

# ASANet: Asymmetric Semantic Aligning Network for RGB and SAR image land cover classification

Pan Zhang <sup>a,\*<sup>1</sup></sup>, Baochai Peng <sup>a,1</sup>, Chaoran Lu <sup>a</sup>, Quanjin Huang <sup>a</sup>, Dongsheng Liu <sup>b</sup>

<sup>a</sup> PIESAT Information Technology Co Ltd., Beijing, 100195, China

<sup>b</sup> Nanjing University of Information Science and Technology School of Remote Sensing and Geomatics Engineering, Nanjing, Jiangsu, 210044, China



## ARTICLE INFO

### Keywords:

Land cover classification  
Multimodal  
Semantic segmentation  
Feature interaction

## ABSTRACT

Synthetic Aperture Radar (SAR) images have proven to be a valuable cue for multimodal Land Cover Classification (LCC) when combined with RGB images. Most existing studies on cross-modal fusion assume that consistent feature information is necessary between the two modalities, and as a result, they construct networks without adequately addressing the unique characteristics of each modality. In this paper, we propose a novel architecture, named the Asymmetric Semantic Aligning Network (ASANet), which introduces asymmetry at the feature level to address the issue that multi-modal architectures frequently fail to fully utilize complementary features. The core of this network is the Semantic Focusing Module (SFM), which explicitly calculates differential weights for each modality to account for the modality-specific features. Furthermore, ASANet incorporates a Cascade Fusion Module (CFM), which delves deeper into channel and spatial representations to efficiently select features from the two modalities for fusion. Through the collaborative effort of these two modules, the proposed ASANet effectively learns feature correlations between the two modalities and eliminates noise caused by feature differences. Comprehensive experiments demonstrate that ASANet achieves excellent performance on three multimodal datasets. Additionally, we have established a new RGB-SAR multimodal dataset, on which our ASANet outperforms other mainstream methods with improvements ranging from 1.21% to 17.69%. The ASANet runs at 48.7 frames per second (FPS) when the input image is 256 × 256 pixels.

## 1. Introduction

LCC in remote sensing images is crucial for monitoring geological disasters (Li et al., 2014), conducting land use analysis (Tu et al., 2020), and facilitating urban planning (Guo et al., 2018). The advent of deep learning has significantly enhanced the intelligent interpretation of remote sensing images (Cao et al., 2023). Numerous existing methods (SLMFNet (Li et al., 2024), FURSformer (Zhang et al., 2023a), MF-DTCFCN (Li et al., 2022c)) primarily rely on single-source satellite data, such as optical imagery, for LCC. However, optical imagery is susceptible to interference from factors like clouds or adverse weather conditions during image acquisition, which can hinder its effectiveness in meeting practical application needs (Araya-López et al., 2018).

Recent advancements in remote sensing technology have enabled the acquisition of data from various imaging modalities across the same geographic area. RGB images provide high spatial resolution and are rich in spectral and textural information; however, they are susceptible to weather conditions (Kazakeviciute-Januskeviciene et al., 2020). In

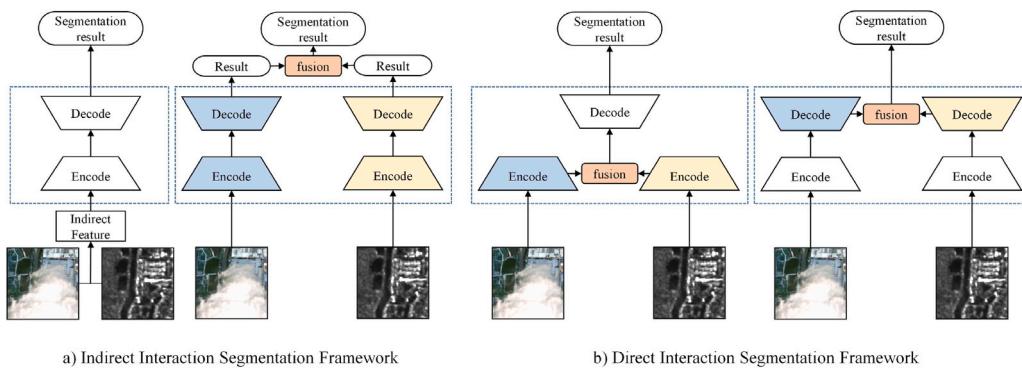
contrast, SAR sensors function in all weather conditions and can penetrate certain types of ground cover (Ma et al., 2017), yielding detailed geometric information about ground objects. In specific applications, RGB and SAR images are highly complementary. The fusion of RGB and SAR images for remote sensing classification is considered a promising strategy to improve classification accuracy (Li et al., 2022a). Nonetheless, segmenting RGB and SAR images for remote sensing purposes presents greater challenges than multimodal segmentation in natural images (Zhang et al., 2020).

Previously, researchers frequently used machine learning techniques (Pei et al., 2014; Stone, 2013) to post-process the segmentation results obtained from both imaging modalities. However, these post-processed results were generally not reusable for further network learning. Currently, deep learning-based multimodal semantic segmentation approaches (CMFNet (Ma et al., 2022b), FTransUNet (Ma et al., 2024), SwinTFNet (Ren et al., 2024)) are increasingly utilized for LCC. As depicted in Fig. 1, these methods can be classified into two

\* Corresponding author.

E-mail addresses: [pzhang@whu.edu.cn](mailto:pzhang@whu.edu.cn) (P. Zhang), [pengbaoc@gmail.com](mailto:pengbaoc@gmail.com) (B. Peng).

<sup>1</sup> Equal contributions.



**Fig. 1.** Comparison of different fusion frameworks: Features are categorized based on whether they engage in direct interaction within the network, being classified into two types: (a) Indirect Interaction Segmentation Framework and (b) Direct Interaction Segmentation Framework.

distinct types based on whether feature interaction occurs within the network architecture.

(a) Feature Indirect Interaction Framework: This framework refers to a segmentation network that lacks direct feature interaction between two modalities during feature extraction or feature aggregation. The Align-CR (Xu et al., 2023a) is employed to generate an indirect image training classification model. Hang et al. (2020) proposed a network that involves fusing the segmentation results from the two branches and then reintroducing them into the network for further learning. While these interaction methods contribute to reducing the discrepancies between modalities, their effectiveness is constrained by the lack of direct connections between modal features. This limitation impedes the network's capability to learn complex feature associations.

(b) Feature Direct Interaction Framework: This framework pertains to a segmentation network that promotes feature interaction between two modalities during feature extraction or feature aggregation. Initial methods, such as those proposed by Ma et al. (2022a), Kang et al. (2022), Hazirbas et al. (2017), generally resorted to simple feature concatenation or summation to integrate feature maps from the two modalities. However, these basic fusion techniques could inadvertently incorporate irrelevant information, potentially disrupting the network's learning process. To address this issue, techniques like SA-Gate (Chen et al., 2020) and MCANet (Li et al., 2022b) have introduced attention mechanisms to compute feature weights between the modalities, thus enabling more effective feature selection and fusion. These symmetric fusion frameworks operate under the assumption that key features are consistently shared across both modalities. Nonetheless, the distinct imaging mechanisms of RGB and SAR suggest inherent differences between the modalities. The traditional symmetric fusion method assigns the same attention weight to different branches, making the features of the two modalities appear closer. However, this approach hinders the learning of complementary features between different modalities.

To fully leverage the complementary features and establish robust feature correlations between RGB and SAR images, we design an Asymmetric Semantic Aligning Network (ASANet). This network is composed of two key modules: the Semantic Focusing Module (SFM) and the Cascade Fusion Module (CFM). The SFM captures the unique characteristics of RGB and SAR images by assigning different attention weights to features during their interaction process. The CFM enables the network to grasp the complex relationships between the two modalities by learning from both channel and spatial perspectives, thereby facilitating the fusion of features from the two modalities. By integrating the SFM and CFM, our network can effectively extract and correlate features from both modalities, significantly diminishing the impact of noise interference.

Furthermore, we observe a scarcity of publicly accessible RGB-SAR datasets, many of which exhibit uneven quality. To address this lack, we initiated a data collection campaign that gathered RGB and SAR imagery from the Pearl River Delta region in China. The collected

images were carefully positioned, corrected, and refined annotations were meticulously applied. Specifically, our dataset comprises regions with cloud cover, offering a challenge for evaluating the effectiveness of RGB-SAR fusion under less-than-ideal observation conditions. Unlike datasets that artificially simulate cloud and fog conditions, ours is more appropriate for intelligent interpretation tasks in real-world application scenarios.

The main contributions of this paper are as follows:

(1) We introduce ASANet, which is distinguished to effectively calibrate and align the information of the two modalities by its ability to leverage the complementarity and consistency of feature information across different modalities.

(2) We have constructed a dataset with fine-grained labels, named PIE-RGB-SAR, which includes RGB and corresponding SAR images. The land categories are organized into six distinct classes.

(3) Our proposed ASANet achieves state-of-the-art (SOTA) performance on the three datasets. Specifically, on the PIE-RGB-SAR dataset, our ASANet outperforms other methods, with an improvement in mean Intersection over Union (mIoU) ranging from 1.21% to 17.69%.

## 2. Related work

### 2.1. Land cover classification

The significance of LCC in remote sensing images cannot be overstated, given its pivotal role in addressing a wide range of practical application requirements (Li et al., 2022a). Within this context, semantic segmentation is extensively employed to guarantee accurate classification of various land cover categories. Initially, researchers relied on machine learning techniques, such as random forest (Rodríguez-Galiano et al., 2012) and K-means clustering (He et al., 2014), to segment high-resolution remote sensing images, which aided in the efficient extraction of surface features.

With the rapid advancement of deep learning techniques, models based on Convolutional Neural Networks (Yang et al., 2018) and the Transformer (Zhou et al., 2023) series have become widely utilized in remote sensing image analysis. For instance, SLCNet (Yu and Ji, 2023), which integrates global and local feature correlations, demonstrates superior performance across three remote sensing datasets. Nevertheless, fully supervised semantic segmentation methods (Costa et al., 2018; Zhao et al., 2023) are increasingly challenged in meeting the demands of large-scale land mapping tasks. Ma et al. (2023) introduced an alternative approach in which an unsupervised domain adaptation method for LCC is developed, leveraging local consistency and global diversity. This method utilizes a subset of annotated source domain data along with unlabeled target domain data to improve the generalization capability of the semantic segmentation model. This adaptation strategy enhances the model's adaptability for LCC in diverse geographical areas. Additionally, a prominent remote sensing big-model (Oscio et al.,

2023) aims to promote the extensive application of semantic segmentation models in real-world scenarios. While these strategies contribute to making the segmentation process more intelligent, challenges persist when dealing with RGB images captured under challenging conditions and extreme weather. Such conditions can pose difficulties for models to learn effectively and segment feature classes accurately.

