remote sensing

MDPI

Article

MFFnet: Multimodal Feature Fusion Network for Synthetic Aperture Radar and Optical Image Land Cover Classification

Yangyang Wang [1,2], Wengang Zhang [2], Weidong Chen [1,*,†], Chang Chen [1] and Zhenyu Liang [2,†]

[1] School of Information Science and Technology, University of Science and Technology of China, Hefei 230037, China; wangyy@nudt.edu.cn (Y.W.); chench@ustc.edu.cn (C.C.)
[2] Electronic Countermeasure Institute, National University of Defense Technology, Hefei 230037, China; zhangwg@nudt.edu.cn (W.Z.); liangzy21@nudt.edu.cn (Z.L.)
* Correspondence: wdchen@ustc.edu.cn
† These authors contributed equally to this work.

Abstract: Optical and Synthetic Aperture Radar (SAR) imagery offers a wealth of complementary information on a given target, attributable to the distinct imaging modalities of each component image type. Thus, multimodal remote sensing data have been widely used to improve land cover classification. However, fully integrating optical and SAR image data is not straightforward due to the distinct distributions of their features. To this end, we propose a land cover classification network based on multimodal feature fusion, i.e., MFFnet. We adopt a dual-stream network to extract features from SAR and optical images, where a ResNet network is utilized to extract deep features from optical images and PidiNet is employed to extract edge features from SAR. Simultaneously, the iAFF feature fusion module is used to facilitate data interactions between multimodal data for both low- and high-level features. Additionally, to enhance global feature dependency, the ASPP module is employed to handle the interactions between high-level features. The processed high-level features extracted from the dual-stream encoder are fused with low-level features and inputted into the decoder to restore the dimensional feature maps, generating predicted images. Comprehensive evaluations demonstrate that MFFnet achieves excellent performance in both qualitative and quantitative assessments on the WHU-OPT-SAR dataset. Compared to the suboptimal results, our method improves the OA and Kappa metrics by 7.7% and 11.26% on the WHU-OPT-SAR dataset, respectively.

Keywords: synthetic aperture radar; land cover classification; multimodal feature fusion

check for updates

Citation: Wang, Y.; Zhang, W.; Chen, W.; Chen, C.; Liang, Z. MFFnet: Multimodal Feature Fusion Network for Synthetic Aperture Radar and Optical Image Land Cover Classification. *Remote Sens.* **2024**, *16*, 2459. https://doi.org/10.3390/rs16132459

Academic Editors: Liangxiu Han, Stefano Pignatti, Sang-Eun Park, Jiali Shang, Wenjiang Huang and Yanbo Huang

Received: 3 March 2024
Revised: 27 June 2024
Accepted: 1 July 2024
Published: 4 July 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Precise land cover classification data are vital for regional development, including natural resource management [1], agricultural planning [2], and change detection [3]. SAR has gained prominence as a principal data source for numerous remote sensing applications, owing to its remarkable penetration capabilities and capacity to measure backscattering irrespective of weather conditions [4–7]. However, it suffers from severe speckle noise during imaging, which cannot be easily removed and significantly distorts the edge information of images. Optical images, while capable of presenting clear edge information during imaging, are often subject to interference from clouds, fog, or precipitation during acquisition [8,9]. SAR and optical images complement each other in terms of edge information. Therefore, leveraging the complementary information from multisource images has become a major trend in improving the accuracy of remote sensing tasks, especially for land classification tasks [10]. Liu et al. [11] employed Sentinel-2A optical imagery and SAR to enhance the precision of land cover classification. However, limited by the feature extraction capability of the model, it cannot effectively delineate the boundary between the two types of land. Hong et al. [12] introduced a novel shared specific feature learning model aimed at leveraging multimodal data for land cover classification. This model incorporates diverse data sources such as hyperspectral and SAR imagery.

However, a notable obstacle emerges in the classification of land cover using optical and SAR imagery, stemming from their distinct imaging modalities. This disparity results in significant variations in data distribution, posing challenges in aligning optical and SAR images. Feature extraction using deep networks and the effective integration of multimodal features can effectively mitigate the problem of domain gaps. Previous studies [13,14] have categorized the integration techniques for optical and SAR data into three primary levels: pixel-level [15], feature-level [16], and decision-level [17]. In general, pixel-level fusion entails directly overlaying optical and SAR data without the need for feature extraction from the images. Dupas, C.A. [18] adopted the Brovey Transform and Intensity-Hue-Saturation (IHS) methodologies to integrate SAR and multispectral images for land cover classification. However, some research suggests that pixel-level fusion may not be suitable for optical and SAR data due to the fundamentally distinct imaging modalities. Feature-level fusion typically involves the incorporation of classification features that are manually crafted or extracted using models [19]. Among the various features, texture features are predominant and widely utilized. Texture features play a crucial role in SAR data analysis, especially in single-polarization SAR data, as microwaves exhibit sensitivity to geometric properties of urban land surfaces [20]. Given its theoretical feasibility and the relatively advanced state of the technology, feature-level fusion remains predominant in existing research. Common methodologies in feature-level fusion include Support Vector Machine [21], Random Forest [22], and deep learning approaches. Quan et al. [23] developed a dual-stream model utilizing feature-level fusion to enhance the effectiveness of land classification. However, the limitations of the PCA method may result in either important features being overlooked or the extraction of excessive redundant features of SAR, thereby reducing the accuracy of the classification results. Zhang et al. [24] separately extracted spectral and textural features from optical and SAR images, and augmented this with related information such as NDVI. Subsequently, they performed object-oriented multiscale segmentation on the fused feature images. The experimental results using an enhanced SVM model demonstrated high classification performance. However, the approach is reliant on manually extracted features, which limits its widespread application. Decision-level fusion entails the classification of land cover using optical and SAR data independently, followed by making decisions grounded on the classification outcomes derived from both sources. Decision methods such as majority voting, entropy weighting, and Dempster–Shafer theory [25–27] are commonly employed. Waske and Benediktsson [28] employed a decision-fusion technique grounded in Support Vector Machines (SVM) to classify both SAR and multispectral images. However, challenges in decision-level fusion include determining which source possesses more dependable classification results and devising strategies to merge the two classification results.

To enhance the alignment of SAR and optical image features and ensure their comprehensive and efficient integration, we propose using a multimodal feature fusion network. Research has shown that SAR imagery, by complementing textural detail information in optical images, aids in land cover classification tasks. Furthermore, the effective interaction between SAR and optical images compensates for their respective limitations, enabling more proficient identification and differentiation of complex terrain types. Specifically, we employ a dual-stream encoder, with one stream utilizing ResNet [29] to extract rich spatial and semantic information from optical images, while the other stream leverages a PidiNet [30] network to capture intricate texture and structural details from SAR images. To maximize the capacity of SAR images to compensate for texture details in optical images, both high- and low-order features are extracted from each. Feature fusion of multimodal data is performed using the iAFF [31] module to facilitate the identification of land boundaries. Additionally, to tackle the challenge of recognizing complex land cover types, we incorporate the Atrous Spatial Pyramid Pooling (ASPP) [32] module, which extracts multiscale contextual information from the fused high-level features. Subsequently, the fused higher- and lower-order features are merged through channel concatenation, serving as the input for the classification module.

Our main contributions are as follows:

- We propose a new multimodal feature fusion network, MFFnet, which consists of a ResNet network specializing in extracting depth features from optical images and another network, PidiNet, for capturing fine-grained texture features of synthetic aperture radar images.
- We employ an iterative attentional feature fusion (iAFF) module to align features from SAR and optical images, which enables deep feature interactions by focusing on features from different modalities in the space and channel.
- The effectiveness and sophistication of our network is verified through extensive comparisons with existing SAR land classification networks and multimodal land classification networks.

The remainder of this paper is organized as follows. We present the details of each module of the MFFnet land cover classification algorithm in Section 2. Section 3 contains extensive experimental results and experimental discussion. The discussion is provided in Section 4. In Section 5, we conclude our paper.

