https://doi.org/10.5194/egusphere-2025-4613
Preprint. Discussion started: 19 November 2025

SimTA: A dual-polarization SAR time series rice mapping model based on deep feature-level fusion and spatio-temporal attention

Li Liu[a, b, c], Jiaxuan Liang[a, b], Dong Ren[a, b, c*] , Jingfeng Huang[d*]

[a] Hubei Key Laboratory of Intelligent Vision Based Monitoring for Hydroelectric Engineering, China Three Gorges University, Yichang 443002, China

[b] College of Computer and Information Technology, China Three Gorges University, Yichang 443002, China

[c] Hubei Engineering Technology Research Center for Farmland Environment Monitoring, China Three Gorges University, Yichang, 443002, China

[d] Institute of Applied Remote Sensing and Information Technology, Zhejiang University, Hangzhou 310058, China

* Corresponding author.

E-mail address: rendong5227@163.com (D. Ren).

https://doi.org/10.5194/egusphere-2025-4613
Preprint. Discussion started: 19 November 2025

Abstract

Accurate large-scale crop mapping is critical for yield prediction, agricultural disaster monitoring, and global food security. Synthetic Aperture Radar (SAR), with its all-weather, day-and-night imaging capability, plays a vital role in remote sensing based crop mapping studies. However, most existing studies fuse VV and VH polarization channels at the data level, overlooking channels' differences in signal-to-noise characteristics and temporal dynamics, which results in rice feature redundancy or conflicts, particularly at rice field edges and in heterogeneous regions, thereby increasing misclassifications error. To address these challenges, this study proposes a novel Spatiotemporal Attention Model (SimTA) for rice mapping. (1) A VV-VH feature-level fusion scheme is designed, integrated with a Content-Guided Attention (CGA) fusion method which effectively exploits the complementary information of the dual-polarized SAR data for achieving deep spatiotemporal dynamics fusion. (2) A Central Difference Convolution Spatial Extraction Conv (CDCSE Conv) Block is designed, enhancing sensitivity to edge variations of rice field by combining standard and central difference convolutions. (3) To achieve efficient spatiotemporal feature integration across SAR time series, a Temporal-Spatial Attention (TSA) Block is developed, utilizing large-kernel convolutions for spatial feature extraction and a squeeze-and-excitation mechanism for capturing long-range temporal dependencies of rice time series. Extensive experiments were conducted by comparing SimTA with different models under five fusion schemes. Results demonstrate that feature-level fusion consistently outperforms other schemes, with SimTA achieving the best performance: OA = 91.1%, F1 Score = 90.9%, and mIoU = 86.2%. Compared to the baseline SimVP, SimTA improves F1 Score and mIoU by 0.8% and 2.1%, respectively. The CGA enhanced feature-level fusion further boosts SimTA's performance to OA = 91.5% and F1 = 91.4%. SimTA bridges the gap between existing VV-VH deep fusion schemes and modern spatiotemporal modeling demands, offering a more accurate and generalizable approach for large-scale rice mapping.

Keywords: rice mapping, Synthetic Aperture Radar, feature fusion, remote sensing, spatio-temporal attention mechanism

https://doi.org/10.5194/egusphere-2025-4613
Preprint. Discussion started: 19 November 2025

1. Introduction

Rice is one of the world's most essential staple crops. Accurate large-scale crop mapping plays a crucial role in ensuring food security, managing agricultural resources and supporting environmental sustainability (Zeng et al., 2017). Synthetic Aperture Radar (SAR), with its weather-independent and high-frequency revisit capabilities, has become an indispensable data source for large-scale rice mapping, particularly in regions where optical imagery is frequently hindered by cloud cover (Qadir et al., 2024; Silva Filho et al., 2024; Yang et al., 2024). The Copernicus programme of the European Space Agency provides open access to dual-polarization SAR data from the Sentinel-1 mission, which features wide swath coverage, relatively small data volumes, and high temporal resolution. These characteristics make Sentinel-1 a valuable resource for long-term, global-scale agricultural monitoring. Consequently, the use of multi-temporal SAR imagery offers significant potential for accurate, large-area rice mapping (Ge et al., 2025; Wang et al., 2022).

Despite its advantages, a significant portion of existing SAR-based rice mapping studies—summarized in Table 1—still rely heavily on data-level fusion scheme, which concatenate, add, or divide VV and VH dual-polarized SAR bands across time series as model input for classification (Ma et al., 2024; Wei et al., 2021; Yang et al., 2022). While this scheme preserves the full backscatter characteristics of SAR signals, it may lack the capacity to distinguish differences in signal-to-noise ratio and temporal dynamics between VV and VH channels, often leading to higher noise sensitivity, lower polarization utilization and increased misclassification in heterogeneous or fragmented landscapes.

Recent studies have highlighted the potential of feature fusion to address these limitations. In multimodal crop mapping studies, particularly involving optical and SAR images, model backbones are commonly categorized into three main fusion schemes: data, feature, and decision fusion (Liu et al., 2024; Orynbaikyzy et al., 2019; Sainte Fare Garnot et al., 2022). Data fusion directly concatenates inputs from different modalities, which simplifies implementation and minimizes early-stage information loss (Skakun et al., 2017; Valero et al., 2021). Decision fusion combines outputs from modality-specific models, offering flexibility but requiring expert knowledge to accurately interpret and integrate results (Gandhi et al., 2023). Feature fusion scheme not only facilitates more nuanced integration of multi-source information but also enhances model interpretability and robustness (Liu et al., 2025; Sainte Fare Garnot et al., 2022; Zhao et al., 2023). Inspired by the success of feature fusion in multimodal crop classification

https://doi.org/10.5194/egusphere-2025-4613
Preprint. Discussion started: 19 November 2025

studies, feature-level fusion scheme can be used to extract and deeply fuse the spatiotemporal features of VV and VH time-series data.

In terms of modeling, deep learning (DL) methods, particularly Long Short-Term Memory Networks (LSTMs) and Convolutional Neural Networks (CNNs), have become prevalent due to their capabilities in capturing spatial and temporal dependencies in SAR images (Wang et al., 2022). Recent progress has seen increasing attention to spatiotemporal attention mechanisms (Fan et al., 2024; Tang et al., 2024), which aim to jointly model dynamic crop growth processes and static spatial structure. While self-attention-based methods (e.g., Transformers) show promise in extracting long-range dependencies, they often entail substantial computational overhead and exhibit poor generalization when applied to time-series remote sensing data with limited samples (Anandakrishnan et al., 2025; Tarasiou et al., 2023; Yan et al., 2024). Moreover, these approaches often overlook the critical integration between spatial and temporal dimensions, or suffer from overfitting due to overly complex structures.

The U-Net and its variants remain the dominant architecture in SAR-based rice mapping (Ge et al., 2025; Li et al., 2022; Xu et al., 2021) in spatial modeling studies, thanks to their encoder-decoder design, multi-scale representation capabilities, and suitability for temporal integration. However, challenges persist in accurately extracting rice field boundaries and distinguishing rice from spectrally similar vegetation types, especially in irregularly shaped and mixed-crop regions. To enhance boundary sensitivity, some researchers have integrated self-attention mechanisms into U-Net-like frameworks, but such designs often come at the cost of computational efficiency (Bai et al., 2021; Liu et al., 2024; Silva Filho et al., 2024).

In summary, there are two key problems: (1) Most existing studies fuse VV and VH polarization channels at the data level (Table 1), overlooking their differences in signal-to-noise characteristics and temporal dynamics, which often results in feature redundancy or conflicts—especially in edge and heterogeneous regions—thereby increasing classification error. (2) Current deep learning models adopt incomplete spatiotemporal fusion methods by overlooking the critical integration between spatial and temporal dimensions, result in underutilizing time-series SAR data and weakening both rice feature representation and the temporal-spatial correlation between polarization modes.

https://doi.org/10.5194/egusphere-2025-4613
Preprint. Discussion started: 19 November 2025

Table 1

Fusion schemes and methods in related SAR based crop mapping studies, with a predominant focus on data-level fusion scheme.

Model	Objectives	Data	Fusion schemes	Fusion methods	Advantage	Limitations
Unet (Wei et al., 2021)	Rice mapping	Sentinel-1 VV VH	Data Level Fusion	VV © VH	Fusion of VV and VH preserves SAR's full scattering characteristics	Higher noise sensitivity, lower polarization utilization
LSTM (Thorp and Drajat, 2021)	Rice mapping	Sentinel-1 VV VH	Data Level Fusion			
TFBS (Yang et al., 2022)	Rice mapping	Sentinel-1 VV VH	Data level Fusion			
STMA (Han et al., 2023)	Crop mapping (Maize, Wheat, Grassland, Peanut, etc.)	Sentinel-1 VV VH	Data level Fusion	VV/VH	Enhance crop biophysical feature recognition	Loss of absolute physical information
BiLSTM (Ma et al., 2024)	Sediment deposition mapping (Sediment deposition, Water, Farmland)	Sentinel-1 VV VH	Data Level Fusion	VV+VH	Lightweight and efficient computation	Feature masking effect and loss of polarimetric discriminability
XM-UNet (Ge et al., 2025)	Rice mapping	Sentinel-1 VH	VH	-	Decrease data size and processing requirements	Limited information dimensionality

Note: VV © VH means the concatenation of multi-temporal VV and VH.

To address these existing limitations and bridges the gap between existing VV-VH deep fusion schemes and modern spatiotemporal modeling demands, this study proposes a novel Spatiotemporal Attention Model (SimTA) for accurate large area rice mapping by precisely modeling the spatiotemporal dynamics and deeply fusing the spatiotemporal features of VV and VH time-series data from Sentinel-1 SAR imageries at the feature level. The main contributions of this study are as follows:

(1) A VV-VH feature-level fusion scheme is designed, integrated with a Content-Guided Attention (CGA) fusion method which effectively exploits the complementary information of the dual-polarized SAR data for achieving deep spatiotemporal dynamics fusion.

(2) A Central Difference Convolution Spatial Extraction Conv (CDCSE Conv) Block of SimTA is designed for effectively enhancing model's sensitivity to edge variations of rice field by combining

https://doi.org/10.5194/egusphere-2025-4613
Preprint. Discussion started: 19 November 2025

standard and central difference convolutions.

(3) A Temporal-Spatial Attention (TSA) Block of SimTA is developed to utilize large-kernel convolutions for fully extracting spatiotemporal features and a squeeze-and-excitation mechanism for capturing long-range temporal dependencies of rice time series by combining spatial static and temporal dynamic attention mechanisms.

2. Materials

2.1 Study area

To validate the spatiotemporal generalization capability of rice phenological characteristics, this study selects two representative rice-growing regions in North America—the Arkansas River Basin and the Sacramento region—as the study areas. The two regions differ markedly in geographical settings, climatic conditions, and rice cultivation practices, which facilitates a comprehensive assessment of the model's adaptability and robustness under diverse ecological conditions. The Arkansas River Basin is used for training, validation and temporal generalizability test, while the Sacramento region serves as the test site for spatiotemporal generalization.

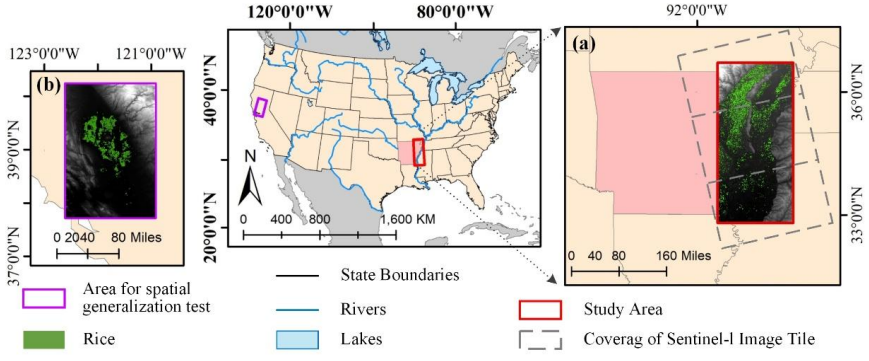


Fig. 1 Geographic location of the study areas. (a) study arca for training, validation and temporal generalizability test, (b) test site for spatiotemporal generalization.

