remote sensing

MDPI

Article

Learning SAR-Optical Cross Modal Features for Land Cover Classification

Yujun Quan [1], Rongrong Zhang [2,*,†], Jian Li [2,†], Song Ji [1], Hengliang Guo [2] and Anzhu Yu [1]

[1] School of Surveying and Mapping, PLA Strategic Support Force Information Engineering University, Zhengzhou 450001, China; qyj5312020@126.com (Y.Q.); jisong_chxy@163.com (S.J.); anzhu_yu@126.com (A.Y.)
[2] School of Geoscience and Technology, Zhengzhou University, Zhengzhou 450001, China; jianli@zzu.edu.cn (J.L.); guohengliang@zzu.edu.cn (H.G.)
* Correspondence: krrzhh@163.com
† These authors contributed equally to this work.

Abstract: Synthetic aperture radar (SAR) and optical images provide highly complementary ground information. The fusion of SAR and optical data can significantly enhance semantic segmentation inference results. However, the fusion methods for multimodal data remains a challenge for current research due to significant disparities in imaging mechanisms from diverse sources. Our goal was to bridge the significant gaps between optical and SAR images by developing a dual-input model that utilizes image-level fusion. To improve most existing state-of-the-art image fusion methods, which often assign equal weights to multiple modalities, we employed the principal component analysis (PCA) transform approach. Subsequently, we performed feature-level fusion on shallow feature maps, which retain rich geometric information. We also incorporated a channel attention module to highlight channels rich in features and suppress irrelevant information. This step is crucial due to the substantial similarity between SAR and optical images in shallow layers such as geometric features. In summary, we propose a generic multimodal fusion strategy that can be attached to most encoding–decoding structures for feature classification tasks, designed with two inputs. One input is the optical image, and the other is the three-band fusion data obtained by combining the PCA component of the optical image with the SAR. Our feature-level fusion method effectively integrates multimodal data. The efficiency of our approach was validated using various public datasets, and the results showed significant improvements when applied to several land cover classification models.

Keywords: multimodal fusion; land cover classification; deep learning; semantic segmentation

Citation: 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. https://doi.org/10.3390/rs16020431

Academic Editor: Timo Balz

Received: 20 December 2023
Revised: 18 January 2024
Accepted: 20 January 2024
Published: 22 January 2024

1. Introduction

The increasing trend of organizations disclosing their policies on satellite-based acquisition of multimodal data has significantly increased the availability of such data. The highly complementary nature of optical images and SAR data renders them invaluable for various applications, including land cover analysis, building damage assessment, earth resource surveys, crop identification, and more [1,2].

Optical remote sensing images, captured in various bands, offer distinct feature information representations. The imaging mechanism of optical images resonates with human visual habits. This alignment makes optical remote sensing images a cornerstone in most current deep learning-based semantic segmentation research. However, these images face limitations due to variable weather, seasons, light, and geographic factors. Moreover, clouds and shadows can obscure feature information. In contrast, synthetic aperture radar (SAR) photographs provide a rich array of landscape information from diverse material and physical viewpoints. SAR, an active sensor, detects backscatter information and is particularly sensitive to geometric features, such as surface roughness, temperature, and complex dielectric constants [3]. Although SAR images can penetrate obstacles and are impervious to water, they present challenges such as shortening, shadows, and speckle

noise, making their imaging characteristics less intuitive for human visual interpretation. The fusion of data from various imaging sources offers a solution to the difficulties posed by single-modal data, especially in land cover classification (LCC) task [4]. These multimodal data offer different perspectives on the same phenomena [1].

Fusing multimodal information for LCC tasks presents significant challenges due to the considerable differences in images captured by separate sensors; pre-processing different modal data and determining effective fusion methods is very difficult [5]. For example, to effectively use fused multimodal data to interpret remote sensing images, it is necessary to perform speckle reduction and register multimodal data because the backscattered signal undergoes coherent processing and SAR images are damaged by multiplicative speckle noise [6]. The registration of multimodal data is a complex task because of the significant geometric distortions between the data. Several studies have been proposed to accomplish the registration task of multimodal data, and these methods can be broadly categorized into area-based, feature-based, and learning-based pipelines [7]. For example, among feature-based pipelines, one study proposed the Harris-PIFD image registration framework with multi-scale features [8]. Moreover, one area-based method has been proposed to use two phases, coarse and fine registration, as well as first- and second-order gradients in the fine registration phase for geometric deformation problems [9]. Learning-based methods provide more possibilities for multimodal data registration, such as integrating deep learning methods with traditional methods to form new registration processes. Additionally, learning-based pipelines convert one modality into another, transforming complex multimodal data registration problems into simpler same source data registration problems. These approaches can also directly train the transformation parameters between regression multimodal data [7]. Overall, some targeted research has focused on improving multimodal data registration. However, after completing this task, it is necessary to determine how to utilize multimodal data with significant differences to perform remote sensing image interpretation more effectively.

Multimodal data fusion methods can be categorized into pixel-level-based, feature-level-based, and image-level-based approaches [2]. Pixel-level fusion uses fusion rules such as intensity–hue–saturation, Gram–Schmidt orthogonalization, bravery transform, high pass filtering, principal component analysis, wavelet transform, and generalized Laplace pyramid to fuse the pixel values of multimodal data [10,11]. Fused images have richer content but are more computationally demanding, and they require strict registration. Feature-level fusion extracts features such as edges, shapes, and textures from the image and then fuses them. Downstream tasks can be accomplished more accurately based on these typical fused features. Decision-based fusion fuses information from classified images and refines some uncertain information using decision rules, and the registration requirements are less stringent [6].

There has been considerable research on three fusion methods based on traditional approaches. To achieve effective LCC using multimodal data, this study proposes a stochastic gradient descent method for image fusion, leveraging three directional color components and a Sobel approach [10]. Addressing the issue of noise in SAR images, the study in [12] introduced a bilateral filter method based on a pixel-by-pixel similarity metric. This method successfully fuses multimodal data using a co-aligned optical image as a reference. In the realm of feature-level fusion methods, Zhang et al. [13] proposed a technique that involves extracting spectral, texture, and spatial features from both optical images and SAR data. Additionally, they incorporated the normalized difference in vegetation index, as well as elevation and slope information. Despite the recognized importance of feature normalization in data processing, most existing normalization methods are not suitable for the fusion of multimodal data, primarily due to the differing imaging mechanisms of optical images and SAR data. To address this challenge, the study proposed the use of a scale normalization algorithm. This algorithm is specifically designed to combine multimodal data effectively, thereby facilitating the evaluation of LCC [14]. To integrate contextual information into the multimodal data fusion process for downstream tasks such as LCC,

this study suggests employing the Markov random field (MRF) approach. Decision-level fusion involves combining various options to generate a common decision [15].