2. Methodology

Figure 1 illustrates the proposed MFFnet, employing a conventional encoder–decoder framework for land cover classification in multimodal data. Specifically, diverse structured dual-stream encoders are employed to extract distinctive features from multimodal data. Subsequently, the iAFF module is utilized to amalgamate low- and high-level features of the multimodal data, fostering profound interactions between them. To better capture contextual information from high-level features, the ASPP module is employed for their processing. Finally, the merged low- and high-level features are then integrated for input into the decoder via channel concatenation. Following dimensionality reduction processing, the ultimate prediction results are derived.

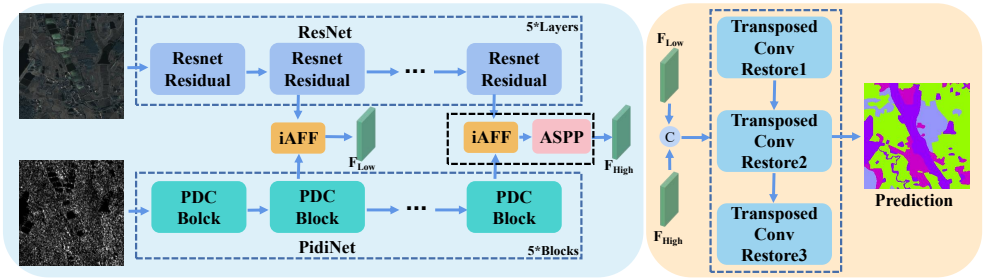

Figure 1. Overall framework of MFFnet, which consists of a two-stream encoder with the ResNet and PidiNet networks for feature extraction and feature fusion, and a decoder to recover features. Key components of the model also include the iAFF and ASPP modules. The *K** means that the number of such structures is *K*.

2.1. Encoder Feature Extraction

2.1.1. Overview

The distinctive imaging modalities of optical and SAR images result in significant variations in the information conveyed to each through their respective features. Optical imagery, acquired through visible light or infrared sensors, highlights vivid color features, enhancing target identification and classification. Nevertheless, susceptibility to shadows or clouds impacts its reliability. Conversely, SAR images provide limited information, predominantly comprising texture features, yet remain unaffected by adverse weather conditions. To mitigate mutual interference, a dual-stream encoder approach is employed for both, with ResNet and PidiNet utilized for feature extraction in the optical and SAR domains, respectively.

2.1.2. Optical Feature Extraction

Figure 1 illustrates the architecture of the ResNet framework, showcasing its application in constructing ultra-deep network structures through the incorporation of residual structures. Traditional convolutional neural networks rely on the sequential arrangement of convolutional and downsampling layers, encountering the gradient explosion issue with increased layer stacking. The introduction of residual structures effectively addresses these challenges while concurrently diminishing model parameterization. Numerous empirical studies attest to ResNet's efficacy in optical image feature extraction. The utilization of ResNet for extracting deep features is primarily attributed to its ability to mitigate gradient vanishing issues through residual learning, enabling the effective training of deeper networks for enhanced feature representation. Its architecture promotes feature reuse and optimization across layers, while consistently delivering top-tier performance across various computer vision tasks. ResNet's modular design and capacity to learn multi-scale features facilitate its seamless integration into sophisticated architectures, making it particularly advantageous for discerning intricate land classifications where a nuanced understanding of diverse feature hierarchies is vital.

The entire ResNet functions as an autonomous branch within the MFFnet encoder, dedicated to processing the input optical image $I \in R^{H\times W\times 4}$ and extracting corresponding features. Specifically, the initial convolutional layer transforms the input image into a feature map F $\in R^{H\times W\times C}$. This feature map undergoes processing through a sequence of residual structures, culminating in the derivation of the ultimate feature map. Each residual structure encompasses distinct convolutional layers and activation functions. With each passage through layers of residual blocks, the dimension of the feature map undergoes reduction, accompanied by a twofold increment in channel count.

The intricate configuration of the ResNet residual block is delineated in Figure 2. ResNet introduces two distinctive mappings: identity mapping and residual mapping. Identity mapping, denoted as shortcut branches, constitutes the lateral component of the structure, while residual mapping comprises the primary component. Within the deep ResNet architecture, three discrete convolutional layers are deployed for the principal branch in each residual block. The initial layer employs a one-dimensional kernel size to diminish the channel count. The second layer is tasked with reducing the dimensionality of the input features. The third layer, employing a kernel size of 1, augments the channel count to twice the input. In the shortcut branch, a convolutional layer is also present, reducing the input dimension and elevating the channel count. It is imperative to emphasize that the outputs from the trunk and shortcut branches must comprise identical shapes to ensure seamless final integration. The comprehensive residual structure can be articulated as follows:

$$Y_L = H(X_L) + F(X_L, W_L), X_{L+1} = R(Y_L) \quad (1)$$

where X_L and X_{L+1} represent the inputs and outputs of the residual structure, respectively. F denotes the backbone function of the residual structure, H stands for the side branches of the residual structure, and R symbolizes the final RELU function.

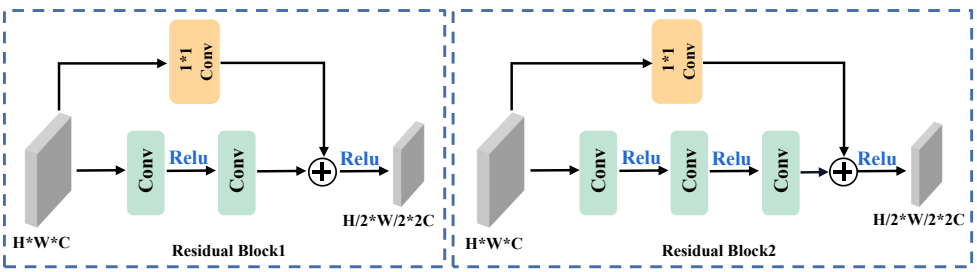

Figure 2. Structure of the ResNet block. On the left is the block for building a shallow network, and on the right is the block for building a deep network. The * in the picture represents ×.

The complete formulation of ResNet aligns with the aforementioned description. The integration of a dual-stream encoder would introduce a surplus of parameters, leading to a notable escalation in computational complexity, especially when opting for a sophisticated feature extractor. The efficacy of the residual structure lies in its ability to yield superior outcomes with reduced learning requirements, thereby expediting the convergence of the neural network. Consequently, ResNet is selected as the optical feature extractor.

2.1.3. SAR Feature Extraction

The SAR imaging mechanism inherently imparts a reduced amount of visual information to SAR images, making it complicated to process them. Traditional feature extraction from SAR images necessitates the utilization of a backbone network endowed with an augmented parameter count. However, this conventional approach exacts a substantial toll on hardware resources. In response, a straightforward, lightweight framework is introduced, named Pixel Difference Network (PidiNet), which incorporates a form of Pixel Difference Convolution (PDC) specifically tailored for efficient edge detection. Given the prevalence of texture features in SAR images and the consequential surge in parameter count facilitated by the dual-stream encoder, PidiNet aligns seamlessly with our objectives. Multiple Pixel Difference Convolutions (PDCs) are employed to instantiate numerous PDC blocks, constituting the backbone network for SAR branches, aligning with their optical counterparts to complete the entire encoder.

In the Pixel Difference Convolution (PDC) methodology, conventional detectors are integrated with CNNs, amalgamating the gradient information acquired from traditional detectors with the semantic features extracted by CNNs. This synergistic fusion culminates in the generation of edges characterized by both robustness and accuracy. Notably, conventional convolution operates on feature values, while PDC is concerned with pixel differences, as elucidated in Equation (2), below.

$$\begin{aligned} F = C(P,\theta) &= \sum_{i=1}^{k*k} w_i \bullet p_i \\ F = C(P,\theta) &= \sum_{p\in \mathrm{P}} w_i \bullet \left(p_i - p_i^{'}\right) \end{aligned} \tag{2}$$

where w denotes the convolution weight and $P = \{p_1, p_2 \ldots p_n\}$ denotes the pixel value.