SAR has garnered considerable attention from researchers due to its microwave-active imaging mechanism, which enables all-weather imaging capabilities. Early studies (Orlíková and Horák, 2019) exploited the sensitivity of different polarization SAR images to feature types for LCC using machine learning. Additionally, the segmentation of water and soil areas (Guo et al., 2022) could be relatively easily achieved by utilizing the water path scattering property of SAR. However, the coherent imaging mechanism of SAR introduces a high degree of variability at the individual pixel level, known as coherent scattering (Lu, 2021), which poses significant challenges for LCC. The GLNS (Liu et al., 2022) addresses this challenge by integrating both local and global features to extract more efficient and discriminative features and addresses the issue of excessive inter-class distances in SAR by incorporating a twofold loss function. Liu et al. (2019) proposed combining representation learning with statistical analysis to capture the statistical properties of SAR in feature space. While these methods offer various perspectives for mitigating the interference caused by coherent scattering in SAR, the LCC task based on SAR remains challenging. This challenge is primarily due to the lack of rich texture and the absence of color information in SAR images.

## 2.2. Multimodal segmentation algorithm

In challenging real-world application scenarios, single modal algorithms often struggle to fully meet the task requirements. For example, in tasks like semantic segmentation for autonomous driving in foggy conditions (Rizzoli et al., 2022), intelligent interpretation of remote sensing images with cloud cover (Li et al., 2023a), or medical image diagnosis relying on multimodal data (Hermessi et al., 2021), the complexities of these real-world scenarios have led to the development of multimodal segmentation algorithms.

As shown in Fig. 1(a), the model trains on the fused images generated from the data of the two modalities to obtain segmentation results, or it fuses the segmentation results of the two modalities and calculates the loss. Xu et al. (2023b) proposed generating the RGB cloud-removed image from the RGB-SAR dataset, which is subsequently used for LCC. Some scholars have also employed methods such as Principal Component Analysis, HSV Color Space, and Generative Adversarial Networks (Bermudez et al., 2019) to create fused images, which are then input into the network for training. The quality of the generated images significantly affects the segmentation results (He and Yokoya, 2018). Hang et al. (2020) introduced a network where the post-processing results are obtained through the weighted summation of the segmentation results from the two modalities. The loss is calculated by comparing these results with the ground truth, providing feedback for network learning. However, this method encounters challenges in learning the complementary relationship between the two modalities. To address these issues, MMFNet (Li et al., 2023b) computes phase features of RGB and SAR images as intermediate features for multi-level training, thereby avoiding direct feature interaction between the modalities. Nonetheless, the multi-level network model presents a complex structure requiring longer training time.

Typically, symmetric Siamese networks are employed to construct the multimodal framework, which facilitates the adaptive learning of feature correlations between the two modalities. As depicted in Fig. 1(b), feature fusion occurs within the encoder or decoder section of the framework, and one or more segmentation heads are deployed to generate the final segmentation outcome. Methods such as FuseNet (Hazirbas et al., 2017) and the network proposed by Gao et al. (2023) fuse feature maps from both modalities at the decoder stage

and then feed them into the segmentation head for the segmentation process. CMFNet (Ma et al., 2022b) and FTransUNet (Ma et al., 2024) propose the use of cross-scale fusion modules, which are designed to integrate features across different modalities and at multiple levels. This fusion strategy integrates the feature extraction component, facilitating the learning of feature correlations. However, it does not provide a feedback mechanism to the branches for feature interaction. The attention weights for feature interactions within SwinTFNet are adjusted exclusively for the RGB branches. CMX (Zhang et al., 2023b) introduces an attention module named CM-FRM, which is specifically designed to learn the weights of two modal channel and spatial features through feature interaction. These interacted feature maps are then passed to the subsequent layer for further learning. The CM-FRM module enhances the network's capability to integrate feature information from another modality within the encoder. Other networks, including ACNet (Hu et al., 2019), SA-Gate (Chen et al., 2020), and CroFuseNet (Wu et al., 2023), also utilize distinct attention modules in their encoder sections to calculate feature weights between the two modalities. These weights are employed to selectively fuse features across the two modalities. While these interactive modules calculate feature weights that highlight the correlation between features of the two modalities and underscore the significance of feature fusion, they often fail to consider the feature independence within each modality. These interaction modules highlight the correlation between features of two modalities and emphasize the importance of feature fusion. However, they often ignore the feature independence within the RGB and SAR modalities, and only output a feature weight that leads to the same feature representation for both modalities to learn.

## 3. Method

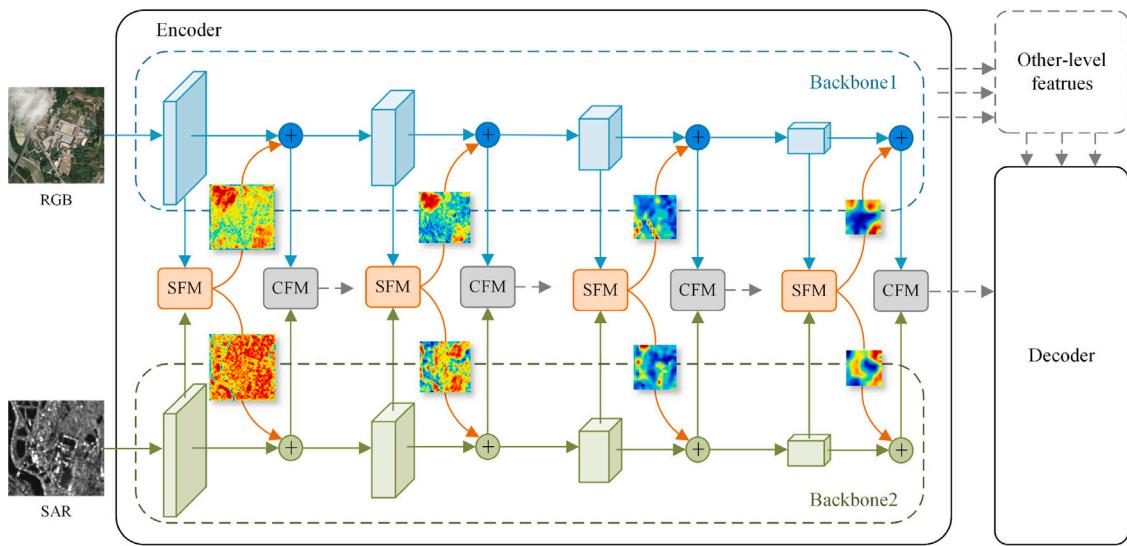
In this section, we present a comprehensive overview of our proposed Asymmetric Semantic Aligning Network, which comprises the Semantic Focusing Module and the Cascade Fusion Module.

### 3.1. Overall network architecture

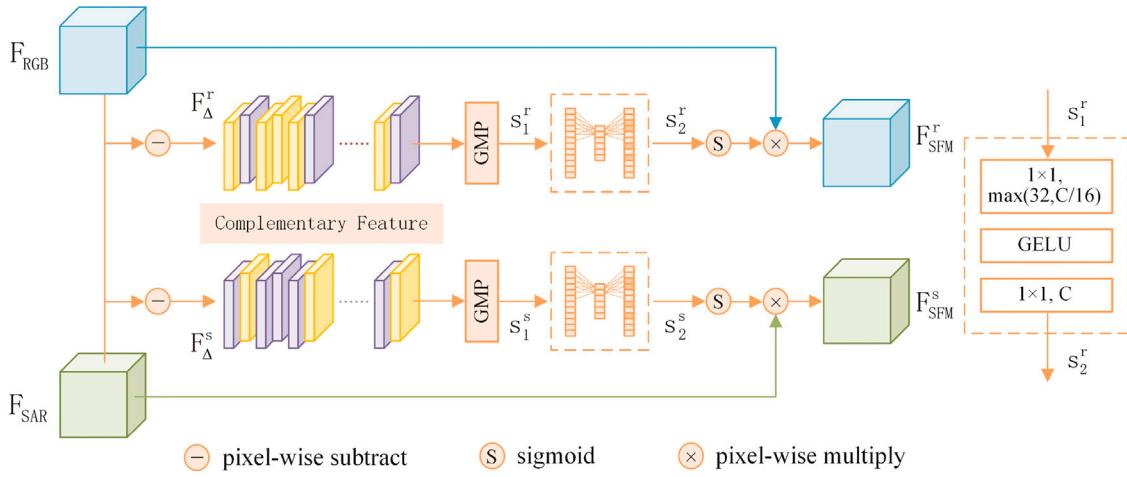
The overall network structure of ASANet is provided in Fig. 2, which is structured on a dual-branch encoder and decoder architecture. In the encoder, we employ two branches where weights are not shared to concurrently extract features from RGB and SAR images, thereby capturing the unique characteristics of each modality in a parallel and interactive manner. The model's residual structure enables it to discern and leverage the substantial differences among various data modalities, resulting in the generation of more effective and complementary features for subsequent stages of feature extraction and interaction. To optimally fuse the two modal features after focusing on the SFM, we propose the CFM. The decoder can be based on nearly any design from SOTA segmentation networks. For broad applicability, we select UPerNet (Xiao et al., 2018) as our preferred decoder design.

### 3.2. Semantic focusing module

Based on the analysis above, it is clear that the feature information required for RGB and SAR images is usually different. Consequently, this study introduces an innovative SFM that obtains different information to adjust the features of parallel branches. Drawing from the channel attention mechanism (Hu et al., 2018), the SFM handles multimodal features along the channel axis. Unlike previous methods which use a single attention matrix to weigh different branches, our approach capitalizes on both consistent and complementary features. The architecture of the SFM is shown in Fig. 3. Its design allows the SFM to compute attention matrices for individual branches, focusing on distinctive and complementary features between them. This provides a robust foundation for subsequent network layers to accurately calibrate and align features from the two modalities.



**Fig. 2.** Overall Network Structure of ASANet: The SFM concentrates on independently complementary features of different modalities, and the CFM calibrates and aligns feature information. ( $\oplus$  denotes the pixel-wise add operation).



**Fig. 3.** Diagram of the structure of the SFM.

Specifically, the process comprises three main steps. We denote the feature maps of the two modalities as  $F_{RGB} \in \mathbb{R}^{C \times H \times W}$  and  $F_{SAR} \in \mathbb{R}^{C \times H \times W}$ , respectively, where  $C$ ,  $H$ , and  $W$  correspond to the number of channels, height, and width of the feature maps. Superscripts  $r$  (RGB) and  $s$  (SAR) are utilized to distinguish between the feature maps of the two modalities.

**Obtain differential features:** As previously mentioned, to capture the complementarity between different modalities, we extract complementary features from each modality separately. Following the principle of differential amplification (Qingyun and Zhaokui, 2022), the SFM performs a pixel-wise subtraction between the  $F_{RGB}$  and  $F_{SAR}$  feature maps to obtain the differential feature map  $F_{\Delta}^r \in \mathbb{R}^{C \times H \times W}$  for the RGB modality. Conversely, for the SAR modality, a subtraction is performed between  $F_{RGB}$  and  $F_{SAR}$  to obtain the differential feature map  $F_{\Delta}^s \in \mathbb{R}^{C \times H \times W}$ . The first step can be described as Eq. (1):

$$\begin{aligned} F_{\Delta}^r &= F_{RGB} - F_{SAR} \\ F_{\Delta}^s &= F_{SAR} - F_{RGB} \end{aligned} \quad (1)$$

where  $-$  denotes the pixel-wise subtraction operation.