The commonly used deep learning methods for multimodal fusion can be broadly categorized into two types: those based on weight sharing and those in which the weights are not shared. Because of the massive differences between SAR data and optical images, many studies have proved that better inference results can be achieved by non-weight-sharing methods [16]. The prevalent deep learning methods are mainly based on feature-level fusion. For example, Zhang et al. [17] suggested constructing a model based on a classical network structure, incorporating a dual encoder and a shared decoder to effectively process both SAR and optical images, furthering the goal of leveraging the complementary aspects of SAR data and optical images. Moreover, this study proposed a triple attention feature network model. This model integrates a self-attention module, a spatial attention module, and a spectral information attention module, which synergistically enhance the utilization of multimodal features. Li et al. [18] proposed a method that involves analyzing distribution histograms of the depth semantic features of both optical and SAR images, revealing the complementary nature of their feature information. They developed a model with a dual-line feature extractor, a multimodal attention module, and a gated heterogeneous data fusion module to improve the accuracy of multimodal fusion segmentation results. Additionally, Li et al. [16] designed a multimodal bilinear fusion network that accomplishes feature extraction using an encoder, a channel selection module for second-order attention, and a bilinear fusion module.

Traditional approaches typically involve manually designed characteristics, such as pixel color in the image space, gradient histograms, and other similar features, rather than depending on domain-specific knowledge [2]. Thus, when using traditional methods relying on parameter settings, whether applying feature-level fusion or decision-level fusion, adjusting the appropriate parameters is a complex task. Moreover, machine learning methods such as support vector machines are limited by their ability to represent manual features, and thus, machine learning does not perform well in terms of robustness and experimental accuracy [19]. The rapid development of deep learning provides a new approach to multimodal data fusion, as this technology can leverage the diverse features of remote sensing images, including spectral, textural, and structural information due to its nonlinear expression and strong performance in feature extraction [7,13]. Therefore, deep learning shows great potential. However, in general, the choice of fusion modality must be based on the specific downstream task, and there is no single fusion modality that works in all situations [19]. For example, in the completion of the underwater detection task, the process suffers from image noise, texture blur, low contrast, color distortion, and impurity particles affecting the optical imaging as a result of blurring atomization, as well as other issues. Therefore, it is often difficult to use single modal data to comprehensively express the characteristics of the object, and it is often necessary to combine multiple features in order to accurately achieve this task. In this case, it is necessary to perform underwater histogram enhancement and Retinex theory enhancement, or simulate the generation of underwater images through generative adversarial networks, underwater restoration, or other processing [20]. At the same time, SAS originates from SAR; as a sensor working in different environments, it can perform underwater imaging, and based on this type of imaging data, multimodal data fusion can be performed to accomplish underwater observation and image interpretation [21,22].

In our research, we have developed modules that can be integrated into popular encoding–decoding network architectures. These enhancements improve feature extraction findings and overcome the limitations of optical images, such as their susceptibility to weather conditions and inability to gather real-time feature information. By using multimodal data, our model overcomes these challenges. The proposed model consists of two primary components. The first step involves processing SAR images to address the significant differences in image disparities between SAR and optical data. We also introduced a two-way network designed for multimodal-based semantic segmentation tasks. An image

input into this dual-input network model undergoes image-level fusion and encoding, resulting in a shallow feature map rich in geometric features. We focused on stitching shallow feature maps due to their similarity in geometric features, such as shapes, found in both optical images and SAR data. An attention module was incorporated to extract the correlation and rich contextual information of the multimodal data effectively. This module ensures that the fusion process goes beyond simple concatenation, preserving the complementary characteristics of the multimodal data [19].

(1) We suggest implementing image-level fusion through the PCA transform prior to performing feature-level fusion of multimodal data.

(2) Our approach introduces a two-input network model designed explicitly for feature-level fusion of multimodal data. In this model, shallowly encoded feature maps are fused, and an attention model is incorporated to achieve the LCC task for multiclass targets effectively.

(3) The efficacy of our proposed method is validated through experimental results on various public multi-class LCC datasets.

The structure of the remainder of this paper is as follows: Section 2 provides an overview of the image-level and feature-level fusion methods we have proposed, as well as the loss function utilized. Section 3 details the experimental datasets and experimental setting details, and offers qualitative and quantitative analyses of the experimental results for different models based on public datasets. This section also includes a comparison with other prevalent multimodal fusion methods. Finally, Section 4 summarizes our study and outlines potential directions for future research in this field.

2. Overview

To improve the inference results of LCC, we propose a multimodal fusion approach that incorporates both image-level and feature-level fusion methods. Figure 1 depicts the comprehensive architecture of this paper. The methodology is divided into two primary sections: image-level and feature-level fusion. We carried out experiments with several public datasets and established networks, aiming to develop a generalized strategy that effectively utilizes multimodal data for LCC. The detailed experimental procedure is described in the following sections.

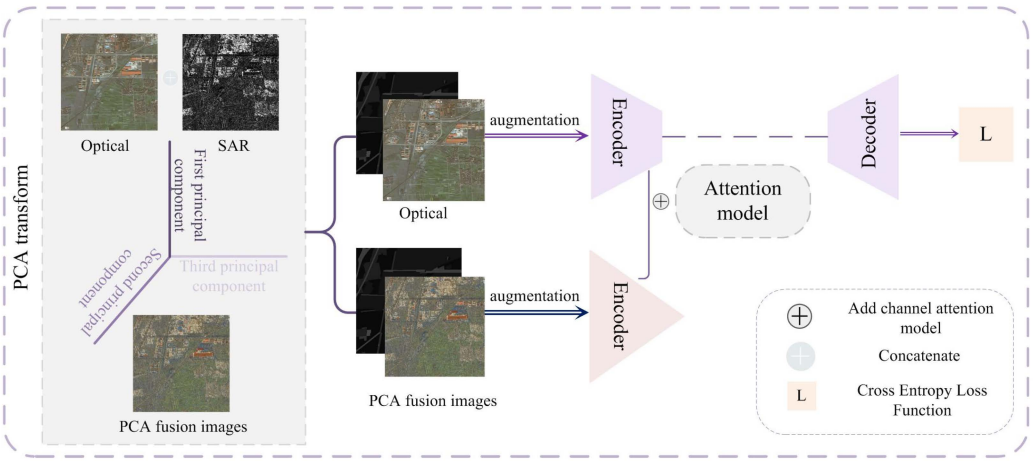

Figure 1. Overview of content in this paper. SAR is first fused with optical image at image-level using PCA transform. Subsequently, shallow feature maps obtained are fused at feature-level, and the attention module is added to complete the encoding–decoding operation, and the task of LCC is accomplished after back-propagation training by means of the cross-entropy loss function.