The Arkansas River Basin in the United States (89°50′46″W–91°17′39″W, 33°4′22″N–36°58′10″N) is selected as the study area to better characterize representative features of rice as shown in Fig. 1(a). The region's temperature and precipitation are suitable for rice cultivation, covering approximately 58,504 km². It is a major rice-producing area in the United States, accounting for about 43% of national rice production; therefore, rice samples within the study area are both generalizable and representative

https://doi.org/10.5194/egusphere-2025-4613
Preprint. Discussion started: 19 November 2025

(Moreira et al., 2013; Wei et al., 2025). Rice is typically sown from April to June and harvested from August to October (see Fig. 2). The region also grows cotton, corn, and soybean, whose climatic and phenological characteristics differ markedly from those of rice. Based on the temporal dynamics of backscatter coefficients (see Fig. 2(b) (c)), rice in the Arkansas River Basin is predominantly dry-seeded. Following sowing, fields are generally not maintained under prolonged deep flooding; the exposed, rough soil surface and early weeds induce strong radar scattering, resulting in relatively high backscatter in the VV and VH polarization channels in remote sensing imagery (VV around −20 dB). As irrigation begins and the crop enters the rapid growth stage (May–July), the dense canopy structure gradually becomes the dominant scatterer, leading to a pronounced increase in VV and VH backscatter.

To conduct spatiotemporal generalization analysis, the Sacramento Valley in California (121°10′–122°15′W, 38°02′–39°20′N) (show in Fig. 1(b)) is selected as the validation area. The region has a Mediterranean climate with hot, dry summers and mild, wet winters. Its distinctive irrigation infrastructure provides favorable conditions for rice cultivation, making it another major rice-growing region in the United States. Unlike the Arkansas River Basin, the Sacramento Valley commonly employs water seeding or continuous flooding, and predominantly cultivates short- and medium-grain rice. Sowing typically occurs from April to May, with harvest from September to October. Due to ample water resources and precise field management, VV and VH backscatter coefficients remain at low levels during the early growth stages (VV around −22 dB). Consequently, differences in cultivation practices and water management between the two regions give rise to markedly distinct VV/VH temporal signatures (Yang et al., 2022).

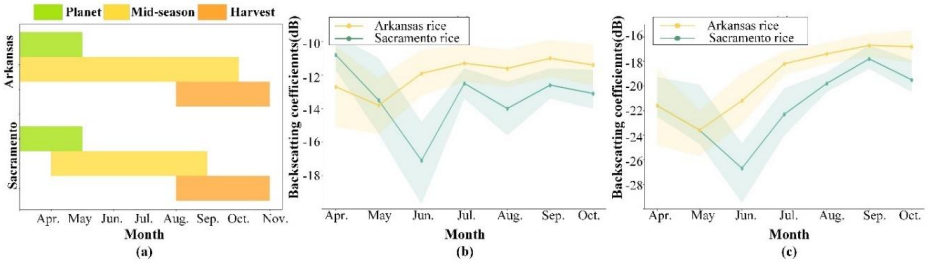


Fig. 2 Calendars and backscattering coefficients curves of rice in study areas. (a) The major calendar of rice. (b) VH Backscatter Coefficients (dB) for Rice in Arkansas and Sacramento. (c)VV Backscatter Coefficients (dB) for Rice in Arkansas and Sacramento.

https://doi.org/10.5194/egusphere-2025-4613
Preprint. Discussion started: 19 November 2025

2.2 Sentinel-1 data sets

We chose Sentinel-1 data to monitor rice growth and its relationship with environmental factors (https://www.earthdata.nasa.gov/data). Sentinel-1 satellite has a revisit period of 12 days and can provide continuous multi-temporal information, which is essential for monitoring the dynamics of rice growth. In addition, its spatial resolution of 10 m is sufficiently detailed to show most of the rice area, ensuring the validity of the data. The interferometric wide-area (IW) mode of Sentinel-1 provides RS imagery with two polarizations (VV and VH). Different polarizations can be chosen to transmit and receive radar signals, allowing for the acquisition of various characteristics of rice features. Each polarization exhibits distinct attributes within the same scene at the same time, providing additional information for crop feature identification and classification. We collected radar images of the rice growth cycle from April to October in 2017, 2018, and 2019 for in-depth analysis.

2.3 Reference datasets

The Crop Data Layer (CDL) dataset is used as the reference data for training, validation and test. The CDL is generated based on Landsat Thematic Mapper imagery and combines Common Land Unit and United States Department of Agriculture (USDA) related ground survey data to form a reliable land cover class dataset which from https://croplandcros.scinet.usda.gov/. The dataset has a spatial resolution of 30 meters and uses the Alber Equal Area Conic projected coordinate system, which provides accurate information on the location of crop fields, the size of their area, and the class to which they belong. The CDL is updated annually through farmer reports, ensuring that the data are current and accurate, making it an important tool for agricultural monitoring and research. By using CDL as reference data, we are able to effectively conduct model training and validation, thus enhancing the reliability and application value of the research results.

2.4 Structure of training, validation and test samples

We cropped the mosaicked, multi-temporal Sentinel-1 imagery covering the study area into a set of small image patches with dimensions of $256 \times 256 \times C \times T$, where T denotes the length of the time series and C represents the number of feature channels derived from the multi-temporal Sentinel-1 data as described in Table 2. The selected multi-temporal Sentinel-1 images encompass the Arkansas during 2017–2019 and the Sacramento area in 2019. Using a non-overlapping sliding-

https://doi.org/10.5194/egusphere-2025-4613
Preprint. Discussion started: 19 November 2025

window approach, we collected the Sentinel-1 image patches and their corresponding CDL samples, yielding a total of 3,026 input image samples of size 256×256×T×C along with their CDL labels. Subsequently, the 2017 and 2018 Arkansas data were split into training and test sets in an 6:4 ratio, while the 2019 Arkansas and Sacramento area images were reserved as test sets to investigate how model generalization across temporal and spatial dimensions affects the accuracy of crop area estimation.

Table 2

Detailed information of the heterogenous datasets

Dataset	Region	Year	Time Series	Channels	Size	Number
Train	Arkansas	2017,2018	13	2	256×256	1061
Validation	Arkansas	2017,2018	13	2	256×256	706
Temporal generalizability	Arkansas	2019	13	2	256×256	884
Spatiotemporal generalizability	Sacramento	2019	13	2	256×256	375

3. Models and experimental setup

The experimental design focuses on three main aspects: VV–VH fusion schemes, the SimTA model, and feature fusion methods (Fig. 3). First, a VV–VH feature-level fusion scheme is proposed (Section 3.1), along with four other fusion schemes for comparison. Next, the SimTA model is introduced, including two novel components—CDCSE Conv Block and TSA Block (Section 3.2)—and six additional DL models are selected for benchmarking. Rice mapping and temporal generalizability experiments are conducted under five fusion schemes and seven models to evaluate the robustness of the proposed VV–VH feature-level fusion scheme (Section 4.1) and assess the performance of SimTA (Section 4.2 and Section 4.3). Subsequently, a CGA deep feature fusion method is developed (Section 3.3) to further enhance the accuracy of SimTA (Section 4.4). In addition, the study employs ablation studies, feature visualization, and Uniform Manifold Approximation and Projection to investigate and validate the proposed model's innovations.

https://doi.org/10.5194/egusphere-2025-4613
Preprint. Discussion started: 19 November 2025

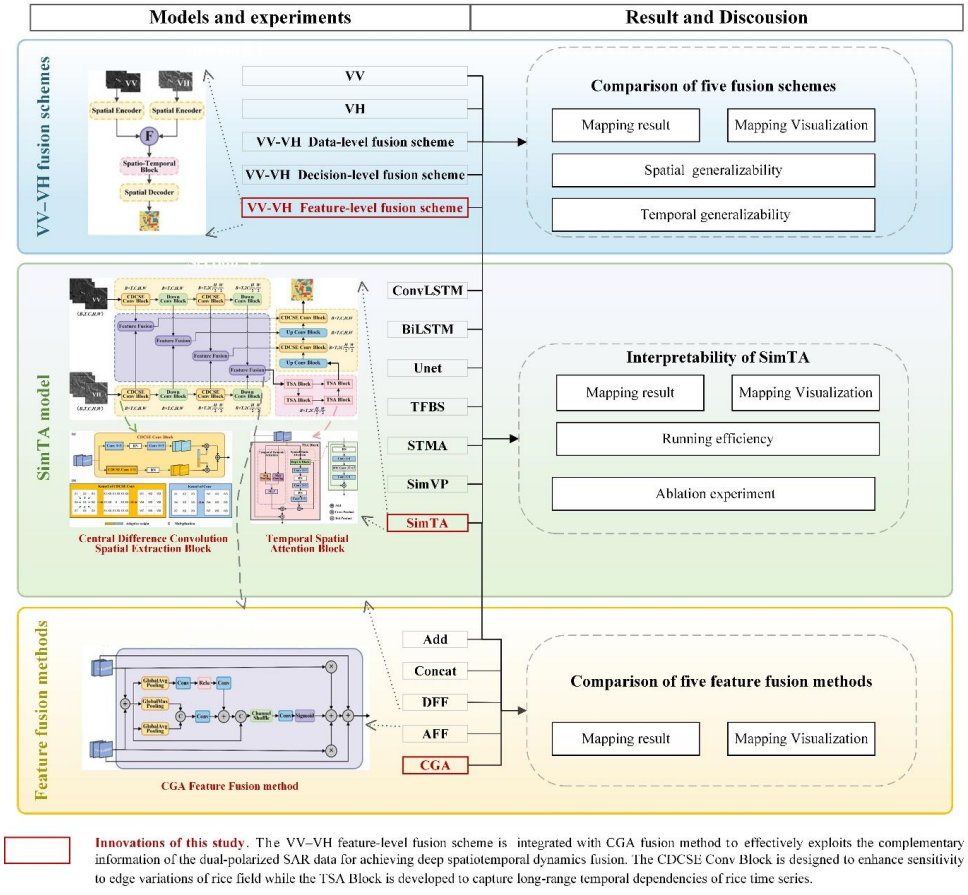


Fig. 3 Experimental design workflow, with the innovations of this study highlighted in box.

3.1 VV-VH feature-level fusion scheme

Given the superior performance of VV and VH polarization modes from Sentinel-1 data in monitoring rice growth conditions, we develop a VV-VH Feature-level Fusion scheme (Fig. 4(d)). In this scheme, representative features are extracted from the raw observations of each polarization, integrated into a unified feature vector, and subsequently processed using pattern recognition techniques to support decision-making. The feature fusion at this level can be achieved through channel-wise summation or concatenation, both of which align with the structure of data-level fusion and fully exploit the complementary information of dual-polarized data to improve feature expressiveness and diversity. More advanced and deeper feature fusion methods will be discussed in Section 3.3.

https://doi.org/10.5194/egusphere-2025-4613
Preprint. Discussion started: 19 November 2025

To verify the VV-VH Feature-level Fusion scheme's performance, we compare it with four other schemes. The VV polarization scheme (Fig. 4(a)) means that both receive and send signals are vertically polarized. Satellite image data mainly reflects the vertical structural characteristics of the target crop surface. In this paper, VV is directly input into the model according to the time series. The VH polarization scheme (Fig. 4(b)) is where the signal is sent vertically polarized and received horizontally polarized. This satellite image allows the model to capture richer features of the target surface and is generally used for scenes with diverse structures. This paper directly feeds the VH into the model in a time series. The VV-VH Data-level Fusion scheme (Fig. 4(c)) performs linear functions or concatenates on channels of raw observations for each polarization, where the most used is stacking on channels, and addition or division is also relatively more used. The VV-VH Decision-level Fusion scheme (Fig. 4(e)) combines model predictions from each polarization via a weighted sum.

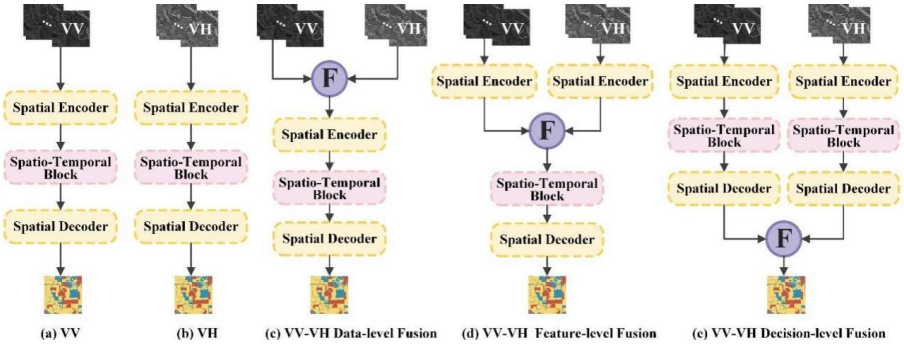


Fig. 4 Illustration of five comparative VV-VH fusion schemes.

https://doi.org/10.5194/egusphere-2025-4613
Preprint. Discussion started: 19 November 2025

3.2 SimTA model architecture

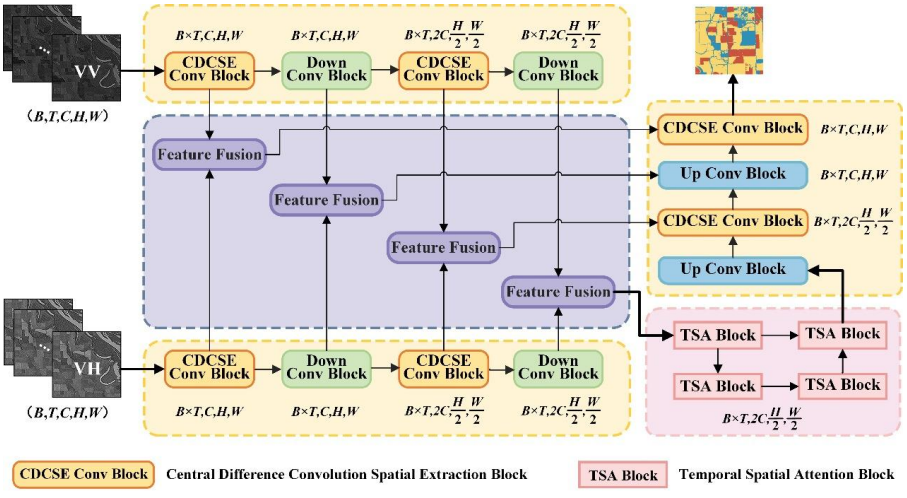


Fig. 5 Overall architecture of the proposed SimTA model for rice mapping using Sentinel-1 VV and VH time-series data. The model first extracts multi-scale spatial features through CDCSE Conv Block, which are then fused with long-range temporal information via the TSA Block. The architecture emphasizes the integration of detailed spatial representations with long-range spatiotemporal dependencies to enhance rice mapping accuracy.

