol. 12, 02 2025.

[2] C. Sarker, L. Mejias, F. Maire, and A. Woodley, "Flood mapping with convolutional neural networks using spatio-contextual pixel information," *Remote Sensing*, vol. 11, no. 19, 2019. [Online]. Available: https://www.mdpi.com/2072-4292/11/19/2331

[3] G. Konapala, S. V. Kumar, and S. K. Ahmad, "Exploring sentinel-1 and sentinel-2 diversity for flood inundation mapping using deep learning," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 180, pp. 163–173, 2021.

[4] Q. Teng, L. Luo, S. Li, L. Xing, K. Shao, S. Wang, and D. Wang, "A framework for flood inundation extraction based on microwave and optical remote sensing images," *Frontiers in Earth Science*, vol. Volume 12 - 2024, 2024. [Online]. Available: https://www.frontiersin.org/journals/earth-science/articles/10.3389/feart.2024.1511834

[5] R. Yadav, A. Nascetti, and Y. Ban, "Deep attentive fusion network for flood detection on uni-temporal sentinel-1 data," *Frontiers in Remote Sensing*, vol. Volume 3 - 2022, 2022. [Online]. Available: https://www.frontiersin.org/journals/remote-sensing/articles/10.3389/frsen.2022.1060144

[6] A. Zope, U. Chaudhuri, and S. Deshpande, "Determining spatial distribution of disaster intensities from news article using llm," in *IGARSS 2025 - 2025 IEEE International Geoscience and Remote Sensing Symposium*, 2025, pp. 1463–1467.

[7] C. Liu, Y. Sun, Y. Xu, Z. Sun, X. Zhang, L. Lei, and G. Kuang, "A review of optical and sar image deep feature fusion in semantic segmentation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 12 910–12 930, 2024.

[8] W. Chen, Y. Zhu, W. Han, D. Liu, G. Mo, and Z. Xing, "Sen2gf3floods: A benchmark multi-source flood dataset with dual-temporal and active learning annotation," *Scientific Data*, vol. 13, 2026. [Online]. Available: https://api.semanticscholar.org/CorpusID:286125592

[9] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," 2020. [Online]. Available: https://arxiv.org/abs/1912.05074

[10] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel squeeze excitation in fully convolutional networks," 2018. [Online]. Available: https://arxiv.org/abs/1803.02579

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015. [Online]. Available: http://arxiv.org/abs/1505.04597

[12] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," 2018. [Online]. Available: https://arxiv.org/abs/1802.02611

[13] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," 2021. [Online]. Available: https://arxiv.org/abs/2105.15203