The Local Binary Pattern (LBP) operator serves as a tool to articulate the local texture features inherent in an image. This method confers notable advantages, including rotational and gray-scale invariance. The PDC seamlessly integrates LBP with CNN [33–35], delineating three distinct methodologies for calculating PDC, as visually represented in Figure 3.

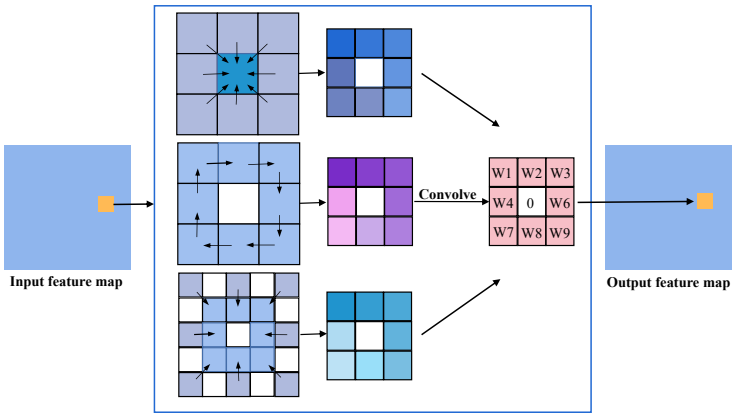

Figure 3. The evolution processes of pixel differential convolution are based on LBP. Shown from top to bottom, they are CPDC, APDC, and RPDC.

The computational process of APDC is delineated in Figure 4. The necessity of calculating a substantial number of pixel variances across diverse inputs leads to an augmented volume of arithmetic operations. To mitigate the computational burden associated with pixel pairs, the PDC layer can be transfigured into a conventional convolutional layer. The aim with this transformation is to alleviate the computational workload, thereby enhancing overall efficiency.

$$\begin{aligned} F &= w_1 * (p_1 - p_2) + w_2 * (p_2 - p_3) + w_3 * (p_3 - p_6) + \ldots \\ &= (w_1 - w_4) * p_1 + (w_2 - w_1) * p_2 + (w_3 - w_2) * p_3 + \ldots \\ &= \hat{w}_1 * p_1 + \hat{w}_2 * p_2 + \hat{w}_3 * p_3 + \ldots \\ &= \sum \hat{w}_i * p_i \end{aligned} \tag{3}$$

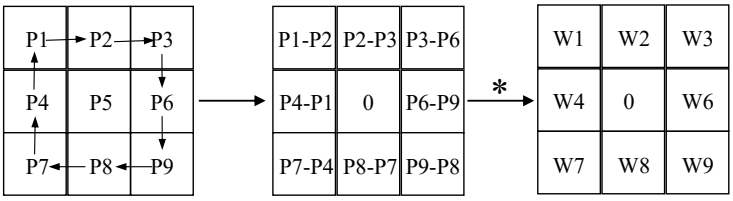

Figure 4. The evolution and derivation process of APDC. The $*$ in the picture represents $\times$.

The computational process of APDC is delineated in Equation (3). The PDC module can be transmuted into the equivalent standard convolution. This transformation modifies the disparity between pixels into disparities between weights, yielding updated weights. Such a modification leads to reduce computation, thereby simplifying the entire PDC. The rationale for converting CPDC and RPDC is explicated in Equations (4) and (5), where RPDC manifests as a convolutional layer with a kernel size of 5. The specific conversion process is illustrated in Figure 5.

$$\begin{aligned} F &= w_1 * (p_1 - p_5) + w_2 * (p_2 - p_5) + w_3 * (p_3 - p_5) + \ldots \\ &= w_1 * p_1 + w_2 * p_2 + w_3 * p_3 + \ldots w_9 * p_9 - \sum_{i=1.2\ldots9} w_i * p_5 \\ &= \hat{w}_1 * p_1 + \hat{w}_2 * p_2 + \hat{w}_3 * p_3 + \ldots \\ &= \sum \hat{w}_i * p_i \end{aligned} \tag{4}$$

$$\begin{aligned} F &= w_1 * (p_1 - p_7) + w_3 * (p_3 - p_8) + w_5 * (p_5 - p_9) + \ldots \\ &= \hat{w}_1 * p_1 + \hat{w}_2 * p_2 + \hat{w}_3 * p_3 + \ldots \\ &= \sum \hat{w}_i * p_i \end{aligned} \tag{5}$$

where p_i represents each pixel value in the feature map, w_i represents the weight of ordinary convolution, and $\hat{w}_i$ represents the weight of the PDC.

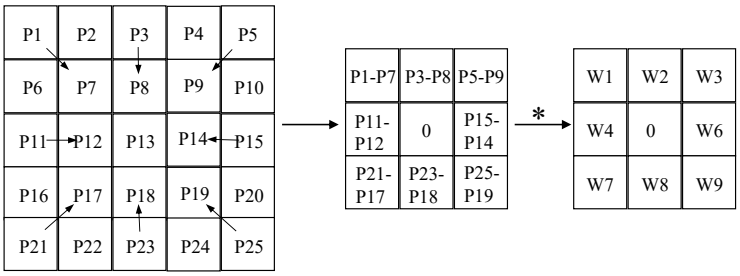

Figure 5. The evolution and derivation process of RPDC. The $*$ in the picture represents $\times$.

PDC was employed to construct the comprehensive SAR branch for feature extraction in SAR images. Each PDC block is depicted in Figure 6. In the architectural design, PDC is incorporated with the RELU function and standard convolution. The PDC layer encompasses multiple PDC blocks arranged in the ACR order to formulate an optimal PDC

layer tailored for the task at hand. The SAR branch comprises five PDC layers generating outputs aligning with those of the ResNet layer, thereby facilitating feature fusion.

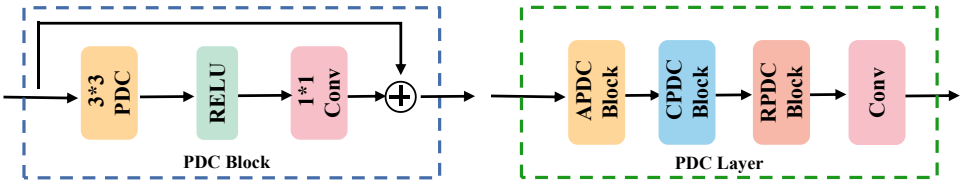

Figure 6. The structural diagram of blocks and layers comprising PDC. The * in the picture represents ×.

2.2. Feature Fusion Module

2.2.1. Overview

Feature fusion, in which features from diverse layers or branches are amalgamated, has become a pervasive element in network systems. Such integration is typically accomplished through elementary linear operations like summation or concatenation. Remote sensing images exhibit diminished resolution in comparison to natural images, possessing an abundance of information but lacking the requisite level of detail for optimal consideration. Concurrently, shallow features proximal to the input harbor richer position and color information, while high-level features in close proximity to the output encapsulate heightened semantic information. The amalgamation of these two feature types in the ultimate feature map amplifies the model's comprehensive understanding of the scene.

2.2.2. Cross-Modal Fusion

In the preceding formulation, we emphasized that optical images encompass visual attributes like colors and boundaries, whereas SAR images exclusively contain texture features. With mere feature superposition or channel splicing, dual-stream features are inadequately integrated. The iAFF method concentrates on features from distinct modalities across both spatial and channel dimensions and is therefore efficacious for scrutinizing intricate scenes in remote sensing.

Introduced in the iAFF method is a multiscale channel attention module (MS-CAM), resulting in a coherent and universal feature fusion scheme. Simultaneously, the entire feature fusion process is implemented through iterative attention to mitigate the bottleneck arising from the initial integration of feature maps. The comprehensive structural representation is illustrated in Figure 7.