The proposed SimTA (Spatiotemporal Attention) network, illustrated in Fig. 5, adopts a UNet-style encoder–decoder architecture tailored for rice mapping from SAR time series. The network leverages skip connections between corresponding encoder and decoder layers to facilitate multi-level feature fusion, enabling the integration of low-level texture and high-level semantic information. To enhance sensitivity to field boundaries, Central Difference Convolution (CDC) is incorporated into both encoder and decoder modules. By computing the intensity differences between central and neighboring pixels, CDC effectively captures edge details of rice fields. This hierarchical stacking of convolutional layers allows the model to progressively extract more abstract and complex spatial features, transitioning from shallow spatial cues to rich semantic representations.

To model temporal dependencies, a Temporal-Spatial Attention (TSA) module is embedded in the bottleneck of the encoder. This module guides the network in identifying key temporal features across the SAR image sequence, allowing for deeper integration of temporal dynamics with spatial context. SimTA thus combines the spatial feature extraction capability of UNet with the dynamic temporal

https://doi.org/10.5194/egusphere-2025-4613
Preprint. Discussion started: 19 November 2025

modeling of attention mechanisms. By jointly optimizing spatial and temporal features, the model captures complex spatiotemporal patterns of rice growth while maintaining computational efficiency.

3.2.1 Central Difference Convolution Spatial Extraction Block (CDCSE Conv Block)

To enhance the model's ability to capture complex spatial patterns and boundary details in rice fields, we design the CDCSE Conv Block, which combines standard convolution with central difference convolution through an adaptive weighting mechanism (Fig. 6). Time-series SAR images with dimensions (B, T, C, H, W) are input into the encoder using a 3×3 convolution (stride = 2), where T denotes 13 acquisition dates and spatial resolution is 256×256. For efficient spatial feature extraction, the temporal (T) and batch (B) dimensions are merged, enabling the model to learn inter-temporal dependencies while simplifying the input structure. Within each encoding layer, two blocks are employed: a standard ConvBlock for basic operations, and the CDCSE Conv Block for spatial enhancement.

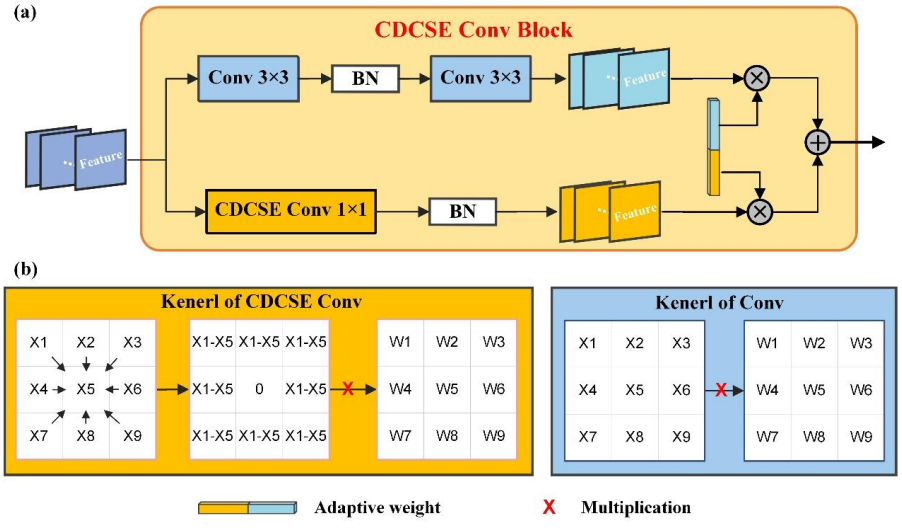


Fig. 6 Architecture of the proposed CDCSE Conv Block with ConvBlock with its convolution kernel schematic; (a) CDCSE Conv Block structure, (b) principle of the convolution kernel.

Standard convolution kernels operate via local weighted summation, which can extract certain textures and features, but limits their ability to capture fine edge features—crucial for high-resolution land cover classification. The specific calculation function is as eq1,

$$y(P_0) = \sum_{P_n} \in R\omega(P_n)(x(P_0 + P_n) - x(P_0)) \quad (1)$$

https://doi.org/10.5194/egusphere-2025-4613
Preprint. Discussion started: 19 November 2025

where the region R denotes the receptive field or the neighborhood defined by the convolution kernel, which contains all offset positions relative to the center point P_0. For every position P_n, compute the difference between the neighborhood point $x(P_0+P_n)$ and the center point $x(P0)x(P0)$, multiply this difference by the corresponding weight $\omega(P_n)$ at that position. Finally, sum all the weighted differences to form the output of this operation. Center difference convolution, as illustrated in Fig. 6(b), enhances gradient sensitivity by computing pixel-wise differences, thus preserving boundary transitions and subtle variations. The adaptive weighting mechanism (He et al., 2024; Li et al., 2020) fuses standard and central difference convolution outputs, dynamically adjusting their contributions based on regional characteristics (Meng et al., 2024). This approach improves spatial detail representation and classification accuracy in heterogeneous agricultural scenes, particularly for rice mapping with intricate edge structures.

3.2.2 Temporal Spatial Attention Block (TSA Block)

To enhance spatiotemporal feature integration in SAR time series, we propose the Temporal-Spatial Attention (TSA) Block, adapted from the spatial attention module of DA-Net (Fu et al., 2019). The TSA Block decomposes attention into two complementary branches:

Spatial Static Attention and Temporal Dynamic Attention, whose outputs are multiplicatively fused (Fig. 7). Spatial Static Attention captures long-range spatiotemporal dependencies using a large receptive field and models global context via the RepLK module (Ding et al., 2022), which applies large-kernel convolutions to capture multi-scale spatial semantics. The subsequent ConvNeXt module (Liu et al., 2022) further refines spatial details and enhances representational capacity for subtle crop growth variations across time.

Temporal Dynamic Attention captures key temporal characteristics by extending channel-wise attention through the SENet (Hu et al., 2020). Here, each time step encodes temporal crop status, and each channel corresponds to different polarimetric or feature dimensions. We follow (Tan et al., 2023) in using global average pooling to compress the feature tensor (T×C,H,W) into (T×C,1,1), then apply fully connected layers to generate reweighting coefficients, restoring it to the original shape. These adaptive weights reflect time-varying channel importance, enabling the model to dynamically focus on informative temporal and spectral responses. Spatiotemporal correlations dominate spatial-only information in remote sensing classification, and the TSA Block effectively exploits this for improved

https://doi.org/10.5194/egusphere-2025-4613
Preprint. Discussion started: 19 November 2025

discrimination across crop types with overlapping phenological stages.

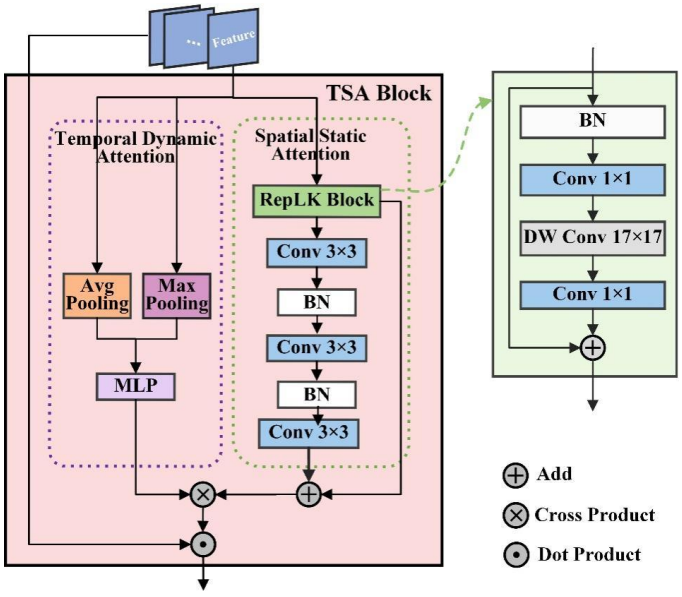


Fig. 7 Specific architecture of the proposed TSA Block.

3.2.3 Loss functions and model settings

This paper uses the cross-entropy loss function, which is well-suited for multi-crop classification tasks. The cross-entropy loss is the most commonly used loss function for pixel-level image semantic segmentation tasks, which seeks the probability of correct classification for each pixel category and pushes the actual labeling probability to one.

$$L = -\frac{1}{N}\sum_{i}\sum_{c=1}^{M} y_{ic}\log(p_{ic}) \tag{2}$$

In eq.2, which is the number of samples, C is the number of classes, y_{ic} is the actual class distribution of sample i, usually denoted as a one-shot coding vector indicating the actual classes, and p_{ic} is the predicted class probability distribution of sample i, usually a probability vector indicating the model's predicted probability for each class.

To train SimTA, we used the AdamW optimizer with a learning rate, which includes a warm-up period from 0 to epoch 5 with a maximum value of 3 to 10, and then the learning rate decays to 5×10^{-4} at the end of training. We trained with a batch size of 2 and in parallel without regularization on an Nvidia

https://doi.org/10.5194/egusphere-2025-4613
Preprint. Discussion started: 19 November 2025

3090 GPU. The entire network was built using the PyTorch deep learning framework with a total training cycle of 40. It takes about 12 minutes to complete a training cycle.

3.2.4 Other models used for comparison

To verify the SimTA's performance, we compare it with six other crop mapping models: ConvLSTM (Masolele et al., 2021), BiLSTM (Ma et al., 2024), Unet (Wei et al., 2021), TFBS (Yang et al., 2022), STMA (Han et al., 2023), and SimVP (Gao et al., 2022). The ConvLSTM and BiLSTM stack spatio-temporal models for temporal image feature classification, using long and short-term memory units. The ConvLSTM uses full connectivity to capture spatiotemporal correlations, but the BiLSTM also improves the ability to capture temporal information through bi-directionality (processing both forward and backward sequences). The Unet, TFBS, STMA and SimVP models use Unet's "encoder-decoder" structure, the structure contains the texture information and semantic information of the image and is used for feature classification. Among them, LSTM is referenced in TFBS, which can be used to establish the dependency of long-range temporal information for each feature. STMA combines spatial self-attention and temporal self-attention to effectively capture the correlations in temporal information and facilitate the fusion of spatiotemporal features. SimVP is used for CNN spatio-temporal modeling to extract multiple crop types from time-series images. This method decouples temporal and spatial information and utilizes a CNN-Inception architecture for crop extraction.

3.3 Content-Guided Attention (GCA) feature fusion method

Compared with traditional shallow fusion methods, we designed a deeper VV–VH Content-Guided Attention (CGA) feature fusion method (Fig. 8(e))(Chen et al., 2024) for improving rice mapping accuracy. CGA employs a content-aware attention mechanism, where an initial spatial attention map is generated for each channel and then refined based on the input feature maps. By leveraging the content of the input features, CGA enhances the network's focus on the unique and complementary characteristics of each polarization channel. This allows for more effective recalibration of the fused features and facilitates the learning of channel-specific attention maps, thereby capturing the distinct distributions and dynamics of VV and VH data more accurately.