**Adaptive refinement learning:** Furthermore, we apply global max-pooling to the differential feature maps  $F_{\Delta}^r$  and  $F_{\Delta}^s$  to obtain channel weights  $s_1^r \in \mathbb{R}^{C \times 1 \times 1}$  and  $s_1^s \in \mathbb{R}^{C \times 1 \times 1}$ . Subsequently, we introduce

multiple convolutional modules resembling channel-wise perceptrons to perform fine-grained learning on  $s_1^r$  and  $s_1^s$ , resulting in global differential weights  $s_2^r \in \mathbb{R}^{C \times 1 \times 1}$  and  $s_2^s \in \mathbb{R}^{C \times 1 \times 1}$ . Eq. (2) describes the second step.

$$\begin{aligned} s_1^r &= \mathcal{F}_{GMP}(F_{\Delta}^r) \\ s_2^r &= \mathcal{F}_2(\mathcal{F}_{GELU}(\mathcal{F}_1(s_1^r))) \\ s_1^s &= \mathcal{F}_{GMP}(F_{\Delta}^s) \\ s_2^s &= \mathcal{F}_2(\mathcal{F}_{GELU}(\mathcal{F}_1(s_1^s))) \end{aligned} \quad (2)$$

where  $\mathcal{F}_{GMP}$  denotes global max-pooling,  $\mathcal{F}_1$  represents a convolutional layer with an output dimension of  $\max(32, \frac{C}{16})$ ,  $\mathcal{F}_2$  is a convolutional layer with an output dimension of  $C$ , and  $\mathcal{F}_{GELU}$  signifies the GELU activation function.

**Correction of parallel branches:** The formula for correcting the original feature maps is as follows:

$$\begin{aligned} F_{SFM}^r &= F_{RGB} \otimes \sigma(s_2^r) \\ F_{SFM}^s &= F_{SAR} \otimes \sigma(s_2^s) \end{aligned} \quad (3)$$

where  $F_{SFM}^r \in \mathbb{R}^{C \times H \times W}$  and  $F_{SFM}^s \in \mathbb{R}^{C \times H \times W}$  represent the output feature maps of the dual branches, respectively.  $\sigma$  is the sigmoid function and  $\otimes$  is the pixel-wise multiplication.

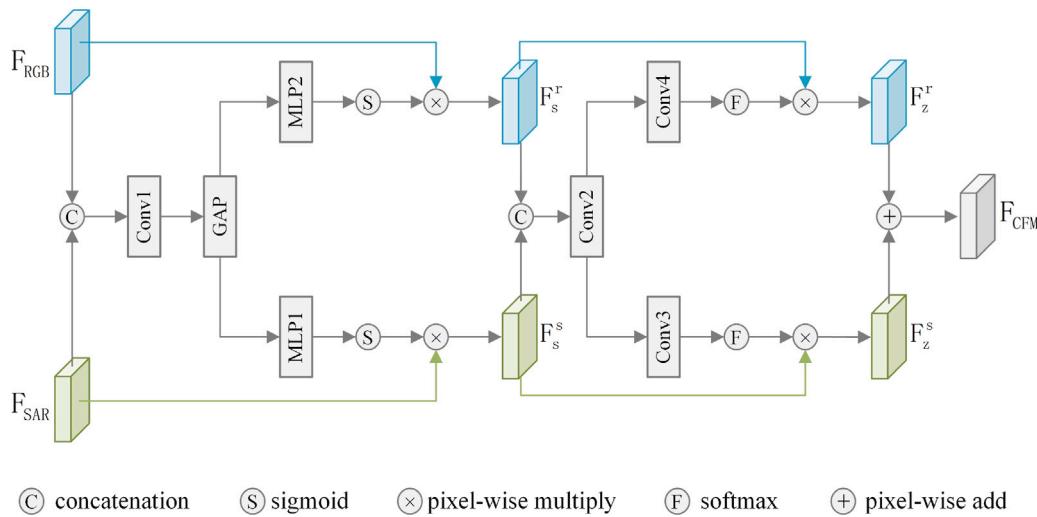


Fig. 4. Diagram of the structure of the CFM.

### 3.3. Cascade fusion module

The feature information in RGB and SAR images is often complementary. The challenge lies in effectively calibrating and aligning the complementary information from both modalities. Traditional feature fusion strategies, such as simple concatenation or summation of feature maps, may inadvertently amplify the noise present in both modalities, thereby negatively impacting the final classification performance. To overcome this issue, inspired by CBAM (Woo et al., 2018), we propose a CFM for feature fusion. As illustrated in Fig. 4, the CFM adopts a concatenated attention mechanism to learn deep representations of RGB and SAR features in both channel and spatial dimensions. This fusion approach can adaptively calibrate and align the complementary information from the two modalities, reducing the interference of noise information and improving the model's performance.

Suppose the input feature maps to the CFM are  $F_{\text{RGB}} \in \mathbb{R}^{C \times H \times W}$  and  $F_{\text{SAR}} \in \mathbb{R}^{C \times H \times W}$ , where C, H, and W represent the number of channels, height, and width of the feature maps, respectively. The superscripts r (RGB) and s (SAR) are used to distinguish between the feature maps of the two modalities.

In the first attention module, we align and calibrate the features from the two branches along the channel dimension. Initially, the feature maps  $F_{\text{RGB}}$  and  $F_{\text{SAR}}$  are concatenated. Subsequently, the Concatenated feature map is fused through convolutional layers, followed by global average-pooling to obtain the feature vector  $s \in \mathbb{R}^{C \times 1 \times 1}$ . GAP generate channel-wise statistics by shrinking concatenated feature through its spatial dimensions. The feature vector  $s$  is then input into two separate Multi-Layer Perceptrons (MLPs) with non-shared weights to capture channel-wise dependencies. After passing through the sigmoid function, different fusion channel weights are obtained for each modality. These weights are then multiplied with the original feature maps to obtain the feature maps  $F_s^r \in \mathbb{R}^{C \times H \times W}$  and  $F_s^s \in \mathbb{R}^{C \times H \times W}$ . The process is described in formulaic terms as follows:

$$\begin{aligned} s &= \mathcal{F}_{\text{GAP}}(\mathcal{F}_1^c(\text{cat}(F_{\text{RGB}}, F_{\text{SAR}}))) \\ F_s^r &= F_{\text{RGB}} \otimes \sigma(\mathcal{F}_{\text{MLP}_1}(s)) \\ F_s^s &= F_{\text{SAR}} \otimes \sigma(\mathcal{F}_{\text{MLP}_2}(s)) \end{aligned} \quad (4)$$

where  $\text{cat}$  refers to concatenating feature maps along the channel dimension,  $\mathcal{F}_{\text{GAP}}$  denotes global average-pooling,  $\mathcal{F}_1^c$  represents a convolutional layer with an input dimension of 2C and an output dimension of c.  $\mathcal{F}_{\text{MLP}_1}$  and  $\mathcal{F}_{\text{MLP}_2}$  signify MLP.  $\sigma$  represents the sigmoid activation function,  $\otimes$  denotes the pixel-wise multiplication operation. CFM learn different channel attention weights, enabling the adaptive selection and calibration of the original features from the two branches.

In the second attention module, we align and calibrate the features from the two branches along the spatial dimensions. Initially, the feature maps  $F_s^r$  and  $F_s^s$  are concatenated, and the concatenated feature map undergoes processing through a convolutional layer to fuse the information, yielding feature map z. This feature map  $z \in \mathbb{R}^{1 \times H \times W}$  is then convolved with two sets of distinct weights to enhance the feature representation. The softmax function is applied to derive the weight scores for the respective modal space dimensions, which represent the importance of each pixel. These scores are used to weight the feature maps  $F_s^r$  and  $F_s^s$ , resulting in the fusion spatial attention weighted feature maps  $F_z^r$  and  $F_z^s$ .

$$\begin{aligned} z &= \mathcal{F}_2^c(\text{cat}(F_s^r, F_s^s)) \\ F_z^r &= F_s^r \otimes \mathcal{F}_{\text{softmax}}(\mathcal{F}_3(z)) \\ F_z^s &= F_s^s \otimes \mathcal{F}_{\text{softmax}}(\mathcal{F}_4(z)) \end{aligned} \quad (5)$$

where  $\text{cat}$  refers to concatenating feature maps along the channel dimension,  $\mathcal{F}_2^c$  represents a convolutional layer with an input dimension of 2C and an output dimension of c,  $\mathcal{F}_3$  and  $\mathcal{F}_4$  represent convolutional layers with both input and output dimensions C.  $\otimes$  represents pixel-by-pixel multiplication, and  $\mathcal{F}_{\text{softmax}}$  is a softmax function. CFM calculate the spatial attention weights for different modalities, adaptively strengthening the effective features and suppressing the invalid ones.

Finally, According to Eq. (6),  $F_z^r$  and  $F_z^s$  are pixel-wise added to produce the composite output feature map  $F_{\text{CFM}} \in \mathbb{R}^{C \times H \times W}$ .

$$F_{\text{CFM}} = F_z^r + F_z^s \quad (6)$$

where + denotes the pixel-wise addition (PWA) operation.

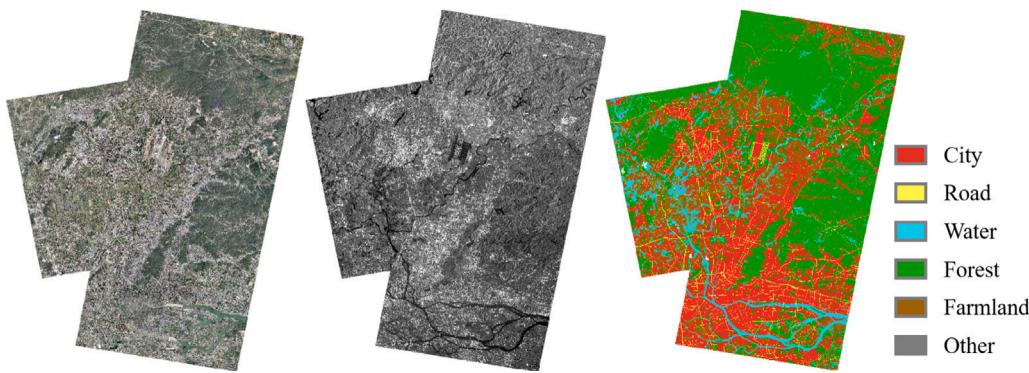
## 4. Experimental and datasets

This section describes the proposed refined annotated semantic segmentation multimodal dataset. Furthermore, a comparative analysis is performed on this dataset, in conjunction with two publicly available datasets, utilizing six other multimodal segmentation algorithms. This comparison aims to highlight the superior performance of the proposed algorithm. The experimental results consistently demonstrate the robustness of our algorithm, affirming its efficacy in handling diverse multimodal datasets.