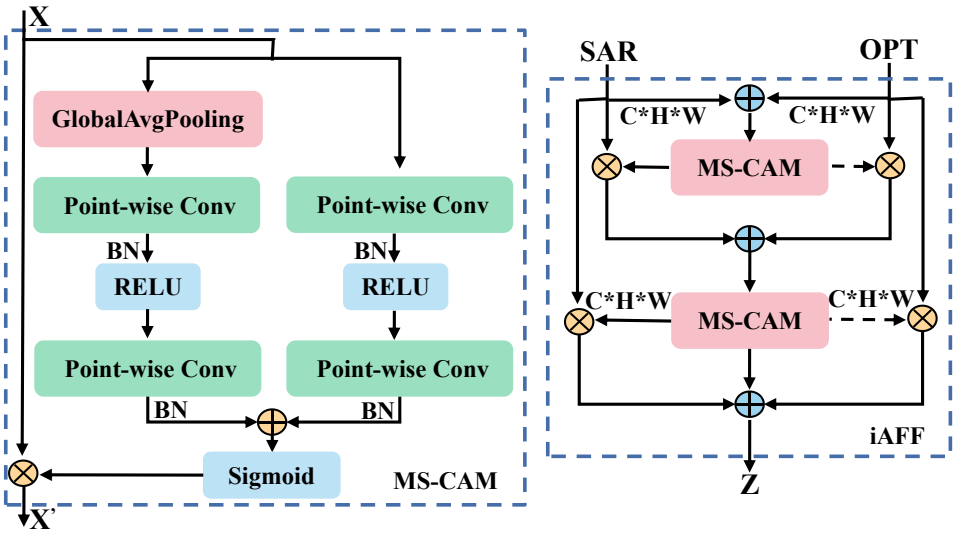

Figure 7. Complete structure of MS-CAM and iAFF. The * in the picture represents ×.

The MS-CAM extends the ParseNet paradigm of employing CNN for the spatial integration of local and global features via attentional fusion. The mitigation of the influence on channel scale is realized through point-by-point convolution, avoiding the necessity for convolution kernels of diverse sizes. MS-CAM does not exclusively govern the entirety of the fusion process, but rather focuses on local and global features.

The channel attention representation of the local features is as follows, implemented through pointwise convolution.

$$F_L = BN(PConv_2(\Re(BN(PConv_1(X))))) \tag{6}$$

where BN denotes batch normalization, $PConv_1$ represents pointwise convolution aimed at decreasing the number of input channels by a factor of r, $\Re$ denotes the RELU function, and the number of channels is finally recovered by $PConv_2$.

Regarding global channel attention, global average pooling of the input is first performed, followed by a pointwise convolution. The entire process and the final output are represented as follows:

$$F_G = G(BN(PConv_2(\Re(BN(PConv_1(X)))))) \tag{7}$$

where G represents the pooling operation, δ represents the Sigmoid function, and $\otimes$ and $\oplus$ denote the corresponding element multiplication and addition, respectively.

The final output $X^{'}$ of MS-CAM can be expressed as follows:

$$X^{'} = X \otimes \delta(F_L \oplus F_G) \tag{8}$$

Using iAFF, the unique characteristics of various modalities are comprehensively discerned through MS-CAM. For optical and SAR features, iAFF initially performs a simple summation of elements. To achieve a comprehensive perception of the two feature maps, the initial feature fusion undergoes iterative attentional fusion. Ultimately, the initially fused features and the iteratively fused features are amalgamated to yield the conclusive fused features. This entire process is denoted as iterative attentional feature fusion, and depicted as follows.

$$F_O + F_S = M(F_O \oplus F_S) \otimes F_O + (1 - M(F_O \oplus F_S)) \otimes F_S \tag{9}$$

The integration of features across the entire dual-stream network is accomplished by iAFF. This ensures the seamless complementarity between optical and SAR features, leading to the generation of more resilient fusion features.

2.2.3. Multiscale Fusion

Shallow features encompass a greater number of pixel points characterized by smaller sensory fields, capturing fine-grained global information regarding areas such as position and color. These features excel at capturing intricate details. In contrast, deeper features possess larger receptive fields and augmented overlapping regions, leading to the consolidation of image information and the generation of more potent semantic representations.

To enhance the amalgamation of features derived from the encoder, we devised a multiscale feature fusion strategy to ensure synergistic integration of low- and high-level features. Initially, we designated the cross-modal fusion features extracted from the second and fifth layers of the dual-flow encoder as F_{low} and F_{high}. Subsequently, the ASPP was introduced to aggregate contextual information from high features characterized by lower resolution and a reduced perceptiveness to details. Specifically, for high features, a 3×3 convolution with sampling factors of 6, 12, and 18 was utilized to extract features from distinct perceptual fields. Simultaneously, pooling and standard convolution, with a kernel size of 1, are executed to amalgamate the channel information, as illustrated in Figure 8. For the feature maps yielded in the five operations, identical sizes are maintained.

Following this, splicing and convolutional dimensionality reduction are implemented to derive the ultimate higher-order features, denoted as F_{high}^{ASPP}.

$$F_{high}^{ASPP} = ASPP(F_{high}) \tag{10}$$

Ultimately, the F_{low} and F_{high}^{ASPP} features are used sequentially as shared inputs for the decoder.

$$F_{low-high} = concat(F_{low}, F_{high}^{ASPP}) \tag{11}$$

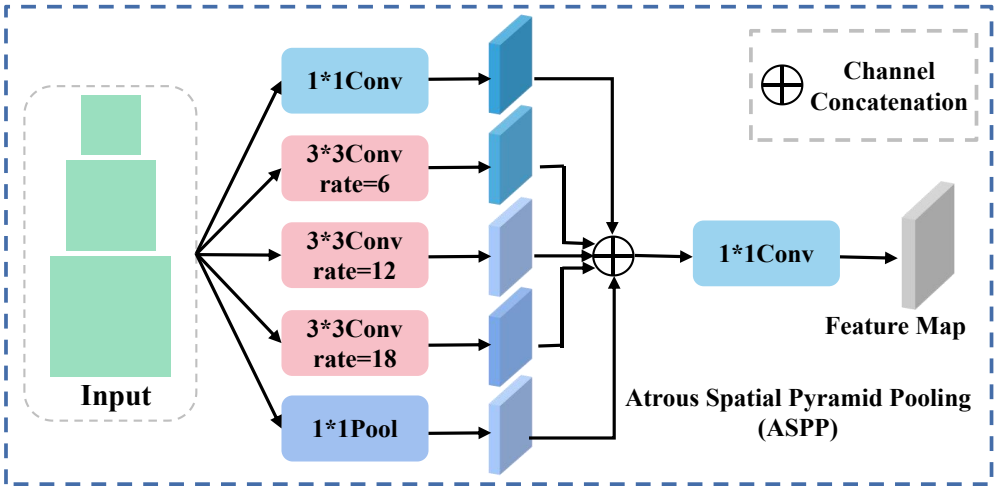

Figure 8. The structure of ASPP. The * in the picture represents ×.

2.3. Decoder Feature Restoration

The decoder serves the purpose of reinstating the feature map's dimensionality to align with the input image. Given the low resolution in remote sensing images, the employment of interpolating upsampling for the features can result in the loss of edge details across various categories. To tackle this challenge, we devised an inverse convolution-based decoding block, as delineated in Figure 9. Initially, for the input features, channels undergo convolution through a 3 × 3 layer to evenly distribute computational load throughout the inverse convolution process. In the second layer, the same number of channels is preserved for a seamless transition, ultimately culminating in the restoration of the feature map dimensions through upsampling via inverse convolution.

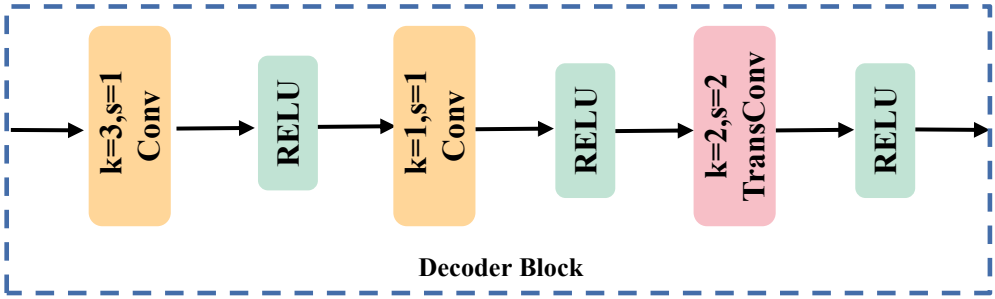

Figure 9. The flowchart of the decoder block.