For comparison, we also implemented four commonly used feature-level fusion methods to evaluate the performance of our CGA-based fusion method. Add (Addition, Fig. 8(a)) method directly adds two

https://doi.org/10.5194/egusphere-2025-4613
Preprint. Discussion started: 19 November 2025

features in a linear manner. The Concat (Concatenation, Fig. 8(b)) method combines features by channel-wise concatenation. Attention Feature Fusion (AFF, Fig. 8(c))(Dai et al., 2021) uses a dual-branch attention module to extract both global and local features with multi-scale channel attention, while residual connections help preserve both shared and modality-specific information. Dynamic Adaptive Fusion (DFF, Fig. 8(d))(Xue and Marculescu, 2023), employs a global-local adaptive mechanism, where dynamic attention guides the selection of informative features, effectively enhancing feature quality by emphasizing useful details and suppressing redundancies.

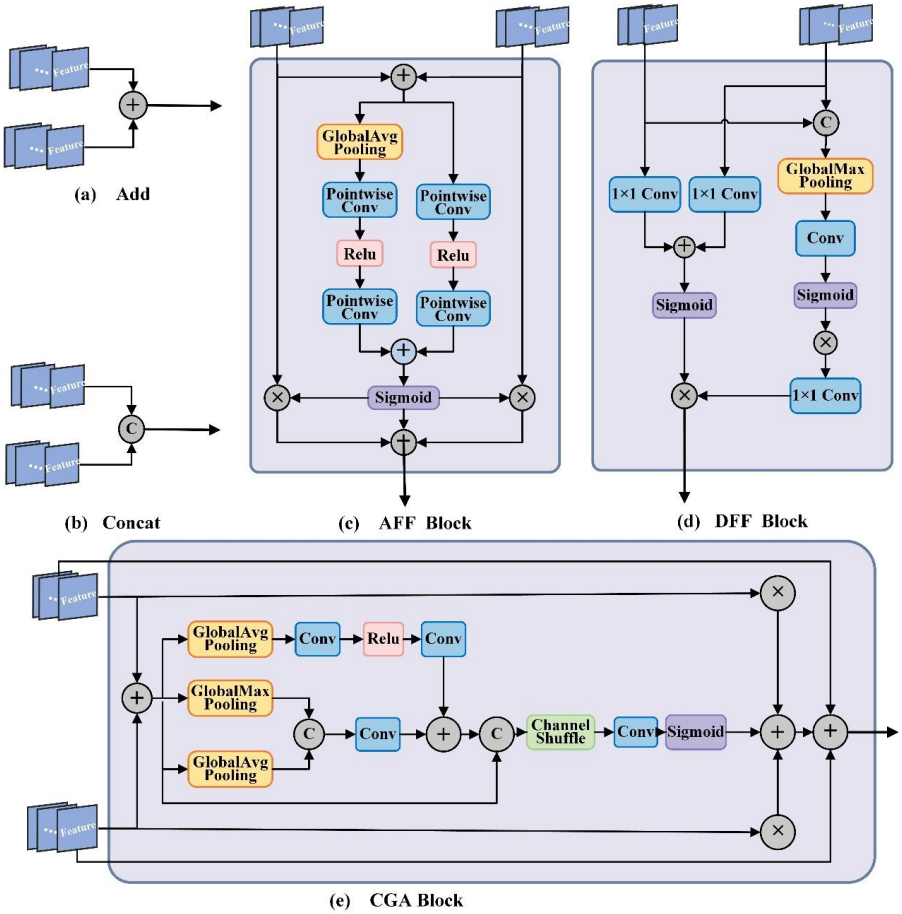


Fig. 8 Illustration of five different methods in Feature-Level Fusion scheme, (a) is Add, (b) is Concat, (c) is Attention Feature Fusion (AFF), (d) is Dynamic Feature Fusion (DFF), (e) is **our** Content-Guided Attention (CGA).

https://doi.org/10.5194/egusphere-2025-4613
Preprint. Discussion started: 19 November 2025

3.4 Evaluation metrics

In this study, F1 scores, mean Intersection over Union (mIoU), and Overall Accuracy (OA) are used as evaluation metrics to assess model performance. These metrics are defined as follows:

$$F1 = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{3}$$

$$\mathrm{m}IoU = \frac{1}{k+1} \sum_{i=0}^{K} \frac{TP}{TP + FP + FN} \tag{4}$$

$$OA = \frac{TP + TN}{TP + FN + FP + TN} \tag{5}$$

TP (True Positive) denotes the number of correctly identified positive samples, specifically the instances where rice is accurately predicted as rice. TN (True Negative) represents the number of correctly identified negative samples, meaning the non-rice instances are correctly classified as non-rice. FN (False Negative) refers to the number of positive samples that are incorrectly classified as negative, i.e., rice samples misclassified as non-rice. FP (False Positive) indicates the number of negative samples mistakenly classified as positive, that is, non-rice samples erroneously predicted as rice.

4. Experiment results and discussion

4.1 Comparison of five fusion schemes

4.1.1 Rice mapping results of five fusion schemes under different models

To compare the rice mapping capability and accuracy of the five fusion schemes, experiments were conducted on the validation set using six different models (ConvLSTM, BiLSTM, Unet, TFBS, SimVP, STMA, and SimTA). The data-level and feature-level fusion schemes adopted the widely used concatenation (Concat) method, while the decision-level fusion schemes employed a weighted summation method. The data-level and feature-level fusion schemes adopted the widely used concatenation method, while the decision-level fusion schemes employed a weighted summation method. The results are shown in Table.B.1.

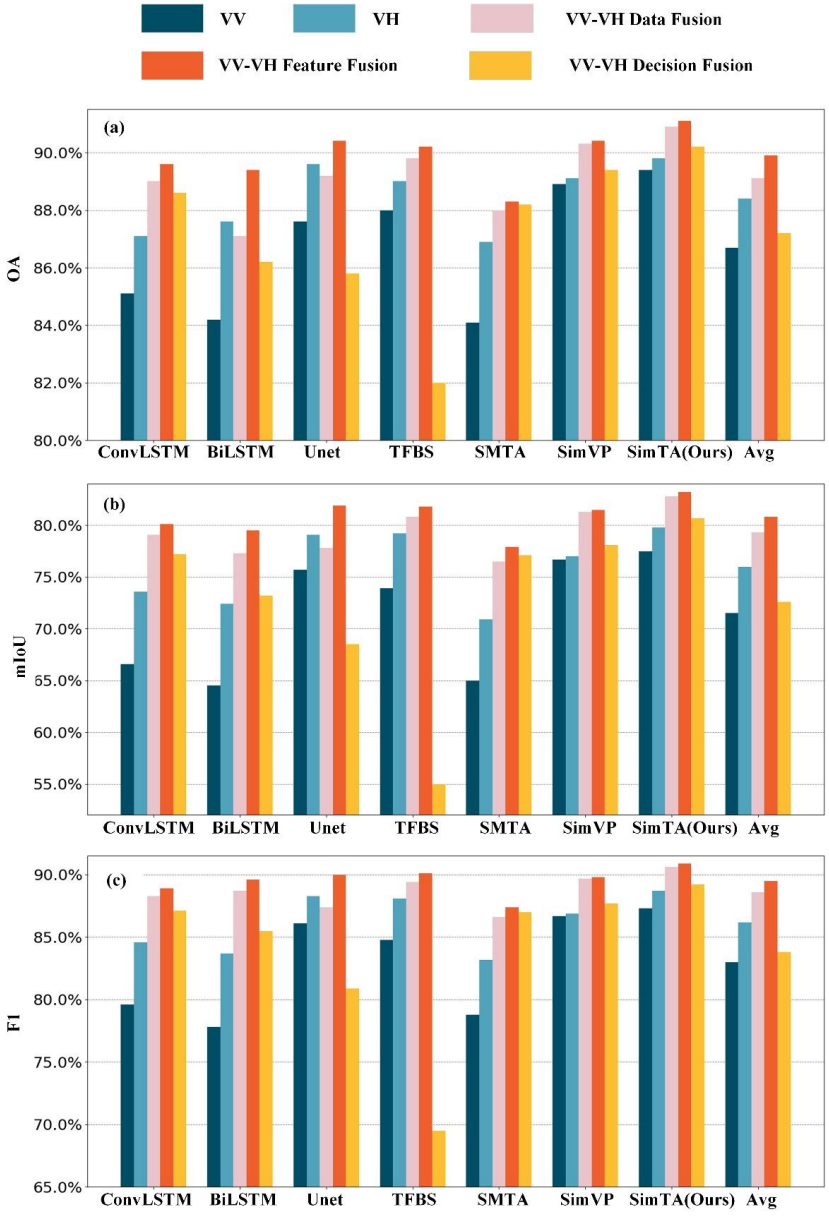
As shown in Fig. 9, there are significant differences in OA, F1, and mIoU for different deep learning models with different fusion, which suggests that the fusion of VV and VH polarization information at the feature extraction stage is more effective than simple data splicing or decision-level fusion in deep learning analysis of remote sensing data. Among them, unpolarized inputs (VV or VH) perform poorly,

https://doi.org/10.5194/egusphere-2025-4613
Preprint. Discussion started: 19 November 2025

especially VV (dark blue), showing the worst results among all schemes (except for the extreme under Unet and TFBS decision-level fusion), the average OA, F1 and mIoU, respectively, while VH (light blue) is slightly better than VV (dark blue). The reason is that the VH polarization contains more information contributing to feature recognition than the VV polarization (Yang et al., 2021; Zhang et al., 2022). However, single-polarized inputs still have a significant performance disadvantage compared to dual-polarization fusion methods. Both data-level fusion (pink) and decision-level fusion (yellow) are effective in improving accuracy (except for Unet and TFBS) but remain overall lower than feature-level fusion (red), which performs the best in all models, especially on TFBS, SimVP, and SimTA (our model). Compared to the suboptimal data-level fusion, the average OA, F1, and mIoU of feature-level fusion across all models increased by 0.7%, 0.9%, and 1.5%, while when the summation method was used, the difference between the two further expanded to improvements of 1.2%, 1.9%, and 3.2% in OA, F1, and mIoU (Fig.A.1; Table.A.1; Table.A.2). This indicates that in deep learning analysis of remote sensing data, the fusion of VV and VH polarization information at the feature extraction stage is more effective than simple data concatenation or decision-level fusion.

Further observing the performance of different models, SimTA performs superiorly under all fusion schemes, especially reaching the highest OA, F1, and mIoU of 91.1%, 90.9%, and 83.2% in the VV-VH feature-level fusion (red) CGA mode, which is a clear advantage over other models. This indicates that SimTA is more robust in spatio-temporal feature extraction and deep fusion of polarized information. In addition, TFBS and Unet also perform relatively well in feature fusion mode, while ConvLSTM and BiLSTM have weaker generalizability, which may be related to their limited time series model capability. It is worth mentioning that the ConvLSTM, BiLSTM, Unet, TFBS, and STMA models involved in the comparison in Section 4.1 all used only the concatenation of data-level fusion in the original published study, and the version of SimTA used for the comparison here used only the same concatenation method of feature-level fusion. While in Section 4.3, more deep feature fusion methods will be further compared, which can further improve rice mapping accuracy; the highest OA, F1, and mIoU are achieved under SimTA with CGA feature fusion with 91.5%, 91.4%, and 84.2%, respectively, which compare to the original publish Unet model (Wei et al., 2021) improved by 1.1%, 1.4%, and 2.3%.

https://doi.org/10.5194/egusphere-2025-4613
Preprint. Discussion started: 19 November 2025



404 **Fig. 9** Rice mapping results of different models under five different fusion schemes with (a) OA, (b)
405 mIoU and (c) F1.

406 **4.1.2 Visualization of five fusion schemes under SimTA**

407 To visually show the differences in rice mapping performance under different fusion schemes, this

https://doi.org/10.5194/egusphere-2025-4613
Preprint. Discussion started: 19 November 2025

study analyzes the crop mapping results by visualization results of the SimTA model and combines them with feature heatmaps which shown in Fig. 10. The heatmaps analysis reveals significant differences in feature focus between the different fusion schemes: the VV-VH feature-level fusion exhibits the most superior performance, while the single-polarized inputs (VV and VH) show obvious limitations.