### 4.1. Datasets

#### 4.1.1. PIE-RGB-SAR

As shown in Fig. 5, the study area for this paper is selected as a part of the Pearl River Delta, China. The RGB image is sourced from



**Fig. 5.** PIE-RGB-SAR dataset: RGB image on the left, SAR image in the middle, and ground truth image on the right.

Google satellite data with a spatial resolution of approximately 0.5 m. The SAR image is derived from the GF3 satellite's ultra-fine stripe mode, with a spatial resolution of 3 m. To ensure the data resolution of the two modes remains consistent, we utilize PIE-Basic software for resampling and aligning the two images. To reduce interference from extreme radiation intensity during the network's learning from SAR images and to facilitate network convergence, radiation values in SAR images are truncated and linearly adjusted to range between 0 and 255.

Considering that the RGB and SAR images were not captured simultaneously, there are regions where feature changes have occurred across different categories. In this dataset, the RGB image serves as the primary source, with the SAR image offering complementary information to aid in identifying the land type associated with each pixel. The dataset is annotated into six categories: city, road, water, forest, farmland, and others. It is noteworthy that the dataset exhibits an imbalance in data distribution, with distinct proportions for each feature type. Specifically, the distribution consists of 30% for city, 3% for road, 7% for water, 42% for forest, 10% for farmland, and 8% for other categories. The images were cropped to a uniform size of  $256 \times 256$ , yielding a total of 4865 images. These images were subsequently partitioned into training and validation sets at a 1:1 ratio, comprising 2433 images in the training set and 2432 images in the validation set.

#### 4.1.2. DDHR-SK

Ren et al. (2022) provided several multimodal datasets, which are divided into five image groups and sourced from three distinct regions. Our experiments were conducted using data from the cloudy group in the Pohang region of South Korea, known as DDHR-SK, where the data division is publicly available. The dataset encompasses five categories: city, road, forest, water, and farmland. The overall preview of the DDHR-SK dataset is shown in Fig. 6. The origin RGB image is sourced from GF2 satellite data. The RGB images of the DDHR-SK dataset were processed using the GNU Image Manipulation Program (GIMP) to simulate a cloudy scene. The SAR image is sourced from GF3 satellite data and has been adjusted to have a resolution of 1 meter. The RGB and SAR images are cropped to a pixel size of  $256 \times 256$ . There are 3087 images for training and 3086 images for verification. This dataset aims to capture the challenges posed by foggy and cloudy conditions and serves as a valuable resource for evaluating multimodal segmentation algorithms.

#### 4.1.3. WHU-OPT-SAR

The WHU-OPT-SAR dataset (Li et al., 2022b) is a multimodal dataset that features a study area in Hubei, China. The dataset comprises a total of 100 RGB and SAR images, each with dimensions of  $5536 \times 3704$  pixels. The RGB images are sourced from GF1 satellite data, while the SAR images are obtained from GF3 satellite data, with a resolution of 5 m. One of the dataset images is shown in Fig. 7. In this paper, the images are cropped to a size of  $512 \times 512$  pixels, resulting in

**Table 1**  
Settings of hyperparameters for each model.

Model	Batchsize	Optimizer	Ir	Momentum/Betas	Weight-decay
UPerNet	8	AdamW	0.0001	(0.9,0.999)	0.05
		SGD	0.01	0.9	0.0005
		SGD	0.01	0.9	0.0005
		SGD	0.01	0.9	0.0005
		SGD	0.01	0.9	0.0005
		AdamW	0.00006	(0.9,0.999)	0.01
		SGD	0.01	0.9	0.0005
		AdamW	0.0001	(0.9,0.999)	0.05

a total of 7040 images for training and 1760 images for validation. The dataset is annotated into seven categories: farmland, city, village, water, forest, road, and other. This dataset provides a valuable resource for evaluating multimodal segmentation algorithms, particularly in the context of diverse land cover types in Hubei, China.

#### 4.2. Experimental setups

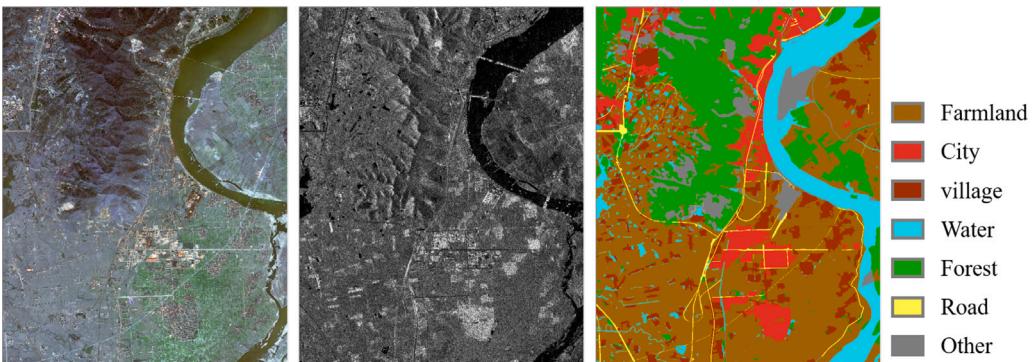
In this paper, all model training and testing are conducted on two Quadro RTX 8000 GPUs, which are running on the Ubuntu 18.04 Linux platform. The batch size per GPU is set to 4, and a total of 80,000 iterations are performed. The model is validated every 8000 iterations. The metrics reported in this paper are indicators of the number of iterations corresponding to the best mIoU. The results for all models are based on the accuracy of our own reproduction. The hyperparameter settings for each model are listed in Table 1. Specifically, pre-trained weights from ImageNet22k are adopted to initialize two ConvNeXtV2-tiny (Woo et al., 2023) backbone networks within ASANet. The primary data augmentation techniques include random scaling, random cropping, and random flipping. To ensure the fairness of the experiment, all experimental configurations outlined in this paper were strictly adhered to throughout the study. The evaluation metrics include mIoU, Kappa, and Overall Accuracy (OA), which are commonly employed in semantic segmentation assessments. We also used floating point operation count (FLOPs), the number of model parameters, and FPS to evaluate the complexity of the model.

#### 4.3. Experimental

To evaluate the models' performance, we conducted a detailed comparison of six multimodal semantic segmentation models: CMX (Zhang et al., 2023b), SA-Gate (Chen et al., 2020), FTransUNet (Ma et al., 2024), CMFNet (Ma et al., 2022b), AFNet (Xu et al., 2021), and FuseNet (Hazirbas et al., 2017). The mIoU performance of various methods on three datasets is presented in Table 2, where our proposed ASANet demonstrates SOTA results. ASANet inference is 15.2 FPS faster than the suboptimal model CMX. Furthermore, we tested the UperNet model, which employs the ConvNeXtV2-tiny backbone,



**Fig. 6.** DDHR-SK dataset: Original RGB image on the left, RGB image with simulated cloudy in the center, and SAR image on the right.



**Fig. 7.** Part of the WHU-OPT-SAR dataset: RGB image on the left, SAR image in the center, and ground truth image on the right.

**Table 2**

The complexity computation metrics for a  $256 \times 256$  image on a single Quadro RTX 8000 GPU are as follows. The mIoU metrics (%) for different methods across three datasets are presented below. The best results are presented in bold, while the second-best results are underlined.

Model	Year	FLOPs	Parameter	Speed	mIoU					
					G	M	FPS	PIE-RGB-SAR	DDHR-SK	WHU-OPT-SAR
FuseNet	2017	66	55	<b>88.8</b>	60.62	48.87	38.01			
SA-Gate	2020	46	121	34.9	73.84	90.89	53.17			
AFNet	2021	65	356	35.9	76.27	91.11	53.57			
CMFNet	2022	77	104	21.6	76.31	89.79	53.72			
CMX	2023	<u>15</u>	<u>67</u>	33.5	<u>77.10</u>	<u>94.32</u>	<u>55.68</u>			
FTransUNet	2024	70	203	20.7	75.72	87.64	54.47			
<b>ASANet (ours)</b>		<u>25</u>	82	<u>48.7</u>	<b>78.31</b>	<b>94.48</b>	<b>56.11</b>			

for single-modal experiments. All models were configured to ensure comparability, facilitating a fair evaluation of both multimodal and single-modal segmentation capabilities.

#### 4.3.1. Results on PIE-RGB-SAR

In Table 3, we compare the performance of single model and six multimodal algorithms. In the evaluation of single model training, the RGB achieves a mIoU that is 10.17% higher than that of SAR, suggesting its significant contribution to feature learning. ASANet (mIoU=78.31%, Kappa=85.27%, OA=89.64%) surpasses other methods in all metrics across both modalities and modules. Section 5.1.2 offers a detailed analysis of these results.

ASANet achieves notable improvements in mIoU over FuseNet, SA-Gate, FTransUNet, CMFNet, AFNet, and CMX, with increases of 15.65%, 4.47%, 2.59%, 2%, 2.04%, and 1.21%, respectively. FuseNet demonstrates overall lower performance. SA-Gate also underperforms on this dataset, attaining a mIoU of only 73.84%. FTransUNet and CMFNet, networks specifically designed for remote sensing imagery, also perform poorly in terms of accuracy. CMX, a fusion algorithm

designed for autonomous driving, achieves a score of 77.1% on this dataset, which is lower than the scores achieved by ASANet.

Table 3 also presents the accuracy of different methods in terms of IoU metrics for individual categories. The ASANet model consistently attains the highest IoU values across all categories. Compared to the SA-Gate network, ASANet shows an 8.87% higher IoU<sub>road</sub> and a 6.8% higher IoU<sub>water</sub>. Against the sub-high precision network CMX, ASANet performs better across multiple categories, with improvements of 0.44%, 2.61%, 1.95%, 0.54%, and 1.53% for the city, road, water, forest, and farmland categories, respectively. It is worth noting that the performance of the networks FTransUNet, CMFNet, and ASANet, which are designed for remote sensing image, is outstanding in water categories. ASANet's performance excels in the road and water categories.

The visualization results of each model on the PIE-RGB-SAR dataset are shown in Fig. 8. The sequence, from left to right, includes the RGB image, SAR image, ground truth label, CMFNet result, AFNet result, CMX result, and ASANet result. This visualization results compellingly evidence the superior performance of ASANet. Notably, our model

**Table 3**

Metrics results (%) for different methods on PIE-RGB-SAR dataset. The best results are presented in bold, while the second-best results are underlined.