As presented in Figure 2, the visualization of multimodal data intuitively showcases the optical image and SAR data of typical geographic elements, such as farmland, buildings, water, and forests. A comparative analysis of the figure reveals that distinguishing features

such as farmland is challenging when solely relying on optical images, whereas SAR data provide a clearer classification. For elements such as buildings and water, features such as texture, color, and shape are more discernible in optical images, but these features are less apparent in the SAR data. Therefore, finding an appropriate method to fuse optical images with SAR data is crucial for enhancing the effectiveness of LCC tasks.

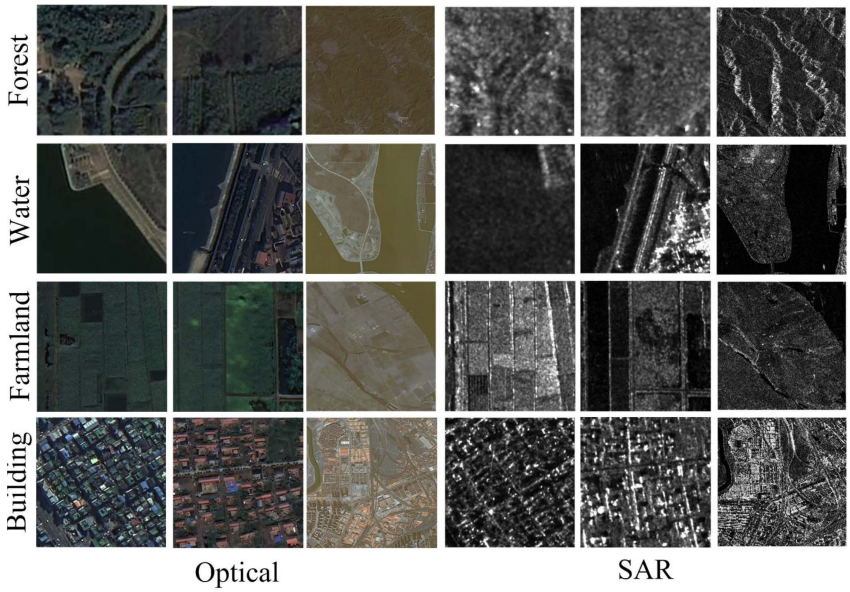

Figure 2. Feature characteristics of the same area from optical and SAR images obtained from different remote sensors, such as the GF-2 and GF-3 satellites.

In optical images, different classes of pixels can display similar spectral characteristics [23]. This phenomenon is evident in Figure 3, which shows spectral reflectance values across different bands. It can be observed that roads and buildings exhibit similar reflectance values in optical images. Similarly, the spectral reflectance values of the farmland and forest are closely aligned in the blue and red bands. As a result, when relying solely on optical images for LCC, features with similar spectral reflectance values, manifesting as similar colors and other features, become challenging to distinguish. However, these features exhibit distinctly different characteristics in SAR data. Consequently, fusing optical images with SAR data proves to be an effective strategy for accomplishing LCC tasks.

Figure 1 demonstrates the proposed process for fusing multimodal data. This process encompasses two key phases to achieve multimodal multitarget semantic segmentation based on remote sensing images: the use of the PCA transform for image-level fusion of optical images with SAR data and feature-level fusion following the encoder stage. For processing optical images, a network model can be chosen from several widely recognized semantic segmentation model structures. These include U-Net [24], UNetPlusPlus [25], EfficientNet-UF [26], and Swin-Transformer [27]. The initial step in the process involves encoding the data using popular encoders designed for SAR data processing. Numerous available encoders come with suitable pre-training weights, enhancing their efficacy. Notable examples of such encoder models include VGG [28] and ResNet [29], among other structures.

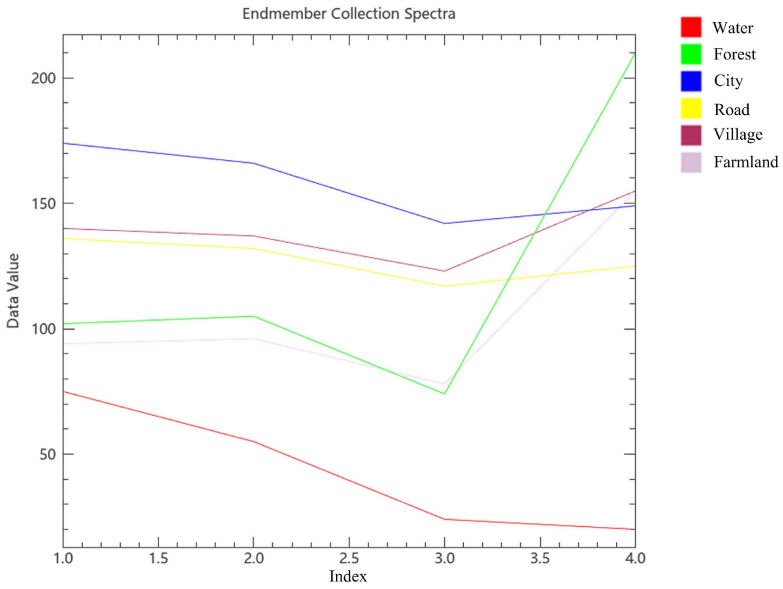

Figure 3. To comparatively analyze the shortcomings of most semantic segmentation tasks using single-modal optical images, spectral reflectance value curves of typical land dividing features in the red, green, blue, and near-infrared bands were plotted to determine the relationship between spectral reflectance of different features. The horizontal axis index denotes the number of bands, and the band number of the used optical images is 4, representing the four bands R-G-B-N, respectively. The vertical axis represents the value of the radiant received by a satellite.

2.1. Image-Level Fusion

The PCA transformation is an effective and widely used method for reducing data dimensionality. It transforms correlated data into a set of uncorrelated features through orthogonal variations. The principal components (PCs) at the forefront represent the greatest differences in the features they contain. This method is notable for its ability to reduce data loss [30].

To achieve feature-level fusion of multimodal data, most current deep-learning-based strategies for multimodal data fusion have developed specialized fusion modules. In this study, the multimodal fusion task was accomplished by first adding an image-level fusion step. The primary steps for processing SAR data are outlined below. First, SAR data significantly differ from optical images, as they employ active microwave transmission for feature detection while simultaneously receiving ground-based echo information. To optimally harness the feature information from both optical and SAR data, we proposes the implementation of PCA fusion for these data types. This strategy is leveraged to reduce the number of parameters necessary for developing relevant deep learning modules, thereby achieving a more efficient feature-level fusion approach. Simultaneously, PCA fusion methods are used to generate three-dimensional images for SAR data input, which can be equipped with pre-training weights bearing optimal initial values. This enhances the efficiency of subsequent training processes. The ultimate goal is to successfully complete the semantic segmentation task using multimodal data while addressing two main challenges: reducing semantic differences between different datasets and accommodating the fact that SAR data typically consist of single-band images. The information from SAR and optical images is then fed into an encoder designed for processing SAR data. These procedural steps are detailed in Figure 4.