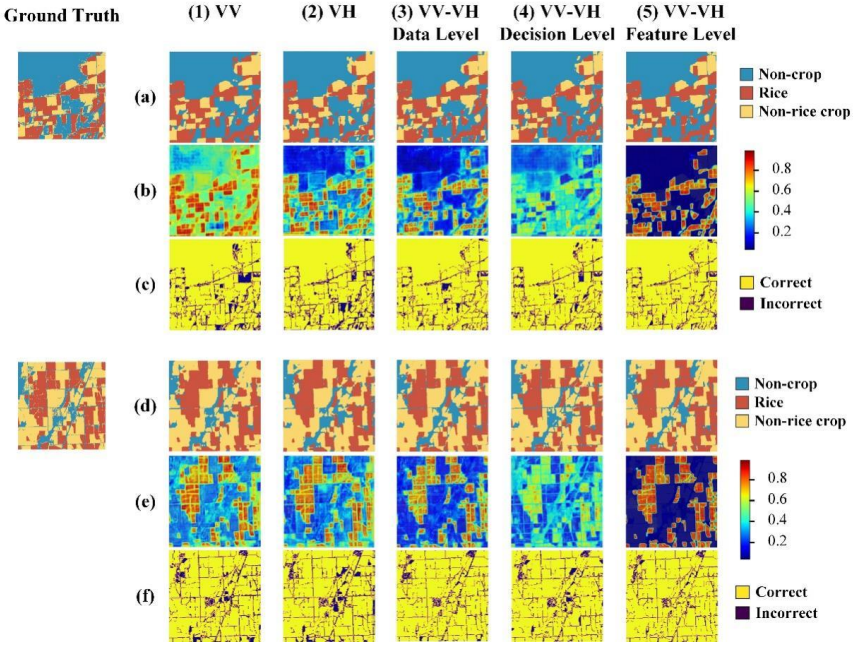


Fig. 10 Visualization of rice mapping and feature comparison results based on the optimal SimTA model under five different fusion schemes, where (a) (d) are the classification result maps, (b) (e) are the heat maps, and (c) (f) are the error maps. The heat value of the heat map is closer to red, the higher the model focuses on this region. The VV–VH feature-level fusion yields clearer heat map boundaries, demonstrating its capability to reduce feature redundancy and conflicts at rice field edges and in heterogeneous areas.

Specifically, the two single-polarization inputs (especially the VV polarization) show a focus on non-target regions in the heat map, a phenomenon that mainly stems from the incomplete information of the single-source data as well as the speckle noise interference inherent in the SAR data (Moreira et al., 2013; Ye et al., 2022; Ye et al., 2024). This lack of information leads to difficulties for the model in fully capturing the features of the target region, which affects classification accuracy. In the VV-VH decision-level fusion scheme (Fig. 10(b4)(e4)), there is a significant deficit in the allocation of the model's

https://doi.org/10.5194/egusphere-2025-4613
Preprint. Discussion started: 19 November 2025

attention to the rice-growing region as shown in the decision-level heatmaps, and this limitation not only affects the accurate identification of the rice area but also reduces its ability to discriminate crops with similar scattering characteristics. It is worth noting that in the decision-level fusion scheme, if the model's attentional weight distribution for rice and other crops tends to be close to each other, it will trigger a more serious model problem, which suggests that the model fails to learn discriminative feature representations among different crop classes adequately, and with the conclusion of (Long et al., 2018) that fusion at the decision level is close.

To further investigate the learning effects of different fusion schemes, we utilize Uniform Manifold Approximation and Projection (UMAP) to visualize the features of different fusions (Mohammadimanesh et al., 2019; Zhao et al., 2025). The results are shown in Fig. 11. UMAP is a commonly used dimensionality reduction technique for visualizing high-dimensional data in two-dimensional space. From Fig. 11, it can be observed that in the feature visualization learned under the SimTA model with different fusion schemes, the data combination of VV-VH significantly enhances the separability between categories. The combination of VV-VH fusion significantly enhances the degree of separation between categories, and classification using multiple fused data features outperforms that using a single polarization input. During VV–VH feature-level fusion, features of the same class tend to form a single cluster instead of interacting extensively with features from other classes. This result confirms that feature-level fusion not only integrates multiple sources of information effectively but also clearly highlights feature differences in complex environments.

In comparison, both feature-level fusion and data-level fusion demonstrate a concentrated focus on rice-growing areas in the heatmaps (Fig. 10 (b3) (e3); (b5) (e5)). Notably, the VV-VH feature-level fusion scheme exhibits a more pronounced effect in classifying the target crops. This is due to the deep neural network structure's ability to learn complex feature representations. Through the fusion of deep network layers, features extracted at different levels can be effectively combined, making features of the same category easier to cluster in high-dimensional space and form distinct category boundaries. The adaptive feature selection and fusion across different network layers allow the model to focus more on selecting features relevant to the target crop while ignoring irrelevant information, effectively filtering out unrelated noise. Therefore, VV-VH feature-level fusion can efficiently integrate VV and VH polarization features, providing a more reliable feature representation foundation for subsequent fine-grained

https://doi.org/10.5194/egusphere-2025-4613
Preprint. Discussion started: 19 November 2025

classification tasks. These findings offer important methodological insights for improving the accuracy of rice mapping and also highlight the critical role of multi-source data fusion in crop classification tasks.

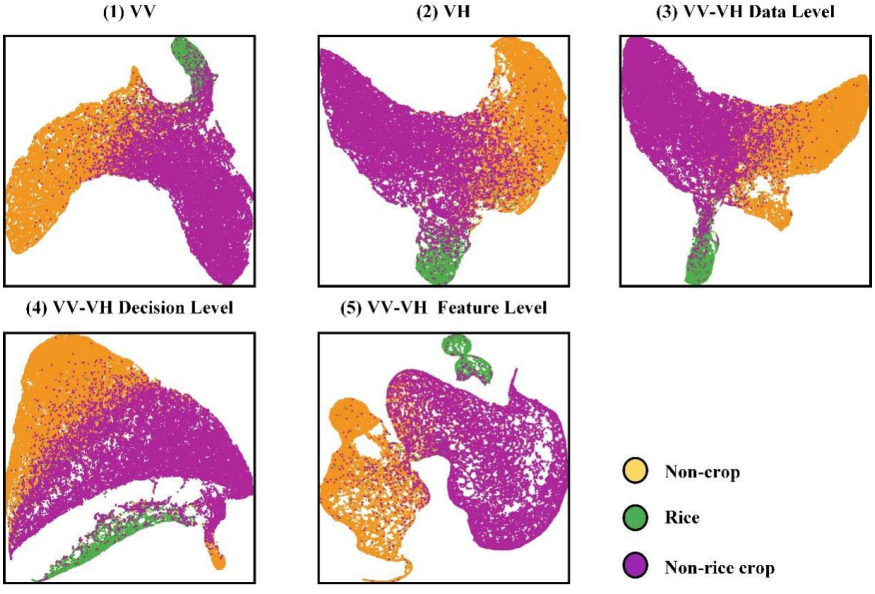


Fig. 11 The UMAP visualization results of five fusion schemes based on SimTA.

4.1.3 Temporal generalizability of five fusion schemes under different models

In the study, the differences in the temporal generalizability of the five fusion schemes were tested and compared using data from 2019 which show in Fig. 12. From it can be seen that all the various fusion schemes exhibit some temporal generalizability. Among them, the feature-level fusion scheme has the best temporal generalizability, followed by data-level fusion, which is significantly better than the other fusion schemes. The average values of F1, mIoU and OA for feature-level fusion under different models are 89.4%, 78.5%, and 87.8%, respectively (Fig. 9-Avg). Compared to VV, VH, decision-level fusion, and data-level fusion, feature-level fusion provides 13.3%, 9.6%, 10.8%, and 0.9% improvement in mIoU.

https://doi.org/10.5194/egusphere-2025-4613
Preprint. Discussion started: 19 November 2025

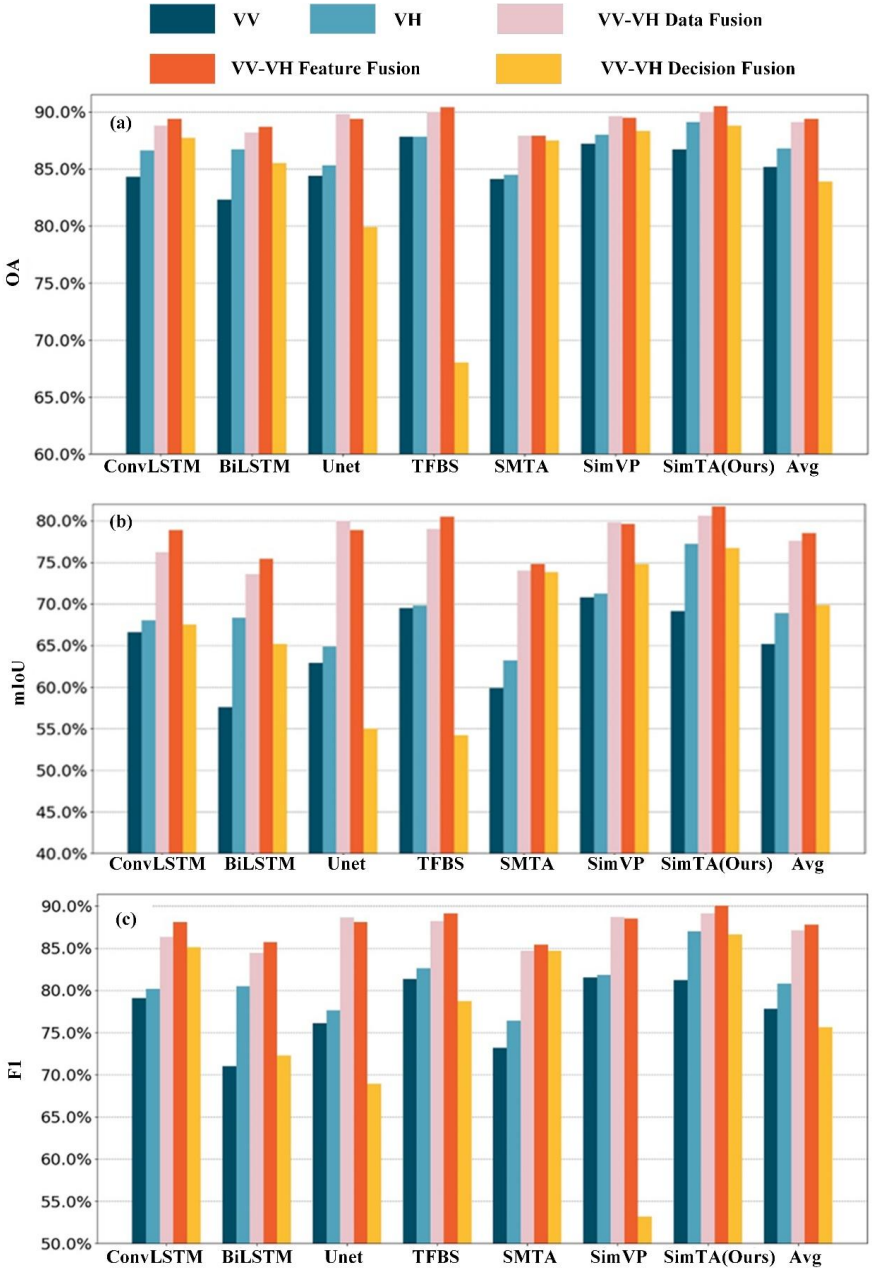


Fig. 12 Temporal generalizability of different models tested in 2019 under five different fusion schemes, (a) OA values; (b) F1 values, (c) mIoU values; Avg: average of all model results.

Notably, the segmentation results in 2019 (shown in Fig. 12) are generally lower than those in 2017 and 2018 (Fig. 9). However, the performance degradation suffered by various fusion schemes is uneven.

https://doi.org/10.5194/egusphere-2025-4613
Preprint. Discussion started: 19 November 2025

For example, the temporal generalizability of VV, VH and decision-level fusion is significantly weaker than that of data-level fusion and feature-level fusion, with their mIoU decreasing by 6.3%,7.1% and 5.0%respectively, compared to the validation set (Fig. 9-Avg) year. In comparison, the mIoU of data-level fusion and feature-level fusion only decrease by 1.6% and 2.3%. This indicates that there are significant differences in the temporal generalizability of different fusion schemes and that VVVH data-level fusion and feature-level fusion are able to utilize the information more comprehensively and retain more details and features, thus effectively capturing the diversity and complexity of the data, and consequently performing well in temporal generalization.

4.2 Interpretability of SimTA

The experimental results in Fig. 9 show that the SimTA model obtained the highest OA,mIoU and F1 under the same "Concat" fusion method. Further, to validate the validity of the model, in this paper, we visualized the comparative results of SimTA's rice mapping, temporal generalizability, and feature extraction (Fig. 13, Fig. 15), carried out the ablation experiments (Table 4), and counted the parameters of the model and the computational efficiency (Table 5).