Model	Modality		Kappa	OA	mIoU	IoU					
	RGB	SAR				City	Road	Water	Forest	Farmland	Other
UPerNet	✓		84.13	88.84	76.90	81.77	60.53	75.62	82.15	64.40	97.27
UPerNet		✓	74.16	83.82	66.73	75.57	33.69	70.61	75.76	49.04	95.73
FuseNet	✓	✓	70.79	79.89	60.62	70.56	34.47	60.41	71.01	29.93	97.31
SA-Gate	✓	✓	82.24	87.59	73.84	80.64	52.93	71.03	80.37	60.95	97.11
AFNet	✓	✓	83.95	88.71	76.27	81.90	58.30	75.03	82.10	64.06	96.34
CMFNet	✓	✓	83.99	88.74	76.31	82.33	57.46	77.28	82.06	61.76	96.96
CMX	✓	✓	<u>84.60</u>	<u>89.18</u>	<u>77.10</u>	<u>82.41</u>	<u>59.19</u>	75.88	<u>82.73</u>	<u>65.22</u>	97.06
FTransUNet	✓	✓	83.53	88.38	75.72	81.86	56.66	77.01	81.67	59.91	97.21
<b>ASANet (ours)</b>	✓	✓	<b>85.27</b>	<b>89.64</b>	<b>78.31</b>	<b>82.85</b>	<b>61.80</b>	<b>77.83</b>	<b>83.27</b>	<b>66.75</b>	<b>97.32</b>

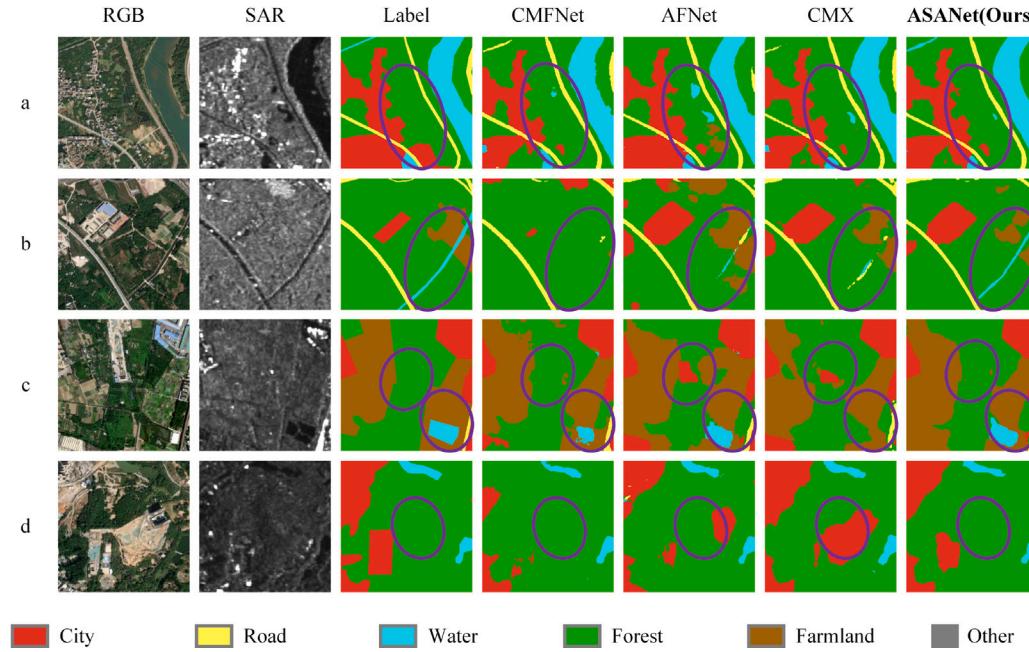


Fig. 8. Visualization results of different methods on PIE-RGB-SAR dataset: four representative multimodal models.

exhibits improved continuity in identifying road and water features. In Fig. 8, groups a, b, and c show that AFNet algorithm is susceptible to SAR images, leading to confusion between forest and farmland. Conversely, ASANet demonstrates resistance to noise in SAR images. Groups c and d of Fig. 8 highlight ASANet's effective use of SAR information, addressing issues of omission and misjudgment present in RGB images. The visualization results underscore that ASANet skillfully integrates information from both modalities, selectively incorporating more effective data for learning.

For a more intuitive comparison, we specifically select images with cloud cover for feature map visualization. As shown in Fig. 9, there is a clear distinction when comparing the RGB and SAR branch features after CMX's stage-3 feature interaction (Fig. 9(c) and (h)) with those after ASANet's stage-3 feature interaction (Fig. 9(d) and (i)). In the case of CMX, the feature blending results in highly similar feature inductions for different branches. Conversely, our network, which employs the SFM strategy, fosters relatively independent features for the two branches, introducing diverse feature information from each branch. When comparing the feature maps in Fig. 9(j) after CMX's stage-3 feature fusion and Fig. 9(e) after ASANet's stage-3 feature fusion, our network consistently detects regional features of the same category with reduced noise. This results in less fragmented small areas of false detection in the segmentation results of ASANet. By utilizing the fusion strategy of SFM and CFM, ASANet excels in self-adapting the learning of

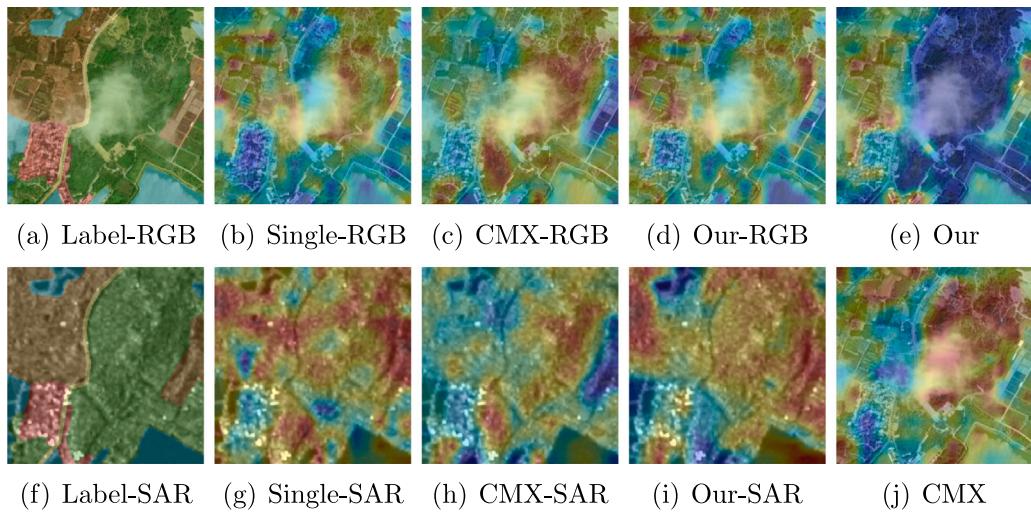
both modalities, effectively incorporating SAR features while mitigating noise information. This strategic approach enhances the robustness and overall segmentation performance of ASANet.

#### 4.3.2. Results on DDHR-SK

In the DDHR-SK dataset, the optical modes consist of RGB images simulated to emulate imaging under cloudy weather conditions, adding variability to the dataset. Notably, ASANet has exhibited the most effective detection performance among other models on this dataset. This demonstrates the robustness and adaptability of ASANet to challenging scenarios, including those with simulated cloud cover in optical images.

Table 4 displays the incremental results for different models and IoU metrics across all classes in the DDHR-SK dataset. When considering mIoU accuracy, the models are ranked in the following order: ASANet, CMX, AFNet, SA-Gate, CMFNet, FTransUNet and FuseNet. Each network exhibits strong performance on this dataset, aided by its high resolution. Notably, it is crucial to emphasize that the simulated images in this dataset depict scenes under cloudy conditions. As a result, the performance gap between RGB and SAR modes is substantially narrowed, suggesting that the network can still effectively learn feature information from RGB imagery despite the challenging cloud cover.

Analyzing the IoU scores for each category, the  $\text{IoU}_{\text{water}}$  for RGB images in the single model approaches saturation, whereas the  $\text{IoU}_{\text{water}}$  for SAR images in the single model is similar to that for RGB images.

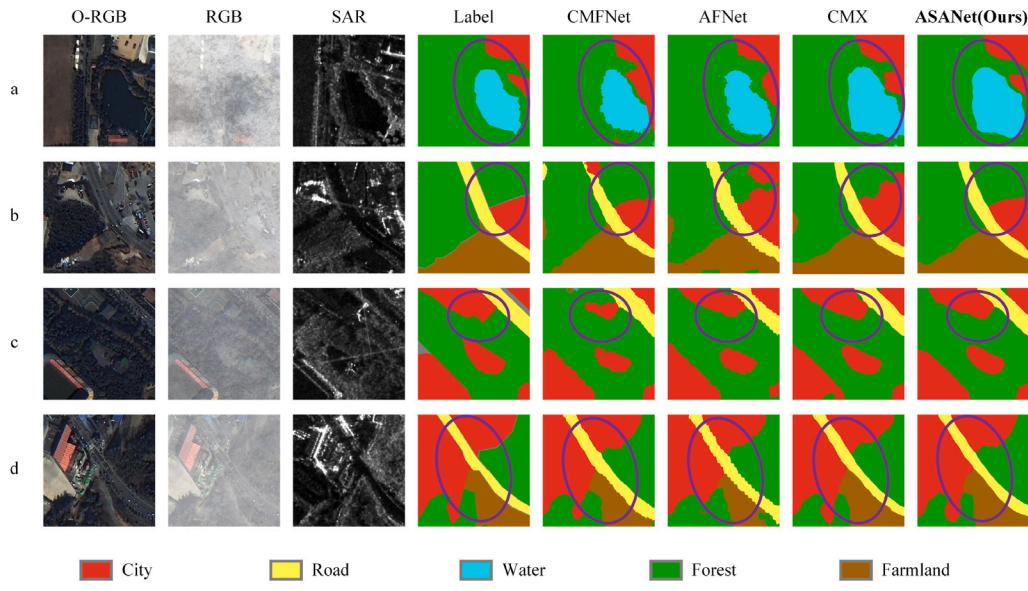


**Fig. 9.** Feature map visualization: (a) and (f) show the superposition ground truth of RGB and SAR images, respectively; (b) and (g) depict the output of the stage-3 backbone network of the RGB and SAR single-mode models, respectively; (c) and (h) represent the output feature map of the RGB and SAR branches after the stage-3 feature interaction of CMX (i.e., the output feature map of CM-FRM). (d) and (i) illustrate the RGB and SAR branch output feature map after the stage-3 feature interaction of our ASANet network (i.e.,  $F_{\text{RGB}}$  and  $F_{\text{SAR}}$  in Fig. 4); (e) shows the output feature map  $F_{\text{CFM}}$  of CFM stage-3 feature fusion in our ASANet network; (j) presents the output feature map of the FFM stage-3 in CMX.

**Table 4**

Metrics results (%) for different methods on DDHR-SK dataset. The best results are presented in bold, while the second-best results are underlined.

