remote sensing

MDPI

Article

HGR Correlation Pooling Fusion Framework for Recognition and Classification in Multimodal Remote Sensing Data

Hongkang Zhang, Shao-Lun Huang * and Ercan Engin Kuruoglu

Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China; zhanghk21@mails.tsinghua.edu.cn (H.Z.); kuruoglu@sz.tsinghua.edu.cn (E.E.K.)
* Correspondence: shaolun.huang@sz.tsinghua.edu.cn

Abstract: This paper investigates remote sensing data recognition and classification with multimodal data fusion. Aiming at the problems of low recognition and classification accuracy and the difficulty in integrating multimodal features in existing methods, a multimodal remote sensing data recognition and classification model based on a heatmap and Hirschfeld–Gebelein–Rényi (HGR) correlation pooling fusion operation is proposed. A novel HGR correlation pooling fusion algorithm is developed by combining a feature fusion method and an HGR maximum correlation algorithm. This method enables the restoration of the original signal without changing the value of transmitted information by performing reverse operations on the sample data. This enhances feature learning for images and improves performance in specific tasks of interpretation by efficiently using multi-modal information with varying degrees of relevance. Ship recognition experiments conducted on the QXS-SROPT dataset demonstrate that the proposed method surpasses existing remote sensing data recognition methods. Furthermore, land cover classification experiments conducted on the Houston 2013 and MUUFL datasets confirm the generalizability of the proposed method. The experimental results fully validate the effectiveness and significant superiority of the proposed method in the recognition and classification of multimodal remote sensing data.

Keywords: remote sensing; multimodal fusion; HGR maximal correlation; ship recognition; land cover classification

Citation: Zhang, H.; Huang, S.-L.; Kuruoglu, E.E. HGR Correlation Pooling Fusion Framework for Recognition and Classification in Multimodal Remote Sensing Data. *Remote Sens.* **2024**, *16*, 1708. https://doi.org/10.3390/rs16101708

Academic Editor: Filomena Romano

Received: 20 March 2024
Revised: 8 May 2024
Accepted: 9 May 2024
Published: 11 May 2024

1. Introduction

In practical applications, there are certain limitations in the information content, resolution, and spectrum of single-mode scenes, making it difficult to meet the application requirements [1–3]. Multimodal image fusion has become an attractive research direction. Recently, with the in-depth research of fusion algorithms, multimodal recognition technology has made rapid progress [4–6]. The multimodal multi-tasking basic model has been widely studied in the field of computer vision. It combines image data with text or speech data as multimodal input and sets different pre-training tasks for different modal branches to enable the model to learn and understand the information between modalities. Multimodal fusion can improve the recognition rate and has better robustness and stability [7], further promoting the development of multimodal image fusion technology. At present, image fusion has been widely used in remote sensing image fusion [8], visible and infrared image fusion [9] multi-focus image fusion [10], multi-exposure image fusion [11], medical imaging fusion [12], etc.

In recent years, marine ship detection has been extensively used in many fields such as fishery management and navigation supervision. Determining how to achieve accurate detection through multimodal fusion of marine ships has great strategic significance in both civil and military fields. Thus, Cao et al. [13] proposed a ship recognition method based on morphological watershed image segmentation and Zemyk moments for ship extraction and recognition in video surveillance frame images. Wang et al. [14] developed a SAR

ship recognition method based on multi-scale feature attention and an adaptive weighted classifier. Zhang et al. [15] presented a fine-grained ship image recognition network based on the bilinear convolutional neural network (BCNN). Han et al. [16] proposed a new efficient information reuse network (EIRNet), and based on EIRNet, they designed a dense feature fusion network (DFF-Net), which reduces information redundancy and further improves the recognition accuracy of remote sensing ships. Liu, Chen, and Wang [17] fused optical images with SAR images, utilized feature point matching, contour extraction, and brightness saliency to detect ship components, and identified ship target types based on component information voting results. It is challenging to realize information separation under maximum use of information without compromising image quality.

Meanwhile, in the past decade, the application of remote sensing based on deep learning has made significant advancements in object detection, scene classification, land use segmentation, and recognition. This is mainly because deep neural networks can effectively map remote sensing observations into the needed geographic knowledge through their strong feature extraction and representation capabilities [18–20]. Existing remote sensing interpretation methods mainly adopt manual visual interpretation and semi-automatic techniques based on accumulated expert knowledge, showing high accuracy and reliability. Artificial intelligence technology represented by deep learning is widely used in remote sensing image interpretation [21] and has greatly improved the efficiency of remote sensing data interpretation. For instance, entropy decomposition was utilized to identify crops from synthetic aperture radar (SAR) images [22]. Similarly, normative forests were used to classify hyperspectral images [23]. Jafarzadeh et al. [24] employed several tree-based classifiers' bagging and boosting sets to classify SAR, hyperspectral, and multispectral images.