4.2.1 Rice mapping results and characteristic visualization

Under all five fusion schemes, VV, VH, data-level fusion scheme, feature-level fusion and decision-level fusion, SimTA has the highest classification accuracy, with OA, F1, and mIoU of 91.1%, 90.9%, and 83.2%. Its mIoU is improved by 3.1%, 3.7%, 1.3%, 1.4%, 1.7% compared to BiLSTM, ConvLSTM, Unet, TFBS, STMA with SimVP at feature-level fusion (Fig. 9). This demonstrates that SimTA has a more significant advantage than other models in crop mapping applications. Among these, in feature-level fusion, SimTA's results are close to those of TFBS, which uses skip connections in the semantic layers of the decoder to transmit semantic information while preserving shallow features, promoting more precise classification. However, in other fusion schemes, TFBS's multi-scale fusion scheme does not show clear advantages, as these schemes only combine a single level, focusing more on the model's performance with respect to two factors. In the field of SAR mapping, due to poor imaging quality, it is crucial to extract as much edge and texture information as possible in the shallow layers. SimTA significantly enhances feature extraction capability by adaptively weighting the combination of central difference convolution and standard convolution, suppressing noise influence while simultaneously

https://doi.org/10.5194/egusphere-2025-4613
Preprint. Discussion started: 19 November 2025

modeling both global and local information. This greatly reduces computational redundancy while ensuring high performance, making it especially suitable for task scenarios such as SAR image mapping, which involves complex noise and sparse information.

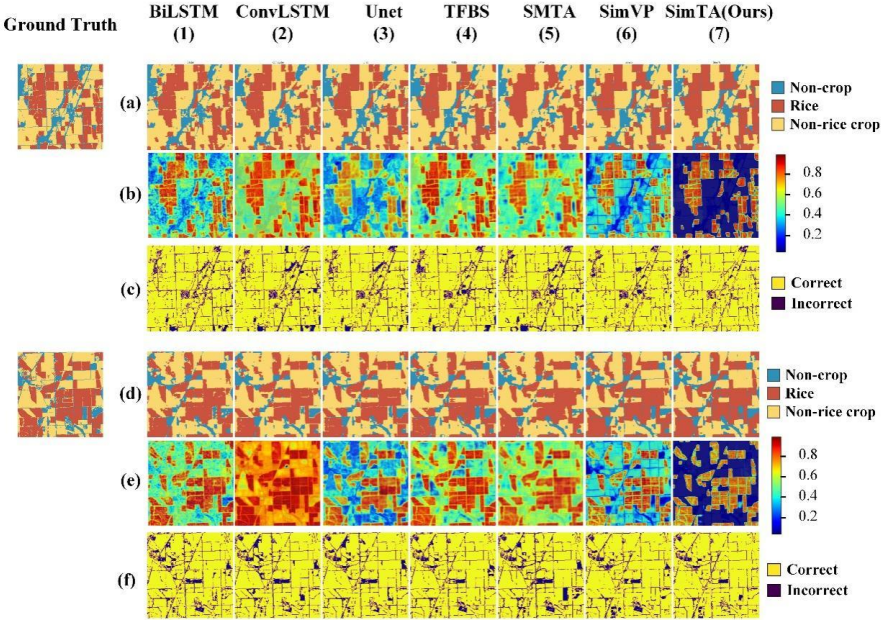


Fig. 13 Visualization of the rice mapping comparison results between SimTA and other five models based on the best VV-VH Feature-Level Fusion. Six different models visualization of the rice mapping comparison results under the optimal VV-VH Feature-Level Fusion, where (a) and (d) are the classification result maps, (b) and (e) are the heat maps, and (c) and (f) are the error maps.

As shown in Fig. 13, the visualization results clearly demonstrate the results of the two regions: (a) and (d) show the result plots of each model, (b) and (e) show the heat map, and (c) and (f) show the error map. The comparison shows that the heat maps of ConvLSTM, BiLSTM, STMA and TFBS show that these models are weak in distinguishing features with similar regions. In contrast, the performance result maps of SimVP, Unet, and SimTA perform better, especially the SimTA model, which is almost completely unconcerned with the rest of the region. SimTA enhances the ability to capture spatial details through the CDCSE Conv Block and strengthens the sequential representation of the spatio-temporal information through the TSA Block. In addition, SimTA combines RepLK with ConvNext, which can integrate the extraction of global and local features of time-series data at the spatial level, thus demonstrating greater adaptability when dealing with time-varying data. This design enables SimTA to

https://doi.org/10.5194/egusphere-2025-4613
Preprint. Discussion started: 19 November 2025

effectively capture dynamic changes when processing time-series image data of rice growth, and to form a stable and reliable feature representation by comprehensively analyzing the image features at different time points. As a result, SimTA is able to keenly perceive the small differences in rice regions when the growing environment changes, showing significant advantages.

The UMAP results of different models in feature fusion are shown in Fig. 14. It shows an overlap between the rice region and other crop regions, while there is no overlap with non-crop regions. This indicates that the segmentation and discrimination of the rice region share similarities in growth characteristics and environmental conditions with those of other crop regions. In contrast, SimTA shows significantly fewer interactions. Other models exhibit substantial interactions between non-crop regions and other crop regions, which collectively reflect their insufficient sensitivity to edge details. The SimTA model, on the other hand, shows significantly fewer interactions. This observation is consistent with the heatmap results in Fig. 13, demonstrating that SimTA is more focused on the characteristics of the target crop area.

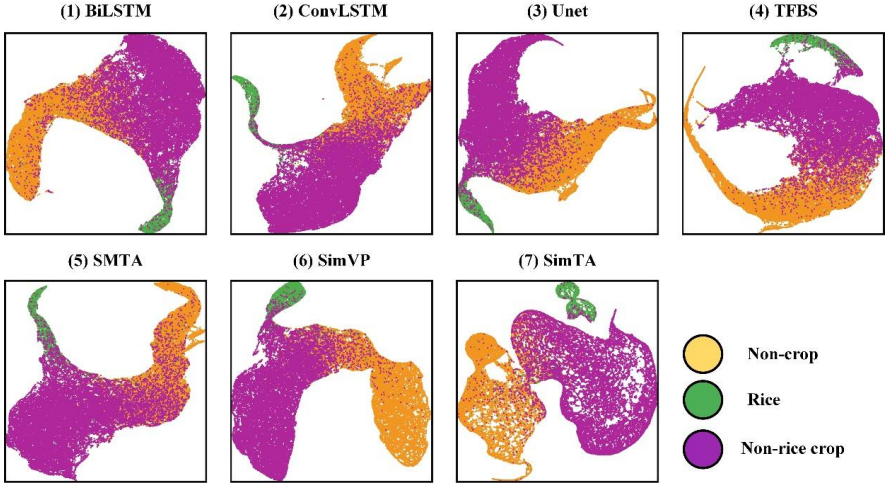


Fig. 14 The UMAP visualization results for different models in feature fusion schemes.

4.2.2 Temporal generalizability and characteristic visualization

The temporal mobility experiment verifies the effectiveness of the SimTA models in terms of temporal generalizability (Table.B.2). Five models show some temporal generalizability, with the overall recognition F1s above 50% and the mIoU metrics all exceeding 50%. Although the time scale

https://doi.org/10.5194/egusphere-2025-4613
Preprint. Discussion started: 19 November 2025

changes and the pixel-level changes caused by agricultural activities still exist, the agricultural enclaves did not substantially change in this time span, so the models perform well in temporal generalizability. SimTA under the fusion of feature-level of VV-VH has the highest classification accuracy, with the F1 score of rice reaching 81.7%. It increased by 2.8%, 6.3%, 2.8%, 1.2%, and 2.1% compared to ConvLSTM, BiLSTM, Unet, TFBS and SimVP, respectively. However, the accuracy of the model decreases over time due to the variability of spatial and radar backscatter in the time domain.

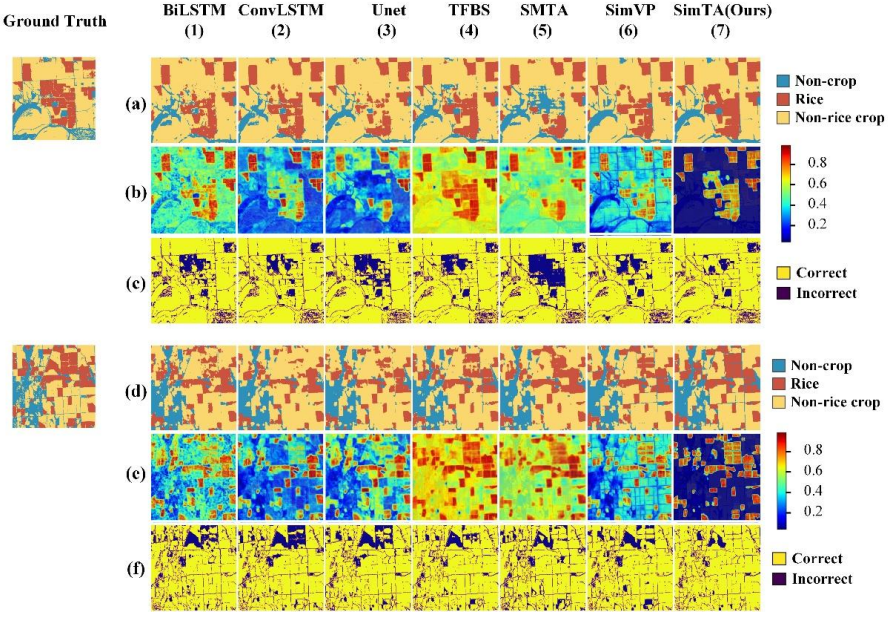


Fig. 15 Visualization of the temporal generalizability comparison results between SimTA and other models based on the feature-level fusion scheme in 2019, where (a) and (d) are the result of crop mapping, (b) and (e) are the heat maps, and (c) and (f) are the error maps.

Fig. 15 shows the visualization of the temporal mobility of the six models in two areas: (a) and (d) show the result plots of each model, (b) and (e) show the heat map, and (c) and (f) show the error map. The comparison shows that ConvLSTM, BiLSTM and Unet are not effective enough to extract features from the rice region compared to the results of the validation set, thus leaving out some rice information. While TFBS is able to extract enough rice information in the test set (shown in Fig. 15(d4)(e4)), it also focuses on irrelevant regions compared to the validation set, which leads to the small objects of crop regions. In contrast, the performance of SimVP and SimTA is much more impressive, especially for the SimTA model, which extracts the features of the rice region completely and pays almost no focus on the

https://doi.org/10.5194/egusphere-2025-4613
Preprint. Discussion started: 19 November 2025

other region. The high visual integrity of SimVP suggests that the model can be effectively applied to other time periods as well, as SimTA emphasizes edge detail and spatio-temporal relationships more than the other models.

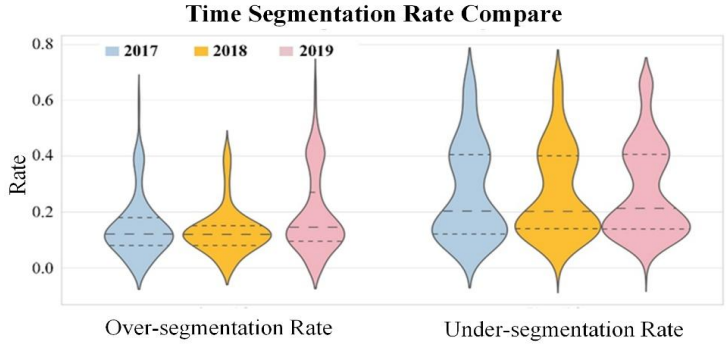


Fig. 16 Violin plot comparing over-segmentation rate with under-segmentation rate based on SimTA in 2017,2018 and 2019.

According to Fig. 16, the analysis of the segmentation error for the rice region between 2017 and 2019 shows that the under-segment rate is always higher than the over-segment rate. This phenomenon mainly stems from the discrete nature of farmland distribution and the localized characteristics of small targets, which makes it difficult for the model to adequately learn its spatial expression law. Specifically, the over-segmentation rate did not change significantly between the three years, indicating that the model has good generalizability for a wide range of rice regions. In contrast, the median misdetection rate in 2019 was significantly higher than that in the previous two years but the upper quartile remained stable. This result reveals a key limitation in the temporal generalizability test: the robustness of the model to extreme misdetection cases did not degrade significantly (upper quartile remained stable), but the shift in the median suggests a distributional bias (such as increased fragmentation of the farmland, adjustments in cropping patterns, or image features) between the 2019 environment and the training data (2017-2018). Inherent challenges such as discrete farmland and small sample targets are further amplified in temporal generalization, resulting in decreased model accuracy.

4.2.3 Spatial generalizability and characteristic visualization

Overall, all fusion schemes demonstrate temporal spatial generalization (see Table 3), with feature-level fusion performing the best, followed by data-level fusion; both significantly outperform the other

https://doi.org/10.5194/egusphere-2025-4613
Preprint. Discussion started: 19 November 2025

schemes. Under feature-level fusion, the average F1 and OA are 91.3% and 92.5%, respectively, corresponding to improvements of 23.2%, 14.6%, 10.8%, 15.1% and 12.1% in F1 relative to VV, VH, decision-level fusion, and data-level fusion, respectively. VV–VH data-level fusion and VV–VH feature-level fusion exhibit stronger generalization advantages when extracting time-series SAR imagery, whereas VV and VH, owing to their single-polarization nature, face difficulty capturing inter-regional differences in rice backscatter features, hindering the accommodation of subtle temporal changes in time-series SAR data across diverse spatiotemporal contexts.