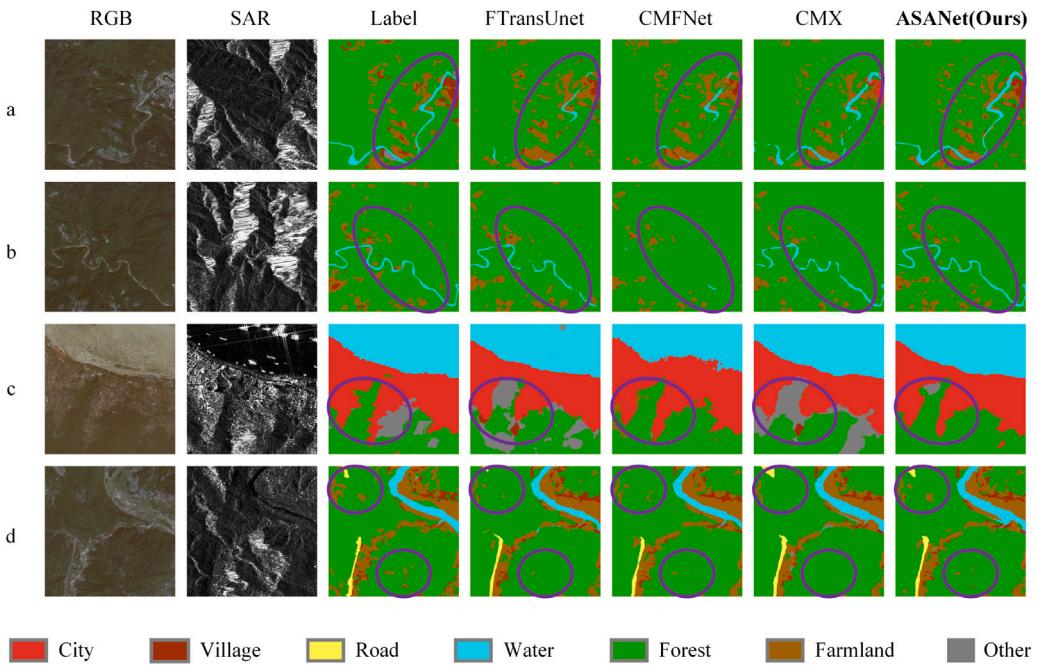
Model	Modality		Kappa	OA	mIoU	IoU				
	RGB	SAR				City	Road	Forest	Water	Farmland
UPerNet	✓		95.39	96.74	91.42	95.09	78.68	92.22	99.15	91.99
UPerNet		✓	92.40	94.63	85.72	92.79	64.70	88.10	98.30	84.74
FuseNet	✓	✓	60.31	73.62	48.87	64.00	1.76	47.80	89.43	41.35
SA-Gate	✓	✓	95.28	96.66	90.89	95.20	76.65	92.36	99.09	91.17
AFNet	✓	✓	95.27	96.65	91.11	95.32	78.91	92.14	98.76	90.42
CMFNet	✓	✓	94.52	96.12	89.79	94.58	76.07	90.91	98.7	88.68
CMX	✓	✓	<u>97.09</u>	<u>97.94</u>	<u>94.32</u>	<u>96.95</u>	<u>85.39</u>	<u>95.12</u>	<u>99.46</u>	<u>94.69</u>
FTransUNet	✓	✓	93.12	95.14	87.64	93.07	71.67	88.71	98.2	86.62
<b>ASANet (ours)</b>	✓	✓	<b>97.22</b>	<b>98.03</b>	<b>94.48</b>	<b>97.04</b>	<u>85.36</u>	<u>95.47</u>	<u>99.49</u>	<u>95.03</u>



**Fig. 10.** Visualization results of different methods on DDHR-SK dataset: four representative multimodal models.

With the inclusion of the SAR modality, both CMX and ASANet exhibit improved accuracy for the water category. Notably, ASANet shows a larger increase, indicating a superior integration of SAR features. Compared to a single model using only RGB images, ASANet exhibits

substantial improvements of 6.45%, 3.33%, and 4.61% for the road, forest, and farmland categories, respectively, after integrating SAR features. This highlights the superior performance of ASANet in cloudy and foggy environments.



**Fig. 11.** Visualization results of different methods on WHU-OPT-SAR dataset: four representative multimodal models.

Visualization results of different models on the simulated DDHR-SK dataset with clouds are presented in Fig. 10. In this dataset, cloud occlusion reduces the texture features in RGB images, while the boundary features of roads and water remain distinct in SAR images. In visualization results, ASANet outperforms other models, producing overall superior results with smoother boundaries and fewer fine details. In Fig. 10, group a, ASANet demonstrates excellent performance by effectively integrating water features from SAR images, achieving superior detection results compared to other models. ASANet achieves higher accuracy with minimal instances of missing and false detections, distinguishing itself from CMFNet, AFNet, and CMX models. For road categories in Fig. 10, groups b, c, and d, the detection and annotation results of ASANet show relatively good overlap, further illustrating the effectiveness of ASANet in cloud-occluded scenarios.

#### 4.3.3. Results on WHU-OPT-SAR

The results of different models on the WHU-OPT-SAR dataset are shown in Table 5. ASANet achieves the best overall results, with mIoU scores that are 15.56%, 2.94%, 1.64%, 2.39%, 2.54%, and 0.43% higher than those of FuseNet, SA-Gate, FTransUNet, CMFNet, AFNet, and CMX, respectively. These results underscore the superiority of ASANet across various metrics on the WHU-OPT-SAR dataset. The visualization results are shown in Fig. 11. By comparing Fig. 11 group a and b, it can be observed that ASANet better combines the advantages of SAR images and demonstrates better connectivity in the water category. Fig. 11 group c shows that CMFNet and our visualization method performed the best overall. However, the detection results of CMFNet contain many misdetected small patches. In Fig. 11 group d, the ASANet model proposed by us accurately categorized the cultivated land more than other models.

## 5. Discussion

As described in Section 4, ASANet achieved the best results on the PIE-RGB-SAR, DDHR-SK, and WHU-OPT-SAR datasets. In this section, the paper discusses the influence of each module and the fusion of different stages on the model's performance. Additionally, we simulate cloud and fog scenarios on RGB images and conduct comparative experiments. All settings in each experiment are consistent with those of the multimodal fusion model in the comparative experiment.

### 5.1. Ablation experiments between modules

In this section, we conduct ablation experiments based on the PIE-RGB-SAR dataset. The results of these ablation experiments between modules are presented in Table 6. Specifically, the third row of results shows the twin network framework with the SAR mode, which uses the PWA (as described in Eq. (6)) to fuse feature maps of the two modes before they are input into the decoder. To evaluate the effectiveness of the CFM, we replace PWA with the CFM. In assessing the effectiveness of the SFM, we integrate the SFM for feature interaction within the network framework, as shown in the third and fourth rows.

#### 5.1.1. Analysis of results for different modes

Comparing the results of the RGB-Single model (first row) with those of the SAR-Single model (second row), it is evident that the overall detection accuracy of RGB is higher than that of SAR. Notably, the  $\text{IoU}_{\text{road}}$  and  $\text{IoU}_{\text{farmland}}$  show significant differences, standing at 26.84% and 15.36%, respectively. The  $\text{IoU}_{\text{city}}$ ,  $\text{IoU}_{\text{water}}$ , and  $\text{IoU}_{\text{forest}}$  exhibit smaller differences, at 6.2%, 5.01%, and 6.39%, respectively. In the multimodal model RGB-SAR-PWA (third row), which incorporates the SAR modality, and in our proposed ASANet model, the IoU indicators for each category demonstrate improvements. These results further confirm that the texture and color features in RGB images contribute to higher accuracy, while the inclusion of SAR provides crucial invisible feature information missing in RGB. Moreover, compared to the RGB-SAR-PWA model (third row), our ASANet model (sixth row) improves mIoU by 0.5%, Kappa by 0.39%, and OA by 0.28%. With the combined influence of both modules, our ASANet achieves the best IoU indicators for all categories, demonstrating superior semantic segmentation performance.

In Fig. 12, we visualized the results of ASANet compared to a single-modal model on the PIE-RGB-SAR dataset. The sequence, from left to right, includes the RGB image, SAR image, ground truth image, RGB-Single result, SAR-Single result, RGB-SAR-PWA result, and RGB-SAR-SFM-CFM (ASANet) result. This visualization illustrates the effectiveness of our model in utilizing both RGB and SAR modalities to enhance semantic segmentation outcomes. The analysis of the visualization results reveals that the SAR-Single model effectively captures edge features of roads and large water bodies but struggles to identify slender roads and water bodies and has difficulty distinguishing between

**Table 5**

Metrics results (%) for different methods on WHU-OPT-SAR dataset. The best results are presented in bold, while the second-best results are underlined.

Model	Modality		Kappa	OA	mIoU	IoU						
	RGB	SAR				City	Village	Water	Road	Forest	Farmland	Other
UPerNet	✓		76.13	<u>84.36</u>	55.84	<b>56.80</b>	<b>49.32</b>	64.56	<u>41.88</u>	83.70	69.95	24.70
UPerNet		✓	69.96	80.59	48.42	55.31	35.74	56.95	28.85	80.76	62.76	18.58
FuseNet	✓	✓	65.67	77.67	38.01	27.94	36.19	49.35	13.84	79.55	59.19	00.00
SA-Gate	✓	✓	74.13	83.13	53.17	53.28	47.39	62.14	39.19	82.20	68.05	19.93
AFNet	✓	✓	74.96	83.69	53.57	54.01	46.93	62.84	39.41	83.03	69.06	19.74
CMFNet	✓	✓	75.14	83.76	53.72	55.36	48.45	62.95	37.91	83.13	68.93	19.25
CMX	✓	✓	<u>76.15</u>	<u>84.36</u>	55.68	56.20	49.02	<u>65.03</u>	41.03	<u>83.72</u>	<u>69.95</u>	<b>24.79</b>
FTransUNet	✓	✓	75.47	83.92	54.47	55.85	48.48	63.86	39.52	83.25	69.44	20.87
<b>ASANet (ours)</b>	✓	✓	<b>76.43</b>	<b>84.56</b>	<b>56.11</b>	<u>56.79</u>	<u>49.17</u>	<b>65.28</b>	<b>42.72</b>	<b>83.89</b>	<b>70.33</b>	24.59

**Table 6**

Metric results (%) of ablation experiments on PIE-RGB-SAR dataset.

Modality	SFM	CFM	PWA	Kappa	OA	mIoU	IoU					
							City	Road	Water	Forest	Farmland	Other
RGB				84.13	88.84	76.90	81.77	60.53	75.62	82.15	64.40	97.27
SAR				74.16	83.82	66.73	75.57	33.69	70.61	75.76	49.04	95.73
RGB+SAR		✓		84.88(+0.75)	89.36(+0.52)	77.81(+0.91)	82.55	61.50	76.94	82.90	65.68	97.27
RGB+SAR	✓			85.03(+0.90)	89.47(+0.63)	78.05(+1.15)	82.67	61.63	77.76	82.98	66.04	97.20
RGB+SAR	✓	✓		85.03(+0.15)	89.46(+0.10)	78.01(+0.20)	82.58	<b>61.80</b>	76.78	83.06	66.52	97.30
RGB+SAR	✓	✓	✓	<b>85.27(+0.24)</b>	<b>89.64(+0.17)</b>	<b>78.31(+0.26)</b>	<b>82.85</b>	<b>61.80</b>	<b>77.83</b>	<b>83.27</b>	<b>66.75</b>	<b>97.32</b>