For the input, the decoder engages in a quantitative prediction for each semantic class. Presuming that this encompasses n classes for classification, the decoder obtains semantic masks $M \in R^{h \times w \times n}$ for each meaningful entity, attributing n − 1 classes to

each semantic label. Finally, the resultant semantic segmentation prediction graph can be formally expressed as

$$M_{Final} = \arg\max(soft\max(M, c = -1), c = -1) \tag{12}$$

where $c = -1$ denotes the execution of softmax and argmax operations in the final dimension, i.e., the channel.

3. Experiments

3.1. Training Details and Dataset

The algorithm, developed on the PyTorch (version 2.3.1) framework, converges within 60 training iterations with a maximum learning rate of 0.001, scaled down by one-tenth every twenty rounds. Computational efficiency is enhanced by conducting experiments on an RTX3090Ti GPU.

Given the sparsity of multimodal datasets for land cover classification datasets, we selected a widely-used WHU-OPT-SAR(WHU) dataset [33] to showcase the performance of the proposed algorithm. SAR images from this dataset were also utilized for unimodal experiments, enabling comprehensive quantitative evaluation.

This dataset is the inaugural and most extensive land-use classification dataset, integrating optical and SAR images with comprehensive annotations. Covering approximately 50,000 km^2 in Hubei Province, China, the dataset includes SAR images measuring 5556 × 3704 pixels, as illustrated in Figure 10. The study area is characterized by a subtropical monsoon climate, encompassing altitudes that vary from a low of 50 m to a peak of 3000 m. Featuring a diverse range of topographies and vegetation, the dataset is segmented into seven primary classes: farmland, city, village, water, forests, roads, and others. Representative instances of these classes are depicted in Figure 11. The WHU dataset comprises a total of 29,400 images, which were divided in our experiments into train, val, and test sets in a ratio of 8:1:1. Specifically, the training set contained 23,520 images, while the validation and test sets each comprised 2940 images.

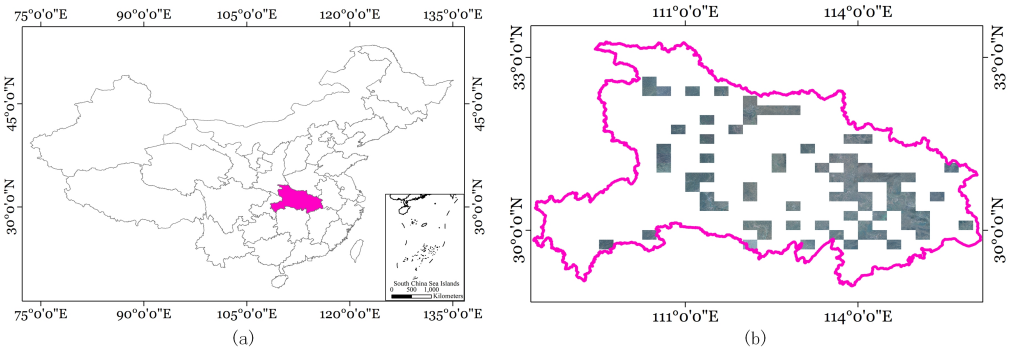

Figure 10. The details of the WHU dataset. (**a**) Geographic location of the dataset in China. (**b**) Distribution of images in the dataset.

3.2. Loss Function

To guarantee the convergence of the overall network training, cross-entropy is adopted as the loss function, measuring the deviation between the model's predicted probability distribution and the actual distribution. Mathematically, cross-entropy can be expressed as

$$Loss = -\sum_{i}^{K} y_i \cdot \log(\hat{y}_i) \tag{13}$$

where y_i represents the true probability distribution for class i, and $\hat{y}_i$ represents the predicted probability distribution for class i.

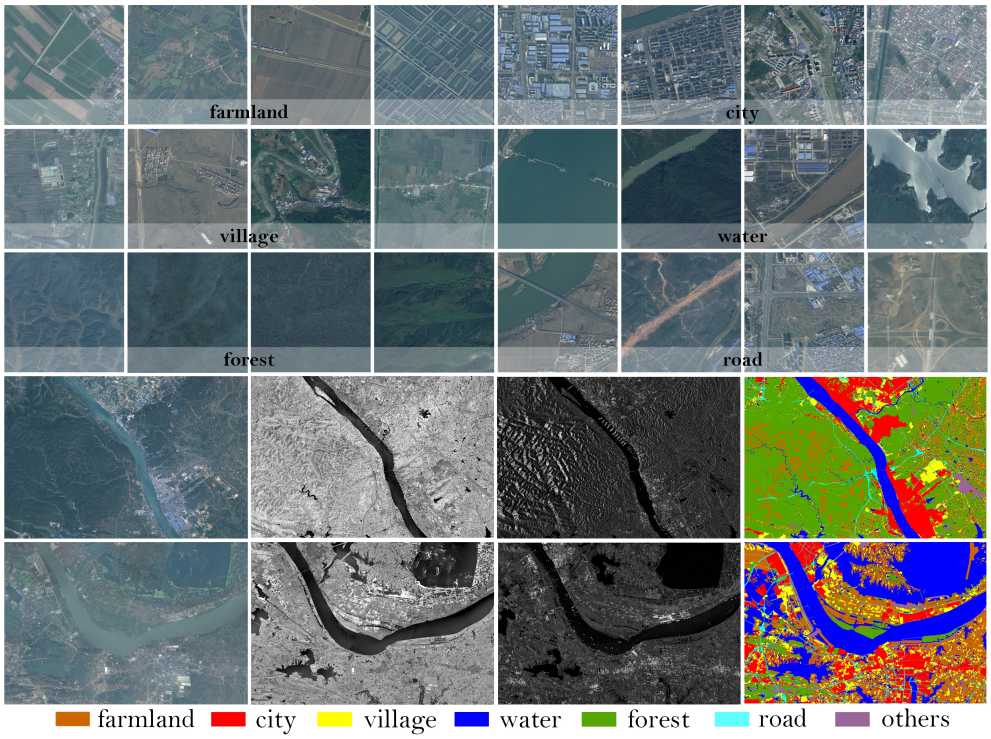

Figure 11. Examples of classes in the WHU dataset. The upper section showcases representative scenes from the dataset, while the lower section displays optical, near-infrared, and SAR images, along with their respective labels.

3.3. Quantitative Analysis

To evaluate the model's efficacy in land cover classification tasks, three widely used metrics are adopted: Overall Accuracy (OA), Kappa Coefficient, and Intersection over Union (IoU). OA measures the proportion of pixels correctly classified within the entire dataset.

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \tag{14}$$

where TP, TN, FP, and FN represent True Positive, True Negative, False Positive, and False Negative, respectively.

Kappa serves as a statistical measure to evaluate the agreement between predicted and actual classifications based on a consideration of the potential occurrence of agreement by chance.

$$Kappa = \frac{(P_0 - P_e)}{1 - P_e} \tag{15}$$

where P_0 denotes the percentage of samples correctly classified in the categories. P_e represents the expected agreement probability, which is also the probability of misclassification by a random classifier.

IoU measures the overlap between predictions and ground truth by dividing the intersection area by the combined area.

$$Iou = \frac{Pre \cap GT}{Pre \cup GT} \tag{16}$$

where Pre stands for prediction and GT signifies the true image.

The mIoU is computed as the mean of the IoU by comparing each Pre with the corresponding GT across all pixels in an image.

$$mIoU = \frac{1}{C} \sum_1^c IoU_C \tag{17}$$

where C denotes the total count of distinct categories.