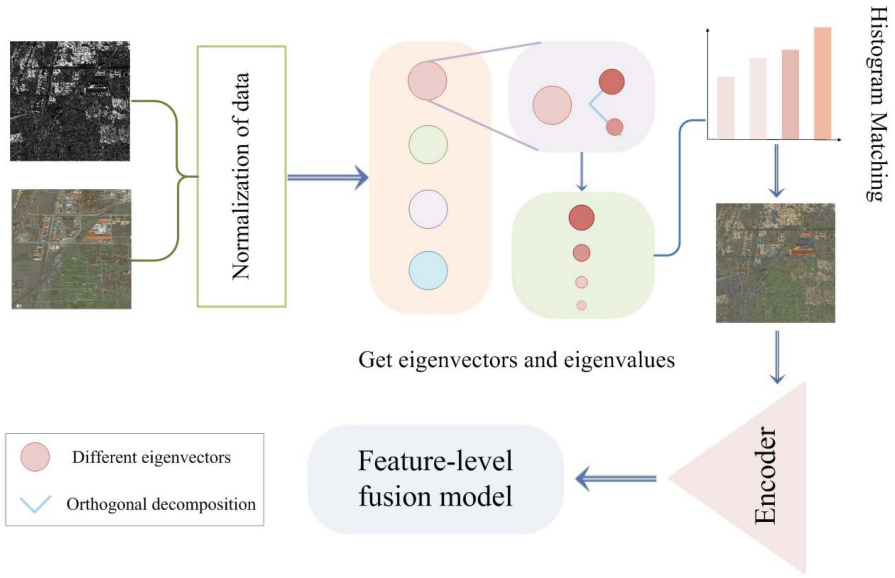

Figure 4. The optical image was fused with the SAR image using PCA. The feature vectors of the optical image and SAR data are calculated and sorted to obtain essential features, after which feature maps are matched, and the PCA inverse transformation is performed to complete the fusion of the images. Subsequently, the optical image is fused with the encoding operation. This figure demonstrates that after applying the PCA approach, the shape features are enhanced. Furthermore, the spatial detail information and spectral similarity aspects are preserved to a large extent in the original image.

2.2. Feature-Level Fusion

To achieve feature-level fusion, we performed separate encoding operations for the optical and fused multimodal images, which had previously undergone image-level fusion. During the encoding process, feature maps rich in geometric features acquired from the shallow layers of the encoder were fused. Additionally, some existing linear fusion methods that perform element-wise summation tend to assign identical weights to different modalities, overlooking the fact that each modality contributes differently to various land classes. In contrast, attention module is employed to assign distinct weights to features based on their respective contributions [2,18,31]. Thus, following this, a channel attention module was added to allocate more weight to the channels containing important information, thus completing the encoding operation. Figure 5 provides a visual representation of this specific fusion method.

In the realm of semantic segmentation models, the choice of encoder for processing optical images is typically predefined. For our research, we selected resnet18 [29] as the encoder for processing SAR data. Additionally, we incorporate an attention module in our method, specifically the channel attention module. Its primary function is assigning proportional weights to different channels, based on the amount of information they contain, after the optical image and SAR data have finished the concatenation process.

After completing the feature fusion step, the probability of the LCC is obtained using softmax. The commonly used loss function in LCC is cross entropy loss [32–34], and some studies explored the influence of inference results after adding the label smoothing to the loss function [35]. To make the experimental inference results comparable, the most commonly used cross-entropy loss function is selected and defined by the following equation:

$$loss = -\sum_{c=1}^{M}[y_{ic}log(p_{ic})] \quad (1)$$

where M denotes the number of categories, y_{ic} takes 0 or 1, 1 if the true category of the sample is c and 0 otherwise, and p_{ic} denotes the probability that sample i belongs to category c.

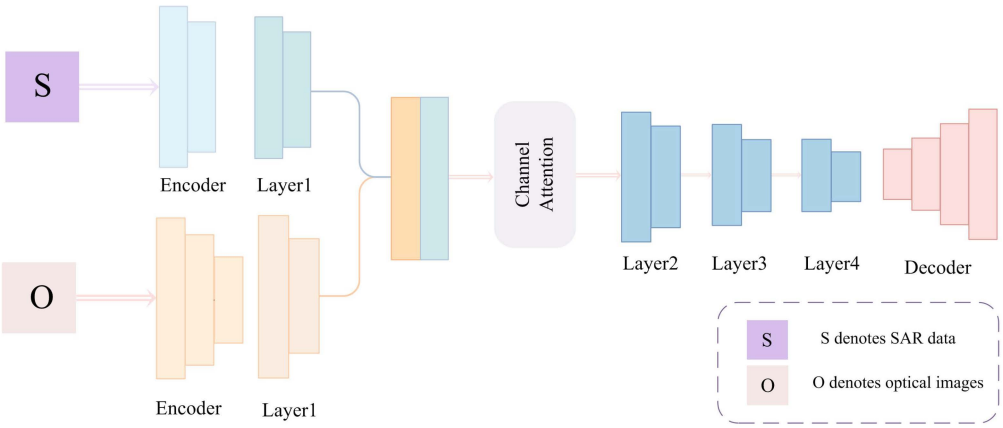

Figure 5. We concatenate shallow feature maps containing rich geometric features and add a channel attention module. We also execute distinct encoding processes for optical images and images that have been fused by PCA approaches. After completing the encoding and decoding processes, the cross-entropy loss is used to achieve semantic segmentation task.

3. Experimental Setup

3.1. Datasets

The WHU-OPT-SAR dataset [36] contains 100 optical images, each with dimensions of 5556 × 3704 pixels, and corresponding SAR images covering the same geographical area. This area spans approximately 50,000 km^2 in Hubei Province. Characterized by a wide variety of plants and rich terrain, this dataset offers a substantial number of remote sensing images. Additionally, pixel-level annotations are provided, which are invaluable for obtaining precise labels for deep learning-based training. To maintain consistency and fairness in comparison with other datasets, we selected only the first three RGB bands from this dataset for our training. The images were resized to uniform dimensions of 512 × 512 pixels. In total, we generated 5000 images of this size. These were then split into training and validation sets in a ratio of 9:1. To ensure robustness and generalizability of the model, the images were randomized before being allocated to the respective sets.