Table 3

Spatial generalization accuracy metrics for different fusion schemes under SimTA

	VV	VH	Data Fusion	Decision Fusion	Feature Fusion
OA	84.2%	86.2%	89.5%	87.5%	92.5%
F1	68.1%	76.7%	79.4%	76.2%	91.3%

Fig. 17 shows a comparative visualization of rice mapping and feature representations across the five fusion schemes for the optimal SimTA model, which (a) and (d) depict the classification outcomes, while (a) and (c) illustrate the error maps and (b) and (d) display the corresponding heat maps. VV and VH polarizations exhibit limited sensitivity to regional variations in rice backscatter, leading to the omission of rice areas with pronounced heterogeneity and consequently impacting classification performance. By contrast, both feature-level fusion and data-level fusion facilitate a more pronounced delineation of rice extent. This enhancement stems from the complementary sensitivities of VV and VH to the underlying scattering mechanisms; when fused, they provide a more comprehensive characterization of regional rice backscatter signatures, thereby improving spatial generalization and yielding more complete and higher-precision classifications.

https://doi.org/10.5194/egusphere-2025-4613
Preprint. Discussion started: 19 November 2025

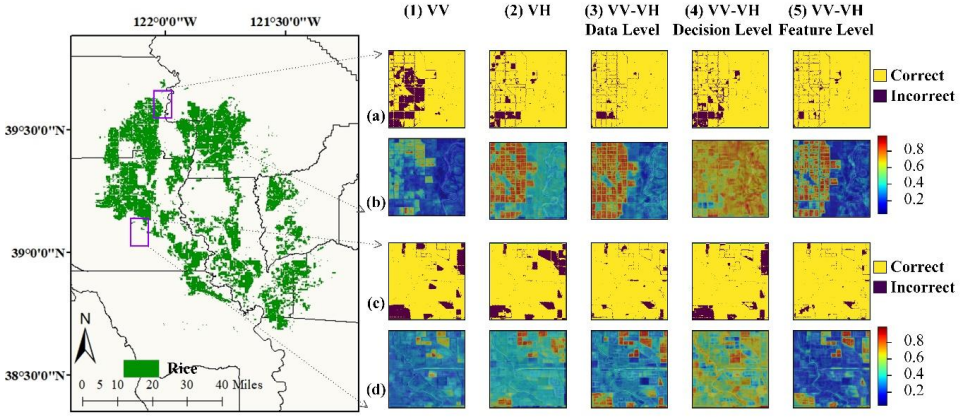


Fig. 17 Visualization of rice mapping and feature comparison results based on the optimal SimTA model under five different fusion schemes, where (a) (c) are the error maps, (b) (d) are the heat maps.

4.2.4 Ablation experiment

In order to investigate the effectiveness of the two modules and assess their impact on the model performance, as shown in Table 4, we conducted ablation experiments on the base model SimTA. We systematically analyzed the performance changes under each configuration by gradually adding CDCSE Conv Block and TSA Block to the base model (Baseline). The experimental results show that the overall accuracy of the Baseline model is 88.6%, the average intersection and merger ratio is 77.3%, the OA reaches 90% after adding feature-level fusion, and the mIoU reaches 0.810. After CDCSE Conv Block is added, the overall accuracy is improved to 90.3%, and the mIoU increases to 0.816, which indicates that CDCSE Conv Block significantly enhances the performance. Further, when TSA Block is added, the model's performance is improved to 90.7% OA and 82.6% mIoU, showing that TSA Block is equally effective. Notably, the model performs best when both CDCSE Conv Block and TSA Block are used, with an overall accuracy of 91.1% and an improved mIoU of 83.3%, showing the synergistic effect of these two modules. In addition, the ablation experiments also show that the introduction of both CDCSE Conv Block and TSA Block can effectively improve the classification effect in different categories of IoU metrics, especially the significant enhancement on Rice IoU and OtherPaddy IoU, which further validates the effectiveness and complementarity of the modules.

https://doi.org/10.5194/egusphere-2025-4613
Preprint. Discussion started: 19 November 2025

Table 4

Indicators of ablation experiments.

	OA	mIoU	F1	RiceIoU
BaseLine	90.3%	81.3%	89.7%	77.0%
BaseLine+ Feature Fusion	90.4%	81.5%	89.8%	80.0%
BaseLine+ CDCSE Conv Block	90.5%	81.9%	90.2%	80.7%
BaseLine+ TSA Block	90.7%	82.6%	90.4%	81.6%
SimTA (Ours)	91.1%	83.2%	90.9%	82.2%

TSA Block**:** It refers to the proposed spatiotemporal attention module detailed in this study. Compared to the Inception module for spatiotemporal sequence extraction used in SimVP, its fundamental advancement lies in its enhanced spatiotemporal modeling capability and its dynamic temporal feature selection mechanism. While Inception solely employs multi-scale convolution to fuse spatial features within individual time steps—thus lacking sensitivity to the temporal dimension—it is unable to effectively capture the temporal patterns associated with crop phenology changes. In contrast, the TSA Block achieves precise modeling of crop growth dynamics through the integration of spatial static attention and temporal dynamic attention: static attention captures long-term spatial dependencies such as field structure, whereas dynamic attention accentuates feature variations across different growth stages. The RepLK module enables the network to attend to relationships among distant pixels within the same temporal period, and when combined with ConvNeXt, it effectively captures local detailed information, allowing for the identification of characteristic spatial distribution patterns during the rice growth cycle. The dynamic attention mechanism further emphasizes the feature disparities among various growth phases.

CDCSEConv Block: Compared to conventional convolutional approaches and existing crop mapping models, this block exhibits notable advantages. Its core strength lies in enhancing the sensitivity to local feature variations and enabling adaptive feature fusion. Traditional convolutions rely on weighted summation for feature extraction, which can lead to blurred boundaries and loss of fine crop details. Center Difference Convolution (CDC) enhances edge sensitivity by computing pixel gradient differences, thereby strengthening the detection of boundary features such as ridges and ditches. Additionally, the

https://doi.org/10.5194/egusphere-2025-4613
Preprint. Discussion started: 19 November 2025

adaptive weighting strategy dynamically balances the contributions of local detail—primarily captured by CDC—and global contextual information—primarily derived from standard convolution. This configuration allows for precise detection of small-scale variations while maintaining boundary continuity in complex agricultural scenes. The design is optimized to address the characteristics prevalent in agricultural remote sensing, such as highly localized abrupt changes, critical boundary information, and complex noise conditions, resulting in significant improvements in small-plot identification and phenological period detection accuracy.

4.2.5 Running efficiency

The parameters and computational efficiencies of the models in this study are shown in Table 5. Our model (SimTA) has 8.48M and 47.08 in the number of parameters and FLOPs, respectively, with a slightly higher number of parameters compared to the baseline model, SimVP of 5.1M, which allows it to have a stronger feature representation. Although its FLOPs are also higher than the baseline model's 28.59, the SimTA model remains within an acceptable range in terms of computational burden when compared to more complex models (TFBS's 50.98 FLOPs), thus proving its superiority in terms of performance and efficiency.

Table 5
Comparison of different model parameters and computational efficiency.

Model	ConvLSTM	BiLSTM	Unet	TFBS	STMA	SimVP	SimTA (Ours)
Parameters(M)	0.06	50.41	32.08	7.7	304.6	5.1	8.89
Flops(G)	20.22	26.41	57.58	50.98	51.6	28.59	47.83

4.3 Comparison of five feature fusion methods

To assess the performance differences of different feature-level fusion methods in rice crop mapping, this study systematically tested multiple fusion methods using a validation set on the SimTA dataset, and the experimental results are shown in Table 6. From the quantitative analysis results, it can be seen that the linear combination (Add) and channel concatenation (Concat) showed poor results. Especially Add showed the worst results with overall accuracy OA, F1, mIoU of 90.9%, 90.8% and 82.8%, which indicates that there are still some mismatched features in multi-scale fusion between Add and Concat. However, DFF does not have a significant increase compared to Concat, which may be due to the loss of some information due to the complex structure. Both IAFF and CGA can effectively improve the result

https://doi.org/10.5194/egusphere-2025-4613
Preprint. Discussion started: 19 November 2025

accuracy, especially the result accuracy of CGA reaches the highest OA, F1, and mIoU of 91.5%, 91.4%, and 84.2%, respectively. This demonstrates that choosing the right features on the channel and using a suitable scheme for fusing the two polarization modalities will be more effective than a simple fusion approach.

Table 6

Comparison of SimTA model rice mapping results under five different feature level fusion methods.

Fusion Method	OA	mIoU	F1	RiceIoU	Parameter(M)
Add	90.9%	82.8%	90.8%	82.0%	8.90
Concat	91.1%	83.2%	90.9%	82.2%	8.89
DFF	91.1%	83.3%	83.3%	82.3%	8.92
AFF	91.4%	83.9%	91.2%	83.3%	8.90
CGA	**91.5%**	**84.2%**	**91.4%**	**83.6%**	**8.91**

Fig. 18 shows the visualization results of SimTA in this paper in the feature-level fusion task. The figure shows that different fusion methods show significant differences in the edge detail and small target detection task. Specifically, the AFF and CGA methods based on the attention mechanism perform particularly well in detecting edge details with small target regions. In contrast, the linear combination (Add) method lost some detailed information during the fusion process, failing to effectively detect edge details with small target rice regions (Fig. 10 (b1)(b2)). The channel concatenation (Concat) method outperforms the linear combination method in detail retention because it retains the complete feature information and reduces the neglect of small objects. However, the Concat method cannot dynamically adjust the importance of the features, resulting in some redundant information being retained, which affects the model's classification performance. In contrast, the AFF and CGA methods based on the attention mechanism not only can dynamically screen important features but also reduce the loss of shallow information through the residual structure, which shows block structure in (Fig. 8(d)(e)), thus achieving a better balance between detail retention and feature expressiveness. In particular, the CGA method, with its integration of global contextual information, not only preserves more detailed information but also significantly enhances the model's focus on the rice region (Fig. 10(g1)(g2)). The CGA method outperforms other fusion methods in small target detection and edge detail processing tasks. In summary, the CGA method demonstrates significant advantages in feature fusion tasks and provides more reliable technical support for rice crop mapping.

https://doi.org/10.5194/egusphere-2025-4613
Preprint. Discussion started: 19 November 2025

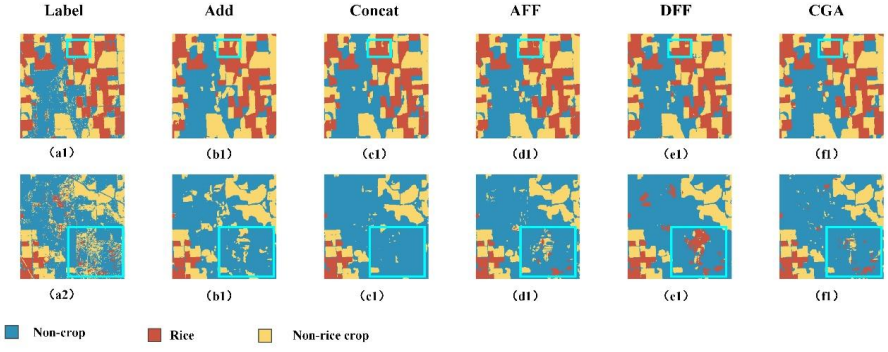


Fig. 18 Visualization of SimTA model rice extraction distribution under five different feature level fusion methods.

5. Conclusion

In this study, we propose a high-precision deep learning network SimTA developed based on Sentinel-1 time series images, and explore the effects of different polarization methods and fusion schemes on the accuracy of rice mapping. The SimTA network 's CDCSE Conv Block combines the advantages of center difference convolution and ordinary convolution, and is able to enhance the weight of the edge information features so as to recognize the size and shape of the rice crop area more accurately. Meanwhile, a TSA Block is introduced into the network design, including spatial static attention and temporal dynamic attention. Spatial static attention is used to capture the spatial global features at each time step, while temporal dynamic attention focuses on the change of temporal information, thus fully utilizing the dynamic characteristics of time series images. In the experiments with three different fusion schemes and single-polarization approaches, the OA, mIoU and F1 of SimTA under the feature-level fusion scheme reached 91.1%, 83.1% and 90.8%, respectively, which significantly improved over other models. In addition, the experimental results show that the feature-level fusion scheme outperforms data-level fusion and decision-level fusion, while the dual-polarization approach outperforms the single-polarization approach. This is due to the fact that feature-level fusion can fully combine spatial and temporal information of the dual-polarization data in the feature extraction stage, which can more accurately capture the boundary and shape features of the rice area. In more complex remote sensing tasks like crop classification or disaster monitoring or when optical data acquisition is difficult, VV-VH deep fusion of dual-polarization features will maximize the segmentation accuracy of the deep learning

https://doi.org/10.5194/egusphere-2025-4613
Preprint. Discussion started: 19 November 2025

model. However, due to the low resolution of SAR images and the presence of noisy patches, in the future we will combine SAR with optical data and thermal infrared data for mapping.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We are very grateful for the support by the Advanced Computing Center of the China Three Gorges University. We express our deepest gratitude to your suggestions.