3.4. Land Cover Classification Experiments

3.4.1. Baselines

Generally, to verify the positive impact of multimodal data on feature recognition, we first compare several unimodal semantic segmentation algorithms, with SAR images used as input. Secondly, we perform a quantitative comparison with some other typical multimodal algorithms to demonstrate the effectiveness of our algorithm.

Single-modal algorithms:

- SwinUnet [36]: A method utilizing a transformer architecture akin to Unet for medical image segmentation is proposed. This methodology integrates input patches into a transformer-based U-shaped architecture, enabling the effective capture of both local and global features.
- SegNet [37]: An integrated convolutional neural network framework, encompassing an encoder-decoder architecture and culminating in a pixel-wise classification layer, is introduced.
- TransUnet [38]: TransUnet presents a novel integration of transformer and U-Net functionalities. In this approach, the transformer component processes the feature maps encoded by a CNN as input sequences.
- DANet [39]: The presented methodology incorporates attention mechanisms into the enhanced Fully Convolutional Network (FCN), enabling the modeling of semantic dependencies across both spatial and channel dimensions.
- EncNet [40]: A context-encoding module is incorporated into the proposed approach to capture the semantic context. Feature maps are highlighted based on class dependencies, enhancing the model's ability to discern relevant contextual information.
- Deeplabv3 [41]: DeepLabv3 is harnessed to encode intricate contextual information, while a succinct decoder module is leveraged for precise boundary reconstruction.

Multimodal algorithms:

- ESANet [42]: An efficient semantic segmentation method designed for multimodal RGB images and depth images is proposed. The feature extraction and fusion of the two inputs occur in stages within the encoder. The decoder incorporates a Multi-Scale Supervision strategy, employing the jump-junction fusion method. Notably, all upsampling is conducted using a learned method rather than bilinear upsampling.
- RedNet [43]: This method employs classical two-stream encoder and decoder architecture, utilizing residual blocks as fundamental building components for both the encoder and decoder. The algorithm integrates encoder outputs and decoder outputs through hopping connections, incorporating an Agent Block operation before each connection layer to decrease the encoder channel size, thereby facilitating fusion.
- TFNet [44]: TFNet is a two-stream fusion algorithm that merges data from distinct modality domains in reconstructing images. In this method, all pooling operations are replaced with a convolution operation. Additionally, transposed convolution is employed for upsampling, and to maintain a symmetric structure in the encoder, feature mapping is conducted for layers in pairs.
- MCANet [45]: The approach is rooted in Deeplabv3, presenting a multimodal segmentation algorithm utilizing optical and SAR images as input. A cross-modal attention module is introduced to enhance and fuse features from both optical and SAR modalities.

3.4.2. Comparison Experiments

To evaluate the superiority of our multimodal algorithm over other unimodal approaches and to emphasize the positivity of multimodal data, we examined seven scenarios for visualization and quantitative comparisons, as illustrated in Figure 12. Highlighted in the figure are deficiencies in certain categories of unimodal algorithms, such as poor edge continuity and significant pixel confusion, leading to numerous false predictions in water

and other unidentified regions. The introduction of optical images enhances algorithms by leveraging color and edge features, resulting in the improved identification of village and water regions. Our approach demonstrates a reduction in misclassified regions, leading to an enhancement in detection accuracy. Table 1 shows that the three main indicators—OA, Kappa, and mIoU—experience increases of at least 7%, with substantial improvements for the two categories of water and village. Overall, our method demonstrates strong accuracy and robustness across all categories, surpassing other unimodal algorithms.

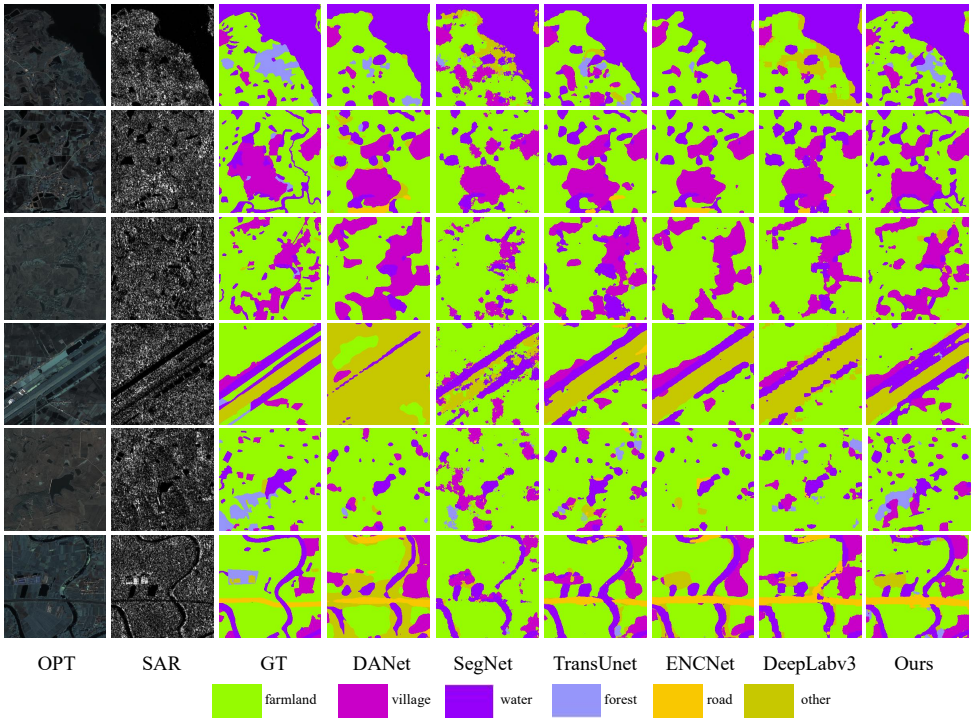

Figure 12. Land cover classification results with other single-modal algorithms on the WHU dataset. Optical, SAR, Ground Truth, and DANet, SegNet, TransUnet, ENCNet, DeepLabv3, and our proposed method are arranged sequentially.

Table 1. Evaluation metrics for each single-modal on the WHU dataset are presented. **Bold** and *italic* formatting is used to denote the best and second best results, respectively.

Method	OA	Kappa	mIoU	IoU						
				Farmland	City	Village	Water	Forest	Road	Others
SwinUnet	70.53	57.80	34.04	56.22	30.84	32.87	32.64	77.27	6.04	2.42
SegNet	73.21	60.51	36.00	59.35	29.55	36.03	27.76	77.88	15.48	5.93
TransUnet	73.40	61.87	38.09	61.22	*32.49*	36.15	*33.01*	77.82	*15.63*	9.88
DANet	69.18	56.82	34.31	61.09	26.41	34.62	23.71	74.61	12.82	6.92
EncNet	*73.67*	61.73	35.34	**63.72**	25.33	32.88	32.46	*78.70*	10.67	3.64
Deeplabv3	73.59	*62.33*	*38.44*	*62.46*	29.01	*38.76*	32.67	77.96	15.43	*12.82*
MFFnet (our network)	**81.29**	**73.59**	**50.64**	61.23	**34.34**	**49.04**	**73.26**	**82.82**	**24.96**	**28.84**

To comprehensively assess the efficacy of our proposed algorithm, we undertook a comparative study with several multimodal approaches. As shown in Figure 13, the forest and farmland masks generated by our method demonstrate superior edge coherence and effectively represent the challenging unknown category (others).The observed superiority of our approach can be attributed to its capacity to effectively utilize crucial edge details

in distinguishing forests and farmlands, irrespective of the visual properties of the optical surface. By integrating this information at a superior level, our methodology achieves more refined classification results. Quantitative comparisons with the multimodal algorithm are summarized in Table 2, indicating significant improvements in various metrics. OA, Kappa, and mIoU show enhancements of 1.72%, 2.26%, and 2.23%, respectively. In particular, our algorithm excels in terms of IoU for specific categories, including farmland (1.5%) and the unknown category (others) (5.32%). In addition, we also measured the model computation and parameters of the comparison experiments, as shown in Table 3, below.

Table 2. Evaluation metrics for each multimodal on the WHU dataset are presented. **Bold** and *italic* formatting is used to denote the best and second best results, respectively.