The DDHRNet dataset [37] comprises satellite images obtained using the GF-2 and GF-3 satellites, which provide both optical images and SAR data of identical geographic locations. To capture a wide array of topographic and vegetative features, the dataset includes images from diverse locations such as Xi'an and Dongying in China and Pohang in South Korea. After pre-processing, the ground resolution of the satellite images in this dataset is 1 m, offering a detailed view of the terrain. The remote sensing image data utilized for the experiments varied in resolution depending on the location: images from Pohang and Dongying were processed at a resolution of 256 × 256 pixels, while those from Xi'an were processed at a resolution of 128 × 128 pixels. In order to ensure a comprehensive and unbiased evaluation, the images were divided into training and validation sets. This division was carried out in a ratio of 9:1.

3.2. Implementation Details

During our experimental phase, we utilized three GeForce RTX2080Ti GPUs to conduct training on public datasets. Given the complexity of the model and the uniformity of training hyperparameters across different models, we processed six images simultaneously for each training iteration. The total number of epochs was set to 100. The training was conducted using the PyTorch 1.9.1 deep learning framework, with Python version 3.7.7. The learning rate was set to 0.001, the decay rate was set to 0.1, and the optimizer was

AdamW. For data augmentation, a standardized approach was adopted. This included horizontal flipping of the images and random rotations at angles of 90°, 180°, and 270°. This augmentation method was applied only to the training set, not during the inference phase, to ensure the integrity and realism of the test data. For the LCC experiments, we specifically selected cropped images from the WHU-OPT-SAR dataset and no-cloud images from the DDHRNet dataset.

In this study, we adhered to the same evaluation criteria commonly used in remote sensing image-based LCC methods for assessing the validity of our experimental results. These metrics included precision, recall, accuracy, mIoU, and mPA. The definitions and calculations for these metrics are as follows (Pr. deotes Precision, Ac. denotes accuracy, and Rec. denotes recall):

$$Pr. = \frac{TP}{TP + FP} \times 100\% \tag{2}$$

$$Rec. = \frac{TP}{TP + FN} \times 100\% \tag{3}$$

$$mPA = \frac{1}{ncls} \times Rec. \times 100\% \tag{4}$$

$$Ac. = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \tag{5}$$

$$mIoU = \frac{1}{ncls} \frac{Pr. \times Rec.}{Pr. + Rec. - Pr. \times Rec.} \times 100\% \tag{6}$$

3.3. Experiment and Discussion

Tables 1 and 2 present the experimental results of our model alongside various comparison methods, employing three commonly used accuracy evaluation metrics. Given the limited interpretability of SAR data and the typically unsatisfactory experimental inference accuracy, we established three sets of experiments on the public LCC dataset. These experiments included training single-modal data containing only optical images on the same base model. This approach allowed for a direct comparison between our proposed multimodal data fusion strategy and other prominent multimodal fusion methods. The results from these experiments indicate that our fusion strategy achieves superior inference accuracy, particularly when applied to a model with inherently high inference accuracy. This outcome not only attests to the effectiveness of our fusion strategy, but also demonstrates that it can successfully exploit the complementary information in optical images and SAR data to enhance LCC inference results.

Table 1. In our experiments, the WHU-OPT-SAR dataset was utilized, selecting only the RGB bands from this dataset to ensure a fair comparison with other public datasets. For the fusion of SAR data, we implemented the proposed method on two foundational models: U-Net and EfficientNet-UF (O and S denote optical and SAR, respectively, and E-UF denotes EfficientNet-UF).

Datasets	Method	mIoU%	mPA%	Accuracy%
WHU-OPT-SAR	Deeplab-v3-O+S [38]	39.97	49.95	76.68
	SOLC-O+S [38]	42.80	53.21	79.11
	U-Net-O [24]	50.43	61.43	80.37
	U-Net+Ours	52.87	63.22	80.88
	E-UF-O [26]	51.55	65.05	80.87
	E-UF+Ours	53.37	64.98	81.16

Table 2. The DDHRNet dataset was employed in our experiments. To highlight the effectiveness of our proposed fusion method, we conducted experiments across three different areas, each characterized by a unique style of feature information distribution. Similar to the approach with the WHU-OPT-SAR dataset, the fusion of SAR data was executed using the same two base models: U-Net and EfficientNet-UF (O and S denote optical and SAR, respectively, and E-UF denotes EfficientNet-UF).

Datasets	Area	Method	mIoU%	mPA%	Accuracy%
DDHRNet	Korea	Deeplab-v3-O+S [38]	68.75	79.13	86.23
		SOLC-O+S [38]	74.38	83.05	89.31
		U-Net-O [24]	83.30	90.02	93.68
		U-Net+Ours	87.04	92.05	95.35
		E-UF-O [26]	89.83	93.95	96.52
		E-UF+Ours	91.42	94.91	97.07
	Shandong	SOLC-O+S [38]	72.88	78.33	92.73
		Deeplab-v3-O+S [38]	75.46	83.47	90.37
		U-Net-O [24]	79.38	86.90	91.20
		U-Net+Ours	83.26	89.63	92.79
		E-UF-O [26]	85.62	91.15	94.23
		E-UF+Ours	86.52	91.86	94.63
	Xi'an	SOLC-O+SAR [38]	67.65	73.82	90.87
		Deeplab-v3-O+S [38]	69.50	76.02	91.91
		U-Net-O [24]	78.42	83.40	95.26
		U-Net+Ours	78.99	83.66	95.50
		E-UF-O+S [26]	82.18	86.60	96.28
		E-UF+Ours	83.09	87.45	96.49

The inference results for SOLC and Deeplab-v3 were obtained through the reproduction of the official code. To maintain experimental fairness, the same data augmentation methods applied in our experiments were also used during the training phase of these models. The chosen loss function was cross-entropy, which does not incorporate weights assigned to different classes in the open-source code and is consistent with the loss function introduced in our paper.

The inclusion of SAR data has been found to effectively improve the inference results across different models after utilizing our proposed method, notably improving evaluation metrics for various classes. Specifically, with the WHU-OPT-SAR dataset, there was a significant improvement in the mIoU of the inference results. For instance, when the U-Net network was used for the inference process, the mIoU improved by 2.44%, while in the DDHRNet dataset, improvements were noted as 3.74% for the Korea dataset, 3.88% for the Shandong dataset, and 0.57% for the Xi'an dataset. Moreover, when our proposed method was integrated into the EfficientNet-UF model with the WHU-OPT-SAR dataset, the mIoU improved by 1.82%. In the Korea dataset of DDHRNet, the application of our method with the EfficientNet-UF model led to mIoU improvement of 1.59%. For the Shandong dataset, there was an increase of 0.90%, and in the Xi'an dataset, the mIoU improved by 0.91%. It is noteworthy that while the comparison methods achieved accuracy metrics comparable to those of our method, the mIoU values were generally lower than those achieved with our proposed method. A potential reason for this could be the training strategy used in the open source models, which might have involved different weights according to the distribution of classes. However, given the complexity and variability of class distribution across different datasets and real-world scenarios, and to validate our method across various public datasets, we opted not to set different weights for classes during the training phase. This decision could lead to the comparative methods achieving higher accuracy but lower mIoU values.