Funding

This work was supported by the National Natural Science Foundation of China (No. 42401404, 42201412, 42101336). It was also funded by the Natural Science Foundation of Hubei Province (2024AFB130) and the China Three Gorges University (2024SDSJ107), as well as the Fundamental Research Funds for the Central Universities of China under Grant. 2452023048.

References

Anandakrishnan, J., Sundaram, V.M. and Paneer, P., 2025. STA-AgriNet: A Spatio-Temporal Attention Framework for Crop Type Mapping from Fused Multi-Sensor Multi-Temporal SITS. IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens., 18: 1817-1826.

Bai, Y., Mei, J., Yuille, A.L. and Xie, C., 2021. Are transformers more robust than cnns? Advances in Neural Information Processing Systems, 34: 26831-26843.

Chen, Z., He, Z. and Lu, Z., 2024. DEA-Net: Single image dehazing based on detail-enhanced convolution and content-guided attention. IEEE Trans. Image Process., 33: 1002-1015.

Dai, Y., Gieseke, F., Oehmcke, S., Wu, Y. and Barnard, K., 2021. Attentional feature fusion, Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp. 3560-3569.

Ding, X., Zhang, X., Han, J. and Ding, G., 2022. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 11963-11975.

Fan, L., Xia, L., Yang, J., Sun, X., Wu, S., Qiu, B., Chen, J., Wu, W. and Yang, P., 2024. A temporal-spatial deep learning network for winter wheat mapping using time-series Sentinel-2 imagery. ISPRS-J. Photogramm. Remote Sens., 214: 48-64.

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z. and Lu, H., 2019. Dual attention network for scene segmentation, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 3146-3154.

Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E. and Hussain, A., 2023. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. Inf. Fusion, 91: 424-444.

https://doi.org/10.5194/egusphere-2025-4613
Preprint. Discussion started: 19 November 2025

Gao, Z., Tan, C., Wu, L. and Li, S.Z., 2022. Simvp: Simpler yet better video prediction, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 3170-3180.

Ge, J., Zhang, H., Zuo, L., Xu, L., Jiang, J., Song, M., Ding, Y., Xie, Y., Wu, F., Wang, C. and Huang, W., 2025. Large-scale rice mapping under spatiotemporal heterogeneity using multi-temporal SAR images and explainable deep learning. ISPRS-J. Photogramm. Remote Sens., 220: 395-412.

Han, Z., Zhang, C., Gao, L., Zeng, Z., Zhang, B. and Atkinson, P.M., 2023. Spatio-temporal multi-level attention crop mapping method using time-series SAR imagery. ISPRS-J. Photogramm. Remote Sens., 206: 293-310.

He, S., Tian, J., Hao, L., Zhang, S. and Tian, Q., 2024. Unleashing the full potential of hyperspectral imaging: Decoupled image and frequency-domain spatial–spectral framework. Expert Syst. Appl., 243: 122870.

Hu, J., Shen, L., Albanie, S., Sun, G. and Wu, E., 2020. Squeeze-and-Excitation Networks. IEEE Trans. Pattern Anal. Mach. Intell., 42(8): 2011-2023.

Li, Z., Chen, G. and Zhang, T., 2020. A CNN-Transformer Hybrid Approach for Crop Classification Using Multitemporal Multisensor Images. IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens., 13: 847-858.

Li, Z., Chen, S., Meng, X., Zhu, R., Lu, J., Cao, L. and Lu, P., 2022. Full Convolution Neural Network Combined with Contextual Feature Representation for Cropland Extraction from High-Resolution Remote Sensing Images. Remote Sens., 14(9): 2157.

Liu, J., Xu, R., Duan, Y., Guo, T., Shi, G. and Luo, F., 2025. MDGF-CD: Land-cover change detection with multi-level DiffFormer feature grouping fusion for VHR remote sensing images. Inf. Fusion, 120: 103110.

Liu, R., Ling, J. and Zhang, H., 2024. SoftFormer: SAR-optical fusion transformer for urban land use and land cover classification. ISPRS-J. Photogramm. Remote Sens., 218: 277-293.

Liu, Z., Mao, H., Wu, C., Feichtenhofer, C., Darrell, T. and Xie, S., 2022. A convnet for the 2020s, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 11976-11986.

Long, X., Gan, C., Melo, G., Liu, X., Li, Y., Li, F. and Wen, S., 2018. Multimodal Keyless Attention Fusion for Video Classification. Proceedings of the AAAI Conference On Artificial Intelligence, 32(1).

Ma, X., Jiao, J., Yang, B., Zhao, W., Ling, Q., Zhang, X., Han, J., Du, P., Chen, Y. and Chen, H., 2024. Using SAR imagery to extract flash flood sediment deposition area in the northern Loess Plateau. J. Hydrol., 644: 132045.

Masolele, R.N., De Sy, V., Herold, M., Marcos, D., Verbesselt, J., Gieseke, F., Mullissa, A.G. and Martius, C., 2021. Spatial and temporal deep learning methods for deriving land-use following deforestation: A pan-tropical case study using Landsat time series. Remote Sens. Environ., 264: 112600.

Meng, H., Wagner, C. and Triguero, I., 2024. SEGAL time series classification — Stable explanations using a generative model and an adaptive weighting method for LIME. Neural Netw., 176: 106345.

Mohammadimanesh, F., Salehi, B., Mahdianpari, M., Gill, E. and Molinier, M., 2019. A new fully convolutional neural network for semantic segmentation of polarimetric SAR imagery in complex land cover ecosystem. ISPRS-J. Photogramm. Remote Sens., 151: 223-236.

Moreira, A., Prats-Iraola, P., Younis, M., Krieger, G., Hajnsek, I. and Papathanassiou, K.P., 2013. A tutorial on synthetic aperture radar. IEEE Geosci. Remote Sens. Mag., 1(1): 6-43.

Orynbaikyzy, A., Gessner, U. and Conrad, C., 2019. Crop type classification using a combination of optical and radar remote sensing data: a review. Int. J. Remote Sens., 40(17): 6553-6595.

Qadir, A., Skakun, S., Kussul, N., Shelestov, A. and Becker-Reshef, I., 2024. A generalized model for

https://doi.org/10.5194/egusphere-2025-4613
Preprint. Discussion started: 19 November 2025

mapping sunflower areas using Sentinel-1 SAR data. Remote Sens. Environ., 306: 114132.
Sainte Fare Garnot, V., Landrieu, L. and Chehata, N., 2022. Multi-modal temporal attention models for crop mapping from satellite time series. ISPRS-J. Photogramm. Remote Sens., 187: 294-305.
Silva Filho, P., Persello, C., Maretto, R.V. and Machado, R., 2024. Mapping the Brazilian savanna's natural vegetation: A SAR-optical uncertainty-aware deep learning approach. ISPRS-J. Photogramm. Remote Sens., 218: 405-421.
Skakun, S., Franch, B., Vermote, E., Roger, J., Becker-Reshef, I., Justice, C. and Kussul, N., 2017. Early season large-area winter crop mapping using MODIS NDVI data, growing degree days information and a Gaussian mixture model. Remote Sens. Environ., 195: 244-258.
Tan, C., Gao, Z., Wu, L., Xu, Y., Xia, J., Li, S. and Li, S.Z., 2023. Temporal attention unit: Towards efficient spatiotemporal predictive learning, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 18770-18782.
Tang, P., Chanussot, J., Guo, S., Zhang, W., Qie, L., Zhang, P., Fang, H. and Du, P., 2024. Deep learning with multi-scale temporal hybrid structure for robust crop mapping. ISPRS-J. Photogramm. Remote Sens., 209: 117-132.
Tarasiou, M., Chavez, E. and Zafeiriou, S., 2023. Vits for sits: Vision transformers for satellite image time series, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10418-10428.
Thorp, K.R. and Drajat, D., 2021. Deep machine learning with Sentinel satellite data to map paddy rice production stages across West Java, Indonesia. Remote Sens. Environ., 265: 112679.
Valero, S., Arnaud, L., Planells, M. and Ceschia, E., 2021. Synergy of Sentinel-1 and Sentinel-2 Imagery for Early Seasonal Agricultural Crop Mapping. Remote Sens., 13(23): 4891.
Wang, X., Zhang, J., Xun, L., Wang, J., Wu, Z., Henchiri, M., Zhang, S., Zhang, S., Bai, Y., Yang, S., Li, S. and Yu, X., 2022. Evaluating the Effectiveness of Machine Learning and Deep Learning Models Combined Time-Series Satellite Data for Multiple Crop Types Classification over a Large-Scale Region. Remote Sens., 14(10): 2341.
Wei, P., Chai, D., Lin, T., Tang, C., Du, M. and Huang, J., 2021. Large-scale rice mapping under different years based on time-series Sentinel-1 images using deep semantic segmentation model. ISPRS-J. Photogramm. Remote Sens., 174: 198-214.
Wei, P., Guo, J., Lian, J. and Wang, C., 2025. Combination Manner of Sampling Method and Model Structure: The Key Factor for Rice Mapping Based on Sentinel-1 Images Using Data-Driven Machine Learning. IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens., 18: 8340-8359.
Xu, L., Zhang, H., Wang, C., Wei, S., Zhang, B., Wu, F. and Tang, Y., 2021. Paddy Rice Mapping in Thailand Using Time-Series Sentinel-1 Data and Deep Learning Model. Remote Sens., 13(19): 3994.
Xue, Z. and Marculescu, R., 2023. Dynamic multimodal fusion, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2575-2584.
Yan, S., Yao, X., Sun, J., Huang, W., Yang, L., Zhang, C., Gao, B., Yang, J., Yun, W. and Zhu, D., 2024. TSANet: A deep learning framework for the delineation of agricultural fields utilizing satellite image time series. Comput. Electron. Agric., 220: 108902.
Yang, H., Pan, B., Li, N., Wang, W., Zhang, J. and Zhang, X., 2021. A systematic method for spatio-temporal phenology estimation of paddy rice using time series Sentinel-1 images. Remote Sens. Environ., 259: 112394.
Yang, J., Hu, Q., Li, W., Song, Q., Cai, Z., Zhang, X., Wei, H. and Wu, W., 2024. An automated sample generation method by integrating phenology domain optical-SAR features in rice cropping pattern

https://doi.org/10.5194/egusphere-2025-4613
Preprint. Discussion started: 19 November 2025

mapping. Remote Sens. Environ., 314: 114387.
Yang, L., Huang, R., Huang, J., Lin, T., Wang, L., Mijiti, R., Wei, P., Tang, C., Shao, J., Li, Q. and Du, X., 2022. Semantic Segmentation Based on Temporal Features: Learning of Temporal–Spatial Information From Time-Series SAR Images for Paddy Rice Mapping. IEEE Trans. Geosci. Remote Sensing, 60: 1-16.
Ye, Y., Liu, W., Zhou, L., Peng, T. and Xu, Q., 2022. An Unsupervised SAR and Optical Image Fusion Network Based on Structure-Texture Decomposition. IEEE Geosci. Remote Sens. Lett., 19: 1-5.
Ye, Y., Zhang, J., Zhou, L., Li, J., Ren, X. and Fan, J., 2024. Optical and SAR Image Fusion Based on Complementary Feature Decomposition and Visual Saliency Features. IEEE Trans. Geosci. Remote Sensing, 62: 1-15.
Zeng, Y., Xie, Z. and Liu, S., 2017. Seasonal effects of irrigation on land–atmosphere latent heat, sensible heat, and carbon fluxes in semiarid basin. Earth Syst. Dyn., 8(1): 113-127.
Zhang, B., Wdowinski, S., Gann, D., Hong, S. and Sah, J., 2022. Spatiotemporal variations of wetland backscatter: The role of water depth and vegetation characteristics in Sentinel-1 dual-polarization SAR observations. Remote Sens. Environ., 270: 112864.
Zhao, S., Zhang, X., Xiao, P. and He, G., 2023. Exchanging Dual-Encoder–Decoder: A New Strategy for Change Detection With Semantic Guidance and Spatial Localization. IEEE Trans. Geosci. Remote Sensing, 61: 1-16.
Zhao, Y., Zhang, M., Yang, B., Zhang, Z., Kang, J. and Gong, J., 2025. LuoJiaHOG: A hierarchy oriented geo-aware image caption dataset for remote sensing image–text retrieval. ISPRS-J. Photogramm. Remote Sens., 222: 130-151.