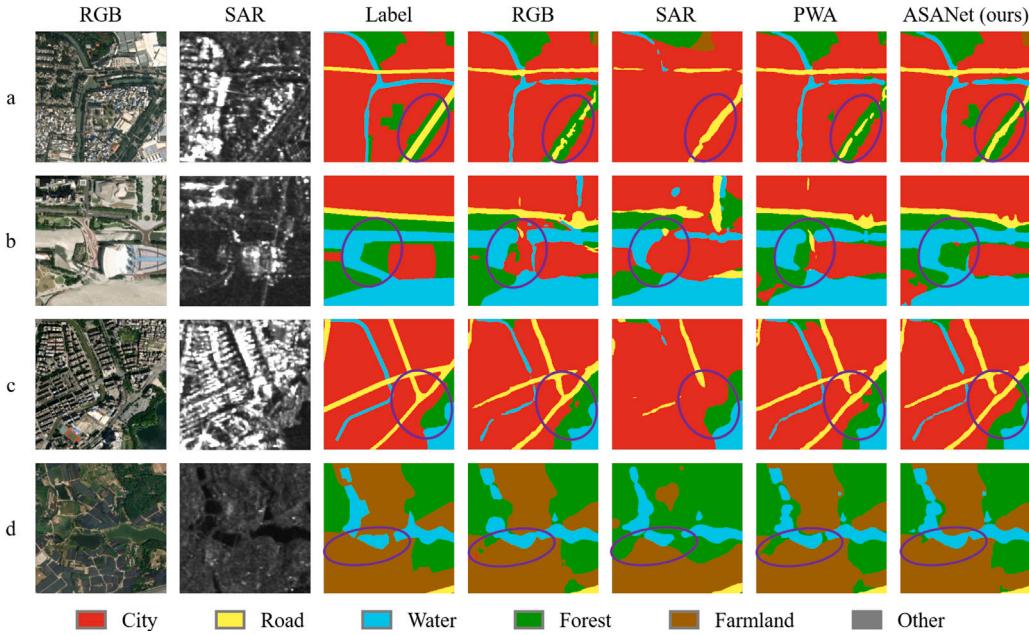


Fig. 12. Visualization results of ablation experiments.

farmland and forest. Conversely, the RGB-Single model exhibits clear texture features, resulting in accurate overall category recognition. However, they are susceptible to interference from similar texture features at the edges, resulting in the appearance of fine image spots. Specifically, observations from Fig. 12, groups a (road), b (water), and c (forest) indicate that the RGB-SAR-PWA model tends to favor learning the features of RGB. In complex scenes, such as those in Fig. 12, groups b and d, where texture and color features in RGB struggle to effectively classify categories, the RGB-SAR-PWA model is prone to introducing noise. For example, as evidenced by the miss segmentation of farmland as forest in Fig. 12 d. In contrast, our model exhibits an overall segmentation effect that mitigates the shortcomings of the two modalities. It successfully fuses complementary features from both modalities, resulting in a superior visualization effect overall. For each category, our model produces more connected segmentation results for

roads and rivers and achieves greater accuracy in extracting cities, farmland, and forests.

### 5.1.2. Analysis of results for different modules

From Table 6, we observe the influence of different modules on the accuracy of each category using RGB-SAR-PWA as the baseline.

**SFM.** When comparing the RGB-SAR-SFM-PWA model (fifth row) with the RGB-SAR-PWA model (third row), the most significant improvement in  $\text{IoU}_{\text{road}}$  and  $\text{IoU}_{\text{farmland}}$  is observed. In comparing the RGB-SAR-CFM model (fourth row) with the RGB-SAR-SFM-CFM model (sixth row), the enhancement in the farmland class becomes more pronounced when incorporating the SFM. Notably, these two categories are precisely those that exhibit poor performance in SAR modes. This indicates that the SFM enables the network model to flexibly adjust to feature differences between different modalities, thereby enhancing the segmentation performance for various categories.

**Table 7**

Metric results (%) of ablation experiments performed between different stages on the PIE-RGB-SAR dataset.

Stage	Kappa		OA		mIoU		IoU						
	1	2	3	4			City	Road	Water	Forest	Farmland	Other	
					84.88	89.36	77.81	82.55	61.5	76.94	82.9	65.68	97.27
✓	✓				84.85(−0.03)	89.35(−0.01)	77.80(−0.01)	82.55	61.62	76.92	82.86	65.51	97.31
		✓			85.09(+0.21)	89.51(+0.15)	78.10(+0.29)	82.70	61.77	77.50	83.07	66.31	97.27
	✓		✓		84.97(+0.09)	89.43(+0.07)	77.92(+0.11)	82.77	61.93	76.94	82.98	65.52	97.37
				✓	84.52(−0.36)	89.11(−0.25)	77.38(−0.43)	82.17	61.15	76.22	82.53	64.92	97.31
✓	✓	✓	✓	✓	85.27(+0.39)	89.64(+0.28)	78.31(+0.5)	82.85	61.80	77.83	83.27	66.75	97.32

**Fig. 13.** Overall image of the simulated cloudy scene: RGB on the left, Simulated RGB in the middle, and SAR on the right.

**CFM.** The RGB-SAR-CFM model (fourth row), which integrates CFM, demonstrates improved accuracy for all foreground categories compared to the RGB-SAR-PWA model (third row). The RGB-SAR-SFM-CFM (sixth row), on the other hand, also proves the validity of CFM compared to the RGB-SAR-SFM-PWA (fifth row) model. However, when compared to the RGB-SAR-PWA (third row) model, the RGB-SAR-SFM-PWA model (fifth row) shows a slight decrease in  $\text{IoU}_{\text{water}}$  of 0.16%, with the improvements in other categories being less marked. This suggests that simple feature addition does not fully harness the complementary potential of the two modalities and introduces additional noise. In contrast, the CFM we propose is specifically designed to select and fuse features from both modalities effectively.

## 5.2. Ablation experiments between different stages

The backbone network consists of four stages that output features at different scales: stage-1, stage-2, stage-3, and stage-4. In this section, we choose a baseline model that utilizes PWA module for fusion in all stages. Furthermore, SFM and CFM modules are introduced at various stages for ablation experiments. The fusion results for each stage are presented in Table 7. Our experimental findings indicate a decrease in accuracy when the SFM and CFM modules are introduced at stage-1 and stage-4. This reduction in accuracy can be attributed to the insufficient semantic information in the feature maps of the lowest stage and the limited local information in the feature maps of the highest stage. The complex feature interactions and fusions that are unique to stage-1 or stage-4 tend to amplify noise from the two modalities, thereby resulting in adverse effects.

The outcomes of incorporating SFM and CFM across different stages indicate that water bodies, forests, and croplands exhibit superior performance at stage-2, with roads and cities performing optimally at stage-3. The local details and texture features in stage-2 benefit significantly from integration. For categories that require expansive receptive fields and rich semantic information, stage-3 offers the most effective fusion. However, at stage-4, the drop in accuracy is more significant, due to the reduced availability of location-specific information.

The network that integrated SFM and CFM modules across all stages achieved the highest overall accuracy. However, the  $\text{IoU}_{\text{road}}$  and  $\text{IoU}_{\text{other}}$  decreased compared to the network where only the stage-3 was enhanced. This suggests that features fused across different stages may negatively impact certain categories. For detection tasks requiring high accuracy for specific categories, combining fusion strategies across different stages can yield superior results.

## 5.3. Comparative experiments under cloudy conditions

To assess the model's effectiveness and superiority in capturing large-scale cloudy scenes, we generated RGB images under simulated cloudy conditions using GIMP. (The foggy filter was used to perform the simulation, with fog color set to white, turbulence set to 1.5, and opacity set to 100.) The simulated image is depicted in Fig. 13. Adhering to the methodology employed in PIE-RGB-SAR, we cropped the images and segmented the dataset. The experimental outcomes for diverse models are summarized in Table 8.

When compared to the original RGB image, the kappa, OA, and mIoU of the simulated S-RGB single-mode image decreased by 2.18%, 1.53%, and 3.72%, respectively. This suggests that foggy conditions can adversely affect detection outcomes. Nevertheless, in the simulated environment, our ASANet achieved the best performance, with the mIoU index 1.07% to 21.06% higher than the corresponding indices of other models. The advantage of ASANet is more obvious in the road and water categories.

## 6. Conclusion

In this study, we introduce ASANet, which asymmetrically fuses RGB and SAR images at the feature level for LCC. We design an SFM module to asymmetrically enhance the interaction between features of the two modalities, adaptively amplifying complementary features in distinct branches. Additionally, we introduce a CFM to perform feature fusion by selectively integrating features from both modalities in both channel and spatial dimensions. Furthermore, we introduce a multimodal LCC dataset that is designed for practical application

**Table 8**

Comparative experimental metrics (%) of different models in a cloudy scene. The best results are presented in bold, while the second-best results are underlined.

Model	Modality		Kappa	OA	mIoU	IoU	City	Road	Water	Forest	Farmland	Other
	RGB	SAR										
UPerNet	✓		81.95	87.31	73.18	79.68	50.49	68.03	80.38	63.99	96.51	
UPerNet		✓	74.16	83.82	66.73	75.59	33.74	70.42	75.98	49.60	95.74	
FuseNet	✓	✓	66.40	77.14	54.60	67.96	19.40	49.24	67.85	27.04	96.10	
SA-Gate	✓	✓	79.34	85.54	69.21	77.67	39.41	64.02	78.01	59.73	96.44	
AFNet	✓	✓	81.40	86.91	72.00	79.43	45.89	69.17	80.04	62.21	95.29	
CMFNet	✓	✓	81.78	87.22	72.77	80.4	48.15	73.11	80.15	58.17	96.63	
CMX	✓	✓	<u>83.52</u>	<u>88.43</u>	<u>75.13</u>	<u>81.51</u>	<u>53.06</u>	<u>72.97</u>	<u>81.89</u>	<u>64.20</u>	<b>97.13</b>	
FTransUNet	✓	✓	81.70	87.13	72.78	80.24	48.09	<u>73.61</u>	80.06	57.75	96.89	
ASANet (ours)	✓	✓	84.15	<b>88.85</b>	<b>76.20</b>	<b>81.76</b>	<b>55.34</b>	<b>74.46</b>	<b>82.48</b>	<b>66.09</b>	<b>97.07</b>	

scenarios. We conduct extensive experiments on this dataset as well as two other publicly available datasets, comparing our ASANet with existing multimodal fusion methods. The experimental results show that our ASANet achieves excellent performance in LCC under both real-world and simulated cloud and fog conditions. The effectiveness of our model is also evident in feature map visualizations, which highlight the model's ability to adaptively utilize information from both modalities, mitigate unimodal limitations, and achieve complementary information fusion. In our future work, we will focus on investigating the application of an asymmetric modal fusion architecture. This approach aims to align the modal information with the corresponding parameters of each branch, thereby optimizing the utilization of modal characteristics.

#### CRediT authorship contribution statement

**Pan Zhang:** Conceptualization, Methodology, Writing – review & editing. **Baochai Peng:** Conceptualization, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **Chaoran Lu:** Writing – review & editing. **Quanjin Huang:** Conceptualization. **Dongsheng Liu:** Funding acquisition, Resources.

#### 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 wish to express our gratitude for the financial support provided by the [Startup Foundation for Introducing Talent of Nanjing University of Information Science and Technology / Grant ID: 2024r102] for this research.