To thoroughly assess the effect of our proposed multimodal data fusion approach on LCC tasks using public datasets, we focused on the mIoU evaluation metrics for specific classes. These classes include farmland, city, village, water, forest, road, others, and background, encompassing all pixels in an image in these eight classes. From the data presented in Table 3, it is evident that the fusion of SAR data using our method effectively improves the distinction between the background and other classes. Moreover, in the

WHU-OPT-SAR dataset, there were marked improvements in the mIoU values for the farmland, city, village, water, and road classes across different models. However, it was observed that in the forest and other classes, the mIoU values experienced a slight decrease when using the EfficientNet-UF and U-Net models. We attribute this phenomenon to the unique spectral characteristics of forests. As green vegetation, forests exhibit high spectral reflectance in the near-infrared and green wavelength bands, making them relatively easy to distinguish from other features in optical images. We provide solutions for accomplishing the LCC task based on remote sensing images obtained from different sensors.

Table 3. In the WHU-OPT-SAR dataset, mIoU evaluation criteria were calculated for all classes to which each pixel in the sample belonged (bg denotes background, O and S denote optical and SAR, respectively, and E-UF denotes EfficientNet-UF).

Method	bg	Farmland	City	Village	Water	Forest	Road	Others
Deeplab-v3-O+S	10.84	62.31	52.75	38.09	58.65	75.14	11.12	10.84
SOLC-O+S	0.39	65.95	56.48	47.20	61.97	76.93	25.73	7.75
U-Net-O	38.64	67.70	54.96	47.31	64.80	77.53	33.56	18.93
U-Net+Ours	56.42	68.52	56.04	47.68	65.32	77.95	33.67	17.34
E-UF-O	40.35	68.88	54.12	47.76	64.83	78.20	37.35	20.88
E-UF+Ours	48.73	68.90	55.15	48.67	66.62	77.73	38.68	22.44

As shown in Table 4, the inference results for different classes such as building, road, farmland, water, greenery, and others within the DDHRNet dataset demonstrate that the mIoU for all six of these classes improved upon applying the proposed method. This outcome effectively validates the efficacy of our approach. In regions such as Korea, Shandong, and Xi'an, which each have distinct feature distribution styles and characteristics, significant enhancements were noted in classes such as roads, farmland, greenery, and others across two commonly used base models. The category of buildings showed notable improvement in both models in Korea and Shandong, but there was no substantial enhancement in the U-Net network in Xi'an. Additionally, the U-Net model did not show significant mIoU improvement in the water class in Shandong. This variance in performance, particularly with the U-Net model, might be attributed to its simpler architecture, which potentially limits its ability to effectively utilize the rich and complementary semantic information available in the fused multimodal data.

However, considering the overall inference results, it is clear that significant mIoU improvement is attainable using our proposed strategy, regardless of whether the eight classes WHU-OPT-SAR dataset or the six classes DDHRNets dataset are used. Most notably, there is a significant improvement in classes such as road, farmland, and greenery.

The visualization results depicted in Figure 6 clearly illustrate that the LCC results obtained using our proposed method exhibit more detailed completeness, and are closer to the ground truth. The fusion of SAR data into the analysis notably enhances the clarity of land classification inference results, particularly in terms of the edge information of various feature targets. In this study, we employed image-level fusion through the application of PCA and the feature level fusion method, which effectively highlights target objects and larger weights to features with more informative content. Moreover, with the addition of SAR data, we enhanced the remote sensing images of ground information, which were acquired using various imaging modalities. The results of extracting features, particularly for elements such as farmland, cities, villages, and water, show considerable improvement. Notably, in urban areas and water, the number of misclassified pixels was significantly reduced with our proposed fusion strategy. Farmland, which is often challenging to differentiate in optical images, showed marked improvement with the addition of SAR data. In the DDHRNet dataset, the extraction results for roads, farmland, buildings, and greenery also saw significant improvements when using our proposed method. It can be concluded from the LCC inference results that using our method can improve model segmentation performance, with different kinds of segmentation results being further

refined to obtain superior LCC classification results. Additionally, our method can make it easier to recognize different classes from four different regions with six and eight classes, indicating that our method has a clear advantage in predicting multiple semantic classes.