Method	OA	Kappa	mIoU	IoU						
				Farmland	City	Village	Water	Forest	Road	Others
MCANet	76.64	67.27	45.07	55.15	28.34	47.49	67.86	80.19	**26.35**	10.12
ESANet	79.42	71.11	46.74	*59.90*	23.83	41.23	*72.98*	82.48	23.24	*23.52*
RedNet	79.45	71.14	*48.41*	58.16	*34.55*	**50.83**	71.04	*82.68*	24.41	17.21
TFNet	*79.57*	*71.33*	47.93	59.73	**38.54**	*50.01*	70.71	81.38	18.20	16.96
MFFnet (our network)	**81.29**	**73.59**	**50.64**	**61.23**	34.34	49.04	**73.26**	**82.82**	*24.96*	**28.84**

Table 3. Results of model computation and parameter quantities for each multimodal on the WHU dataset.

Model	MCANet	ESANet	RedNet	TFNet	MFFnet (Our Network)
GFLOPs	88.69	11.52	23.30	30.90	22.95
Params/M	72.03	54.41	81.94	23.65	21.27

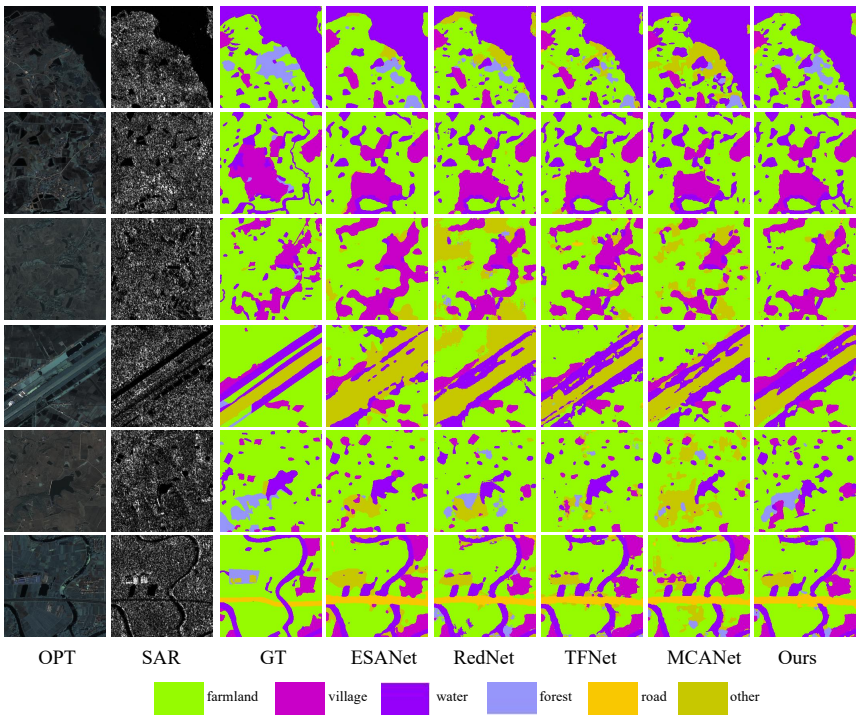

Figure 13. Land cover classification results for the use of other multimodal algorithms on the WHU dataset. Optical, SAR, Ground Truth, and the results of ESANet, RedNet, TFNet, MCANet, and our proposed method are arranged sequentially.

3.4.3. Ablation Experiment

To substantiate the function of the integral parts within MFFnet, we performed some ablation experiments using the WHU dataset. Experiments were performed for each combination of key components (i.e., PidiNet SAR image encoder, iAFF feature fusion module, and ASPP enhancement module). Table 4 contains detailed results of the entire ablation experiments, demonstrating that each component in the MFFnet framework contributes positively to the final results.

The distinctive edge extraction mechanism of PidiNet significantly enhances the metrics of land cover classification. The ASPP module effectively mitigates information loss arising from the downscaling of high-order features and the channel alignment of low-order features. Moreover, simple linear join operations for integrating optical and SAR features result in challenges for peak performance. The integration of the iAFF module demonstrates the imperative nature of cross-modal feature fusion, ensuring the optimal alignment of metrics with the pinnacle performance of the entire network.

Table 4. Ablation experiment of the MFFNet network: ✓ for use, × for no use. **Bold** indicates the best results.

ResNet	PidiNet	ASPP	iAFF	OA	Kappa	mIoU
✓	×	×	×	77.52	68.72	46.29
✓	✓	×	×	80.46	72.47	47.93
✓	✓	✓	×	80.50	72.58	49.26
✓	✓	✓	✓	**81.29**	**73.59**	**50.64**

4. Discussion

In the pursuit of advancing land classification in remote sensing, the integration of multimodal data, especially the synergy of SAR and optical imagery, unlocks the potential for unprecedented accuracy and detail capture. This multidimensional approach, while promising elevated classification accuracies, simultaneously introduces complex challenges rooted in the fundamental disparities between the two modalities. SAR's inherent speckle noise, a persistent obstacle to precise feature extraction, often defeats conventional denoising methods like the Lee filter, underlining the necessity for innovative solutions.

Our study rises to this challenge by harnessing the transformative power of deep learning, which has revolutionized feature extraction across various domains. Deep neural networks, with their hierarchical architecture, excel in distilling high-level, semantically meaningful representations from raw data, thereby conquering the intricacies posed by SARs' noisy texture. By adopting a deep learning framework tailored for speckle noise suppression and feature extraction, we aim to unlock SAR's full potential as a complementary source to optical data, enriching the informational depth for land classification tasks.

However, the true prowess of multimodal data lies in their synergistic fusion, a task fraught with subtleties. The iAFF emerges as our chosen strategy for harmoniously integrating these disparate yet complementary streams of information. Yet, a pivotal gap in the current research is the scarcity of rigorous assessments delineating the actual extent of feature integration between SAR and optical modalities post-fusion.

Our approach brings advances in overcoming the limitations of previous SAR processing techniques by providing a subtle understanding of multimodal feature interactions. We delve into how our deep-learning-driven noise reduction and iAFF-based fusion strategy contribute uniquely to the corpus of knowledge, pinpointing where our results deviate from or align with previous studies.

5. Conclusions

In summary, our work propose MFFnet, a novel dual-stream network designed for enhanced land cover classification using multimodal remote sensing data. By integrating PidiNet for SAR feature extraction, we optimize texture and semantic details, while the

innovative iAFF module facilitates the comprehensive fusion of multimodal features across spatial and channel dimensions. Empirical validations on the WHU dataset affirm MFFnet's superiority, marking a significant step forward in multimodal classification capabilities. Compared to the unimodal model, our method improves on OA and kappa, respectively, by 7.7% and 11.26% on the WHU-OPT-SAR dataset, respectively. Compared to the multimodal model, our method improves the OA and Kappa metrics by 1.72% and 2.26% on the WHU-OPT-SAR dataset, respectively. Notably, we recognize the limitations of SAR images with severe speckle noise in the absence of clear optical edge information, a factor that may affect classification accuracy. Hence, future research avenues entail the development of advanced feature extraction methodologies tailored to SAR's unique noise challenges, further pushing the boundaries of multimodal remote sensing analysis.

Author Contributions: Data curation, Y.W.; Writing—original draft, W.Z. and Y.W.; Writing—review and editing, Y.W., W.C. and Z.L.; Project administration, C.C. All authors have reviewed and approved the final version of the manuscript for publication.

Funding: This work is supported by the Scientific Research Project of the National University of Defense Technology, under Grant 22-ZZCX-07, and Hefei Comprehensive National Science Center.

Data Availability Statement: The WHU-OPT-SAR dataset is openly available on the website https://github.com/AmberHen/WHU-OPT-SAR-dataset.git (accessed on 1 January 2024).

Acknowledgments: The authors extend many thanks to researchers for providing the WHU-OPT-SAR dataset free of charge.