#### References

- Araya-López, R.A., Lopatin, J., Fassnacht, F.E., Hernández, H.J., 2018. Monitoring andean high altitude wetlands in central Chile with seasonal optical data: A comparison between worldview-2 and sentinel-2 imagery. *ISPRS J. Photogramm. Remote Sens.* 145, 213–224.
- Bermudez, J.D., Happ, P.N., Feitosa, R.Q., Oliveira, D.A., 2019. Synthesis of multispectral optical images from SAR/optical multitemporal data using conditional generative adversarial networks. *IEEE Geosci. Remote Sens. Lett.* 16 (8), 1220–1224.
- Cao, Z., Jiang, L., Yue, P., Gong, J., Hu, X., Liu, S., Tan, H., Liu, C., Shangguan, B., Yu, D., 2023. A large scale training sample database system for intelligent interpretation of remote sensing imagery. *Geo-Spatial Inf. Sci.* 1–20.
- Chen, X., Lin, K.Y., Wang, J., Wu, W., Qian, C., Li, H., Zeng, G., 2020. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation. In: European Conference on Computer Vision. Springer, pp. 561–577.
- Costa, H., Foody, G.M., Boyd, D.S., 2018. Supervised methods of image segmentation accuracy assessment in land cover mapping. *Remote Sens. Environ.* 205, 338–351.
- Gao, M., Xu, J., Yu, J., Dong, Q., 2023. Distilled heterogeneous feature alignment network for SAR image semantic segmentation. *IEEE Geosci. Remote Sens. Lett.*
- Guo, Z., Shengoku, H., Wu, G., Chen, Q., Yuan, W., Shi, X., Shao, X., Xu, Y., Shibasaki, R., 2018. Semantic segmentation for urban planning maps based on U-net. In: IGARSS 2018–2018 IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 6187–6190.
- Guo, Z., Wu, L., Huang, Y., Guo, Z., Zhao, J., Li, N., 2022. Water-body segmentation for SAR images: past, current, and future. *Remote Sens.* 14 (7), 1752.
- Hang, R., Li, Z., Ghamisi, P., Hong, D., Xia, G., Liu, Q., 2020. Classification of hyperspectral and LiDAR data using coupled CNNs. *IEEE Trans. Geosci. Remote Sens.* 58 (7), 4939–4950.
- Hazirbas, C., Ma, L., Domokos, C., Cremers, D., 2017. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In: Computer Vision-ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20–24, 2016, Revised Selected Papers, Part I 13. Springer, pp. 213–228.
- He, T., Sun, Y.J., Xu, J.D., Wang, X.J., Hu, C.R., 2014. Enhanced land use/cover classification using support vector machines and fuzzy k-means clustering algorithms. *J. Appl. Remote Sens.* 8 (1), 083636.
- He, W., Yokoya, N., 2018. Multi-temporal sentinel-1 and-2 data fusion for optical image simulation. *ISPRS Int. J. Geo-Inf.* 7 (10), 389.
- Hermessi, H., Mourali, O., Zagrouba, E., 2021. Multimodal medical image fusion review: Theoretical background and recent advances. *Signal Process.* 183, 108036.
- Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141.
- Hu, X., Yang, K., Fei, L., Wang, K., 2019. ACNet: Attention based network to exploit complementary features for rgbd semantic segmentation. In: 2019 IEEE International Conference on Image Processing. ICIP, IEEE, pp. 1440–1444.
- Kang, W., Xiang, Y., Wang, F., You, H., 2022. CFNet: A cross fusion network for joint land cover classification using optical and SAR images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 15, 1562–1574.
- Kazakeviciute-Januskeviciene, G., Janusonis, E., Bausys, R., Limba, T., Kiskis, M., 2020. Assessment of the segmentation of RGB remote sensing images: a subjective approach. *Remote Sens.* 12 (24), 4152.
- Li, J., Hong, D., Gao, L., Yao, J., Zheng, K., Zhang, B., Chanussot, J., 2022a. Deep learning in multimodal remote sensing data fusion: A comprehensive review. *Int. J. Appl. Earth Obs. Geoinf.* 112, 102926.
- Li, Z., Jiao, Q., Liu, L., Tang, H., Liu, T., 2014. Monitoring geologic hazards and vegetation recovery in the Wenchuan earthquake region using aerial photography. *ISPRS Int. J. Geo-Inf.* 3 (1), 368–390.
- Li, W., Sun, K., Li, W., Wei, J., Miao, S., Gao, S., Zhou, Q., 2023a. Aligning semantic distribution in fusing optical and SAR images for land use classification. *ISPRS J. Photogramm. Remote Sens.* 199, 272–288.
- Li, X., Zhang, G., Cui, H., Hou, S., Chen, Y., Li, Z., Li, H., Wang, H., 2023b. Progressive fusion learning: A multimodal joint segmentation framework for building extraction from optical and SAR images. *ISPRS J. Photogramm. Remote Sens.* 195, 178–191.
- Li, X., Zhang, G., Cui, H., Hou, S., Wang, S., Li, X., Chen, Y., Li, Z., Zhang, L., 2022b. MCANet: A joint semantic segmentation framework of optical and SAR images for land use classification. *Int. J. Appl. Earth Obs. Geoinf.* 106, 102638.
- Li, X., Zhao, H., Wu, D., Liu, Q., Tang, R., Li, L., Xu, Z., Lyu, X., 2024. SLMFNet: Enhancing land cover classification of remote sensing images through selective attentions and multi-level feature fusion. *PLoS One* 19 (5), e0301134.
- Li, R., Zheng, S., Duan, C., Wang, L., Zhang, C., 2022c. Land cover classification from remote sensing images based on multi-scale fully convolutional network. *Geo-Spatial Inf. Sci.* 25 (2), 278–294.
- Liu, X., He, C., Zhang, Q., Liao, M., 2019. Statistical convolutional neural network for land-cover classification from SAR images. *IEEE Geosci. Remote Sens. Lett.* 17 (9), 1548–1552.
- Li, X., Wu, Y., Liang, W., Cao, Y., Li, M., 2022. High resolution SAR image classification using global-local network structure based on vision transformer and CNN. *IEEE Geosci. Remote Sens. Lett.* 19, 1–5.
- Lu, c., 2021. Geological Hazards Detection and Deformation Analysis in Mountainous Areas of Guizhou Based on Space-borne InSAR Technology. Wuhan University.
- Ma, W., Karakuş, O., Rosin, P.L., 2022a. AMM-FuseNet: Attention-based multi-modal image fusion network for land cover mapping. *Remote Sens.* 14 (18), 4458.
- Ma, X., Wu, P., Wu, Y., Shen, H., 2017. A review on recent developments in fully polarimetric SAR image despeckling. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 11 (3), 743–758.

- Ma, X., Zhang, X., Pun, M.O., 2022b. A crossmodal multiscale fusion network for semantic segmentation of remote sensing data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 15, 3463–3474.
- Ma, X., Zhang, X., Pun, M.O., Liu, M., 2024. A multilevel multimodal fusion transformer for remote sensing semantic segmentation. *IEEE Trans. Geosci. Remote Sens.*
- Ma, A., Zheng, C., Wang, J., Zhong, Y., 2023. Domain adaptive land-cover classification via local consistency and global diversity. *IEEE Trans. Geosci. Remote Sens.*
- Orlíková, L., Horák, J., 2019. Land cover classification using sentinel-1 SAR data. In: 2019 International Conference on Military Technologies. ICMT, IEEE, pp. 1–5.
- Oscio, L.P., Wu, Q., de Lemos, E.L., Gonçalves, W.N., Ramos, A.P.M., Li, J., Junior, J.M., 2023. The segment anything model (sam) for remote sensing applications: From zero to one shot. *Int. J. Appl. Earth Obs. Geoinf.* 124, 103540.
- Pei, D., Li, Z., Ji, R., Sun, F., 2014. Efficient semantic image segmentation with multi-class ranking prior. *Comput. Vis. Image Underst.* 120, 81–90.
- Qingyun, F., Zhaokui, W., 2022. Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery. *Pattern Recognit.* 130, 108786.
- Ren, B., Liu, B., Hou, B., Wang, Z., Yang, C., Jiao, L., 2024. SwinTFNet: Dual-stream transformer with cross attention fusion for land cover classification. *IEEE Geosci. Remote Sens. Lett.*
- Ren, B., Ma, S., Hou, B., Hong, D., Chanussot, J., Wang, J., Jiao, L., 2022. A dual-stream high resolution network: Deep fusion of GF-2 and GF-3 data for land cover classification. *Int. J. Appl. Earth Obs. Geoinf.* 112, 102896.
- Rizzoli, G., Barbato, F., Zanuttigh, P., 2022. Multimodal semantic segmentation in autonomous driving: A review of current approaches and future perspectives. *Technologies* 10 (4), 90.
- Rodríguez-Galiano, V.F., Ghimire, B., Rogan, J., Chica-Olmo, M., Rigol-Sánchez, J.P., 2012. An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS J. Photogramm. Remote Sens.* 67, 93–104.
- Stone, J.V., 2013. Bayes' Rule: A Tutorial Introduction to Bayesian Analysis. Sebtel Press.
- Tu, Y., Chen, B., Zhang, T., Xu, B., 2020. Regional mapping of essential urban land use categories in China: A segmentation-based approach. *Remote Sens.* 12 (7), 1058.
- Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S., 2023. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16133–16142.
- Woo, S., Park, J., Lee, J.-Y., Kweon, I.S., 2018. Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 3–19.
- Wu, W., Guo, S., Shao, Z., Li, D., 2023. CroFuseNet: A semantic segmentation network for urban impervious surface extraction based on cross fusion of optical and SAR images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 16, 2573–2588.
- Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J., 2018. Unified perceptual parsing for scene understanding. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 418–434.
- Xu, J., Lu, K., Wang, H., 2021. Attention fusion network for multi-spectral semantic segmentation. *Pattern Recognit. Lett.* 146, 179–184.
- Xu, F., Shi, Y., Ebel, P., Yang, W., Zhu, X.X., 2023a. High-resolution cloud removal with multi-modal and multi-resolution data fusion: A new baseline and benchmark. *arXiv preprint arXiv:2301.03432*.
- Xu, F., Shi, Y., Yang, W., Zhu, X., 2023b. Multi-modal multi-task learning for semantic segmentation of land cover under cloudy conditions. In: IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 6274–6277.
- Yang, C., Rottensteiner, F., Heipke, C., 2018. Classification of land cover and land use based on convolutional neural networks. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* 4, 251–258.
- Yu, D., Ji, S., 2023. Long-range correlation supervision for land-cover classification from remote sensing images. *IEEE Trans. Geosci. Remote Sens.*
- Zhang, Z., Liu, B., Li, Y., 2023a. FURSformer: Semantic segmentation network for remote sensing images with fused heterogeneous features. *Electronics* 12 (14), 3113.
- Zhang, J., Liu, H., Yang, K., Hu, X., Liu, R., Stiefelhagen, R., 2023b. CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers. *IEEE Trans. Intell. Transp. Syst.*
- Zhang, R., Tang, X., You, S., Duan, K., Xiang, H., Luo, H., 2020. A novel feature-level fusion framework using optical and SAR remote sensing images for land use/land cover (LULC) classification in cloudy mountainous area. *Appl. Sci.* 10 (8), 2928.
- Zhao, S., Tu, K., Ye, S., Tang, H., Hu, Y., Xie, C., 2023. Land use and land cover classification meets deep learning: A review. *Sensors* 23 (21), 8966.
- Zhou, W., Persello, C., Li, M., Stein, A., 2023. Building use and mixed-use classification with a transformer-based network fusing satellite images and geospatial textual information. *Remote Sens. Environ.* 297, 113767.