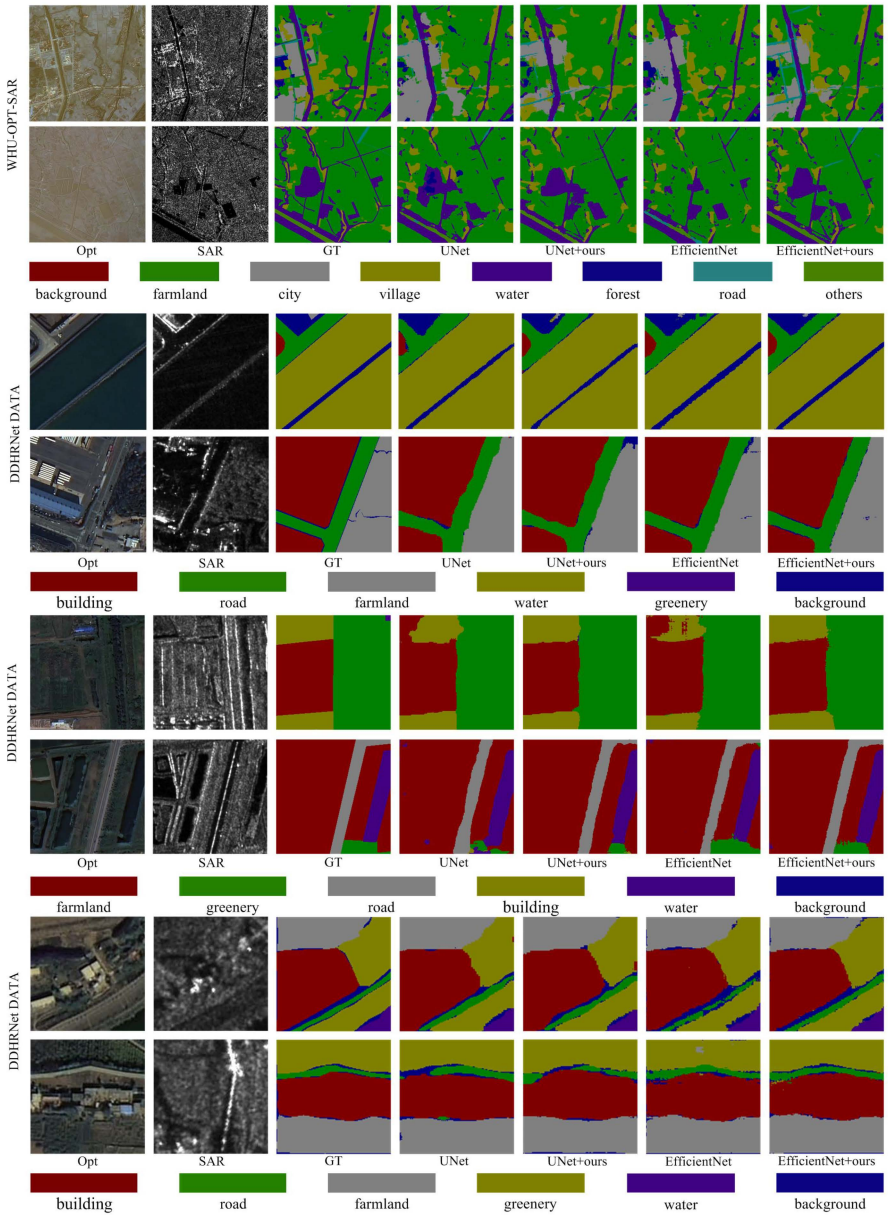

Figure 6. The inference results, both with and without the implementation of the proposed method, were visualized using the WHU-OPT-SAR and DDHRNet datasets. These visualization results were then thoroughly analyzed and compared against the original, SAR, and ground truth images to assess the efficacy of the proposed method.

We incorporated a channel attention module to enhance the feature and multimodal data fusion process. This module effectively strengthened the extraction of important features while simultaneously inhibiting the extraction of unimportant information. The visualization results indicate that our proposed method notably improved the accuracy of target feature extraction by significantly reducing the erroneous extraction of background information.

Table 4. In the DDHRNet dataset, the mIoU evaluation criteria were meticulously calculated for all the classes that are specified in the dataset's labeling system (O and S denote optical and SAR, respectively, and E-UF denotes EfficientNet-UF).

Area	Method	Building	Road	Farmland	Water	Greenery	Others
Korea	Deeplab-v3-O+S	79.61	63.03	66.83	85.51	80.70	36.83
	SOLC-O+S	85.35	54.32	82.19	97.51	77.24	49.67
	U-Net-O	91.61	72.83	89.17	98.34	87.18	60.69
	U-Net+Ours	94.04	79.50	91.81	98.42	90.86	67.61
	E-UF-O	95.73	84.85	94.30	98.91	92.96	72.20
	E-UF+Ours	96.30	87.51	95.12	99.10	93.73	76.78
Shandong	SOLC-O+S	89.92	70.73	91.74	82.96	95.95	6.00
	Deeplab-v3-O+S	87.13	49.41	84.27	96.38	82.51	53.05
	U-Net-O	84.35	74.84	78.02	90.67	88.80	59.62
	U-Net+Ours	87.07	81.04	82.14	90.62	90.60	68.10
	E-UF-O	90.41	83.15	85.73	93.56	90.45	70.44
	E-UF+Ours	91.52	84.08	86.13	93.72	92.07	71.58
Xi'an	SOLC-O+S	86.67	57.49	90.33	81.13	91.98	0.12
	Deeplab-v3-O+S	88.97	60.36	91.45	82.83	93.31	0.10
	U-Net-O	93.45	81.37	96.09	90.01	96.19	13.39
	U-Net+Ours	93.38	81.58	96.13	90.64	96.29	15.91
	E-UF-O	94.31	84.84	96.83	91.73	95.96	29.41
	E-UF+Ours	94.56	85.30	97.16	92.00	97.05	32.48

4. Conclusions and Discussion

Because SAR data and optical images contain complementary feature information, we have achieved LCC using a multimodal approach that fuses SAR data with optical images. Multimodal fusion methods can generally be categorized into three types: image-level, feature-level, and decision-level. Considering the distinct nature of the information in SAR data and optical images, this paper proposes using the PCA algorithm to retain the important components of both as one branch of a dual input system for semantic segmentation. We spliced shallow feature maps obtained from different images using two encoders into the depth dimensions. To more effectively capture important features from SAR data and optical images, a channel attention module was added after concatenating the shallow feature maps. This addition helps to extract important information from the fused SAR features and facilitates the subsequent decoding process. To validate the effectiveness of our methods, we conducted experiments on multimodal fusion using two public datasets encompassing various geographical areas. We utilized the U-Net and EfficientNet-UF networks as our base models and incorporated our proposed multimodal fusion method. The quantitative data suggest that our method notably improves the classification accuracy of different features, significantly improving the results for multi-class features. Visualization results indicate that our approach can extract better features and define boundaries more effectively. Comparing our method with other multimodal fusion approaches, it is evident that our method achieves superior accuracy. In future research, our aim is to fuse hyperspectral images, optical images, and SAR data, as well as multimodal data with greater variance in the resolution of remote sensing images. We plan to leverage large or world models more effectively to minimize sample use and enhance inference results. The fusion of multimodal data requires pre-processing tasks, such as registration and scatter noise reduction, which have a significant impact on the inference results of the downstream tasks and are cumbersome to perform. In our future work, we will consider exploring more adaptive registration methods for downstream tasks or using fewer samples to complete the LCC task.

Author Contributions: Methodology, Y.Q. and R.Z.; investigation, J.L. and S.J.; resources, H.G. and A.Y.; writing—original draft preparation, Y.Q. and R.Z.; writing—review and editing, J.L. and S.J.; visualization, Y.Q. and R.Z.; supervision, H.G. and A.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This researchwas funded by the National Natural Science Foundation of China under Grants 42101458, 42171456, 42130112, 41901285, and 42277478.

Data Availability Statement: The raw data will be made available on the request.