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

References

1. Letsoin, S.M.A.; Herak, D.; Purwestri, R.C. Evaluation Land Use Cover Changes Over 29 Years in Papua Province of Indonesia Using Remote Sensing Data. In *IOP Conference Series: Earth and Environmental Science*; IOP Publishing: Bristol, UK, 2022.
2. Dahhani, S.; Raji, M.; Hakdaoui, M.; Lhissou, R. Land cover mapping using sentinel-1 time-series data and machine-learning classifiers in agricultural sub-saharan landscape. *Remote Sens.* **2022**, *15* , 65. [CrossRef]
3. Kaul, H.A.; Sopan, I. Land use land cover classification and change detection using high resolution temporal satellite data. *J. Environ.* **2012**, *1*, 146–152.
4. Xu, F.; Shi, Y.; Ebel, P.; Yu, L.; Xia, G.S.; Yang, W.; Zhu, X.X. GLF-CR: SAR-enhanced cloud removal with global–local fusion. *ISPRS J. Photogramm. Remote Sens.* **2022**, *192*, 268–278. [CrossRef]
5. Xu, X.; Zhang, X.; Zhang, T. Lite-yolov5: A lightweight deep learning detector for on-board ship detection in large-scene sentinel-1 sar images. *Remote Sens.* **2022**, *14*, 1018. [CrossRef]
6. Zhang, T.; Zhang, X.; Ke, X.; Liu, C.; Xu, X.; Zhan, X.; Wang, C.; Ahmad, I.; Zhou, Y.; Pan, D.; et al. HOG-ShipCLSNet: A novel deep learning network with hog feature fusion for SAR ship classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5210322. [CrossRef]
7. Xu, X.; Zhang, X.; Shao, Z.; Shi, J.; Wei, S.; Zhang, T.; Zeng, T. A group-wise feature enhancement-and-fusion network with dual-polarization feature enrichment for SAR ship detection. *Remote Sens.* **2022**, *14*, 5276. [CrossRef]
8. Kang, W.; Xiang, Y.; Wang, F.; You, H. CFNet: A cross fusion network for joint land cover classification using optical and SAR images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 1562–1574. [CrossRef]
9. Li, X.; Ling, F.; Cai, X.; Ge, Y.; Li, X.; Yin, Z.; Shang, C.; Jia, X.; Du, Y. Mapping water bodies under cloud cover using remotely sensed optical images and a spatiotemporal dependence model. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *103*, 102470. [CrossRef]
10. Ye, Y.; Zhang, J.; Zhou, L.; Li, J.; Ren, X.; Fan, J. Optical and SAR image fusion based on complementary feature decomposition and visual saliency features. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5205315. [CrossRef]
11. Liu, S.; Qi, Z.; Li, X.; Yeh, A.G. Integration of convolutional neural networks and object-based post-classification refinement for land use and land cover mapping with optical and SAR data. *Remote Sens.* **2019**, *11*, 690. [CrossRef]
12. Hong, D.; Hu, J.; Yao, J.; Chanussot, J.; Zhu, X.X. Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model. *ISPRS J. Photogramm. Remote Sens.* **2021**, *178*, 68–80. [CrossRef]
13. Ghassemian, H. A review of remote sensing image fusion methods. *Inf. Fusion* **2016**, *32*, 75–89. [CrossRef]
14. Waske, B.; van der Linden, S. Classifying multilevel imagery from SAR and optical sensors by decision fusion. *IEEE Trans. Geosci. Remote Sens.* **2008**, *46*, 1457–1466. [CrossRef]
15. Kulkarni, S.C.; Rege, P.P. Pixel level fusion techniques for SAR and optical images: A review. *Inf. Fusion* **2020**, *59*, 13–29. [CrossRef]
16. Nirmala, D.E.; Vaidehi, V. Comparison of Pixel-level and feature level image fusion methods. In Proceedings of the 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 11–13 March 2015.

17. Xiao, G.; Bavirisetti, D.P.; Liu, G.; Zhang, X.; Xiao, G.; Bavirisetti, D.P.; Liu, G.; Zhang, X. Decision-level image fusion. In *Image Fusion*; Springer: Singapore, 2020.
18. Dupas, C.A. SAR And LANDSAT TM image fusion for land cover classification inthe brazilian atlantic forest domain. *Remote Sens.* **2000**, *33*, 96–103.
19. Zhang, Y.; Zhang, H.; Lin, H. Improving the impervious surface estimation with combined use of optical and SAR remote sensing images. *Remote Sens. Environ.* **2014**, *141*, 155–167. [CrossRef]
20. Masjedi, A.; Zoej, M.J.; Maghsoudi, Y. Classification of polarimetric SAR images based on modeling contextual information and using texture features. *IEEE Trans. Geosci. Remote Sens.* **2015**, *54*, 932–943. [CrossRef]
21. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
22. Rokach, L. Ensemble-based classifiers. *Artif. Intell. Rev.* **2010**, *33*, 1–39. [CrossRef]
23. Quan, Y.; Zhang, R.; Li, J.; Ji, S.; Guo, H.; Yu, A. Learning SAR-Optical Cross Modal Features for Land Cover Classification. *Remote Sens.* **2024**, *16*, 431. [CrossRef]
24. Zhang, R.; Tang, X.; You, S.; Duan, K.; Xiang, H.; Luo, H. 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.* **2020**, *10*, 2928. [CrossRef]
25. Clinton, N.; Yu, L.; Gong, P. Geographic stacking: Decision fusion to increase global land cover map accuracy. *ISPRS J. Photogramm. Remote Sens.* **2015**, *103*, 57–65. [CrossRef]
26. Zhu, Y.; Tian, D.; Yan, F. Effectiveness of entropy weight method in decision-making. *Math. Probl. Eng.* **2020**, *2020*, 3564835. [CrossRef]
27. Messner, M.; Polborn, M.K. Voting on majority rules. *Rev. Econ. Stud.* **2004**, *71*, 115–132. [CrossRef]
28. Waske, B.; Benediktsson, J.A. Fusion of support vector machinesfor classification of multisensor data. *IEEE Trans. Geosci. Remote Sens.* **2007**, *45*, 3858–3866. [CrossRef]
29. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
30. Su, Z.; Liu, W.; Yu, Z.; Hu, D.; Liao, Q.; Tian, Q.; Pietikäinen, M.; Liu, L. Pixel difference networks for efficient edge detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021.
31. Dai, Y.; Gieseke, F.; Oehmcke, S.; Wu, Y.; Barnard, K. Attentional feature fusion. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2021.
32. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef]
33. Liu, L.; Fieguth, P.; Kuang, G.; Zha, H. Sorted random projections for robust texture classification. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011.
34. Liu, L.; Zhao, L.; Long, Y.; Kuang, G.; Fieguth, P. Extended local binary patterns for texture classification. *Image Vis. Comput.* **2012**, *30*, 86–99. [CrossRef]
35. Su, Z.; Pietikäinen, M.; Liu, L. Bird: Learning binary and illumination robust descriptor for face recognition. In Proceedings of the 30th British Machine Visison Conference: BMVC, Cardiff, UK, 9–12 September 2019.
36. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2022.
37. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
38. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
39. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
40. Zhang, H.; Dana, K.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; Agrawal, A. Context encoding for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
41. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
42. Seichter, D.; Köhler, M.; Lewowski, B.; Wengefeld, T.; Gross, H.M. Efficient rgb-d semantic segmentation for indoor scene analysis. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021.
43. Jiang, J.; Zheng, L.; Luo, F.; Zhang, Z. Rednet: Residual encoder-decoder network for indoor rgb-d semantic segmentation. *arXiv* **2018**, arXiv:1806.01054.
44. Liu, X.; Liu, Q.; Wang, Y. Remote sensing image fusion based on two-stream fusion network. *Inf. Fusion* **2020**, *55*, 1–55. [CrossRef]
45. Li, X.; Zhang, G.; Cui, H.; Hou, S.; Wang, S.; Li, X.; Chen, Y.; Li, Z.; Zhang, L. MCANet: A joint semantic segmentation framework of optical and SAR images for land use classification. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *106*, 102638. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.