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

References

1. Xia, J.; Yokoya, N.; Baier, G. DML: Differ-Modality Learning for Building Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
2. Peng, B.; Zhang, W.; Hu, Y.; Chu, Q.; Li, Q. LRFFNet: Large Receptive Field Feature Fusion Network for Semantic Segmentation of SAR Images in Building Areas. *Remote Sens.* **2022**, *14*, 6291. [CrossRef]
3. Wu, W.; Guo, S.; Shao, Z.; Li, D. 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.* **2023**, *16*, 2573–2588. [CrossRef]
4. 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]
5. Xu, H.; He, M.; Rao, Z.; Li, W. Him-Net: A New Neural Network Approach for SAR and Optical Image Template Matching1. In Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 19–22 September 2021; pp. 3827–3831. [CrossRef]
6. Kulkarni, S.; Rege, P.P. Pixel level fusion techniques for SAR and optical images: A review. *Inf. Fusion* **2020**, *59*, 13–29. [CrossRef]
7. Zhu, B.; Zhou, L.; Pu, S.; Fan, J.; Ye, Y. Advances and Challenges in Multimodal Remote Sensing Image Registration. *IEEE J. Miniaturization Air Space Syst.* **2023**, *4*, 165–174. [CrossRef]
8. Gao, C.; Li, W. Multi-Scale PIIFD for Registration of Multi-Source Remote Sensing Images. *J. Beijing Inst. Technol.* **2021**, *30*, 113–124.
9. Ye, Y.; Zhu, B.; Tang, T.; Yang, C.; Xu, Q.; Zhang, G. A Robust Multimodal Remote Sensing Image Registration Method and System Using Steerable Filters with First- and Second-order Gradients. *arXiv* **2022**, arXiv:2202.13347. [CrossRef]
10. Shakya, A.; Biswas, M.; Pal, M. Fusion and Classification of SAR and Optical Data Using Multi-Image Color Components with Differential Gradients. *Remote Sens.* **2023**, *15*, 274. [CrossRef]
11. Lewis, J.J.; O'Callaghan, R.J.; Nikolov, S.G.; Bull, D.R.; Canagarajah, N. Pixel- and region-based image fusion with complex wavelets. *Inf. Fusion* **2007**, *8*, 119–130. [CrossRef]
12. Gaetano, R.; Cozzolino, D.; D'Amiano, L.; Verdoliva, L.; Poggi, G. Fusion of sar-optical data for land cover monitoring. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 23–28 July 2017; pp. 5470–5473. [CrossRef]
13. 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]
14. Zhang, H.; Lin, H.; Li, Y. Impacts of Feature Normalization on Optical and SAR Data Fusion for Land Use/Land Cover Classification. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 1061–1065. [CrossRef]
15. Maggiolo, L.; Solarna, D.; Moser, G.; Serpico, S.B. Optical-SAR Decision Fusion with Markov Random Fields for High-Resolution Large-Scale Land Cover Mapping. In Proceedings of the IGARSS 2022—2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 5508–5511. [CrossRef]
16. Li, X.; Lei, L.; Sun, Y.; Li, M.; Kuang, G. Multimodal Bilinear Fusion Network With Second-Order Attention-Based Channel Selection for Land Cover Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 1011–1026. [CrossRef]
17. Zhang, D.; Gade, M.; Zhang, J. SOF-UNet: SAR and Optical Fusion Unet for Land Cover Classification. In Proceedings of the IGARSS 2022—2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 907–910. [CrossRef]
18. Li, X.; Lei, L.; Sun, Y.; Li, M.; Kuang, G. Collaborative Attention-Based Heterogeneous Gated Fusion Network for Land Cover Classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 3829–3845. [CrossRef]
19. Li, Y.; Zhou, Y.; Zhang, Y.; Zhong, L.; Wang, J.; Chen, J. DKDFN: Domain Knowledge-Guided deep collaborative fusion network for multimodal unitemporal remote sensing land cover classification. *ISPRS J. Photogramm. Remote Sens.* **2022**, *186*, 170–189. [CrossRef]
20. Lin Sen, Z.Y. Reiview on key technologies of target exploration in underwater optical images. *Laster Optoelectron. Prog.* **2020**, *57*, 060002. [CrossRef]
21. Zhang, X.; Wu, H.; Sun, H.; Ying, W. Multireceiver SAS Imagery Based on Monostatic Conversion. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 10835–10853. [CrossRef]

22. Yang, P. An imaging algorithm for high-resolution imaging sonar system. *Multimed. Tools Appl.* **2023**. [CrossRef]
23. Luo, D.; Li, L.; Mu, F.; Gao, L. Fusion of high spatial resolution optical and polarimetric SAR images for urban land cover classification. In Proceedings of the 2014 Third International Workshop on Earth Observation and Remote Sensing Applications (EORSA), Changsha, China, 1–14 June 2014; pp. 362–365. [CrossRef]
24. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv* **2015**, arXiv:1505.04597.
25. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. *arXiv* **2018**, arXiv:1807.10165.
26. Quan, Y.; Yu, A.; Cao, X.; Qiu, C.; Zhang, X.; Liu, B.; He, P. Building Extraction From Remote Sensing Images With DoG as Prior Constraint. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 6559–6570. [CrossRef]
27. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October2021; pp. 9992–10002. [CrossRef]
28. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2015**, arXiv:1409.1556.
29. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
30. Makar, R.; Shahin, S.; El-Nazer, M.; Wheida, A. Development of a PCA-based land use/land cover classification utilizing Sentinel-2 time series. *Middle East J. Agric. Res.* **2022**, *11*, 630–637.
31. Hazirbas, C.; Ma, L.; Domokos, C.; Cremers, D. FuseNet: Incorporating Depth into Semantic Segmentation via Fusion-Based CNN Architecture. In *Computer Vision—ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, 20–24 November 2016, Revised Selected Papers, Part I 13*; Lai, S.H., Lepetit, V., Nishino, K., Sato, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2017; pp. 213–228.
32. Yao, J.; Cao, X.; Hong, D.; Wu, X.; Meng, D.; Chanussot, J.; Xu, Z. Semi-Active Convolutional Neural Networks for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5537915. [CrossRef]
33. Dong, W.; Zhang, T.; Qu, J.; Xiao, S.; Zhang, T.; Li, Y. Multibranch Feature Fusion Network With Self- and Cross-Guided Attention for Hyperspectral and LiDAR Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5530612. [CrossRef]
34. Quan, D.; Wei, H.; Wang, S.; Gu, Y.; Hou, B.; Jiao, L. A Novel Coarse-to-Fine Deep Learning Registration Framework for Multimodal Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5108316. [CrossRef]
35. Müller, R.; Kornblith, S.; Hinton, G.E. When Does Label Smoothing Help? *arXiv* **2019**, arXiv:1906.02629.
36. 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]
37. Ren, B.; Ma, S.; Hou, B.; Hong, D.; Chanussot, J.; Wang, J.; Jiao, L. 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.* **2022**, *112*, 102896. [CrossRef]
38. Sun, Y.; Zhao, Y.; Wang, Z.; Fan, Y. SOLC. 2022. Available online: https://github.com/yisun98/SOLC (accessed on 22 January 2020).