Compared to single sensors, multi-sensor or remote sensing data provides different descriptions of ground objects, thereby providing richer information for various application tasks. In the field of remote sensing, modality can usually be regarded as the imaging results of the same scene and target under different sensors, and using multimodal data for prediction and estimation is a research hotspot in this field. Given this, an integration method using intensity tone saturation (IHS) transform and wavelet was adopted to fuse SAR images with medium-resolution multispectral images (MSIs) [25]. Cao et al. [26] proposed a method for monitoring mangrove species using rotating forest fusion HSI and LiDAR images. Hu et al. [27] developed a fusion method for PolSAR and HSI data, which extracts features from the two patterns at the target level and then fuses them for land cover classification. Li et al. [28] introduced an asymmetric feature fusion idea for hyperspectral and SAR images. This idea can be extended to the fields of hyperspectral and LiDAR images. Although the above studies have realized the fusion of multiple remote sensing data, designing the loss function specifically needs further investigation. Multimodal data fusion [29,30] is one of the most promising research directions for deep learning in remote sensing, particularly when SAR and optical data are combined because they have highly different geometrical and radiometric properties [31,32].

Meanwhile, the Hirschfeld–Gebelein–Rényi (HGR) maximum correlation [33] has been widely used as an information metric for studying inference and learning problems [34]. In the field of multimodal fusion based on HGR correlation, Liang et al. [35] introduced the HGR maximum correlation terms into the loss function for person recognition in multimodal data. Wang et al. [36] proposed Soft-HGR, a novel framework to extract informative features from multiple data modalities. Ma et al. [37] developed an efficient data augmentation framework by designing a multimodal conditional generative adversarial network (GAN) for audiovisual emotion recognition. However, the values of the transmitted data are changed in the data fusion process.

Inspired by previous studies, the issues of remote sensing data recognition and classification with multimodal data fusion are studied. The main innovations of this article are stated as follows:

(1) An HGR correlation pooling fusion algorithm is developed by integrating a feature fusion method with an HGR correlation algorithm. This framework adheres to the principle of relevance correlation and segregates information based on its intrinsic relevance into distinct classification channels. It enables the derivation of loss functions for positive, zero, and negative samples. Then, a tailored overall loss function is designed for the model, which significantly enhances feature learning in multimodal images.
(2) A multimodal remote sensing data recognition and classification model is proposed, which can achieve information separation under maximum utilization. The model enhances the precision and accuracy of target recognition and classification while preserving image information integrity and image quality.
(3) The HGR pooling specifically addresses multimodal pairs (vectors) and intervenes in the information transmission process without changing the value of the transmitted information. It enables inversion operations on positive, zero, and negative sample data in the original signal of the framework, thereby supporting traceability for the restoration of the original signal. This advancement greatly improves the interpretability of the data.

2. Related Work

To date, multimodal data fusion has been widely used in remote sensing [2,38]. In most cases, multimodal data recognition systems are much more accurate than the corresponding optimal single-modal data recognition systems [39]. According to the fusion level, the fusion strategies between various modalities can be mainly divided into data-level fusion, feature-level fusion, and decision-level fusion [40]. Data-level fusion is aimed at the data without special processing for each mode. The original data of each mode is combined without pretreatment to obtain the data after the mode function. Finally, the fusion data are taken as the input of the identification network for training or identification. Feature-level fusion concatenates the features of each modality into a large feature vector, which is then fed into a classifier for classification and recognition. Decision-level fusion determines the weights and fusion strategies of each modality based on their credibility after obtaining the prediction probability through a classifier, and then it obtains the fused classification results.

The complexity of the above three fusion strategies decreases in sequence, and their dependence on the rest of the system processes increases in sequence. Usually, multimodal fusion strategies are selected based on specific situations. With the improvement in hardware computing power and the increasing demand for applications, the studies of data recognition, which contains massive data information and has mature data collection methods, are constantly enriched.

In order to ascertain the degree of correlation and to identify the most informative features, the Hirschfeld–Gebelein–Rényi (HGR) maximal correlation is employed as a normalized measure of the dependence between two random variables. This has been widely applied as an information metric to study inference and learning problems. In [33], the sample complexity of estimating the HGR maximal correlation functions comes from the alternating conditional expectation algorithm using training samples from large datasets. By using the HGR maximal correlation in [37], the high dependence between the different modalities in the generated multimodal data is modeled. In this way, it exploits different modalities in the generated data to approximate the real data. Although these studies have yielded promising outcomes, it is difficult to achieve accurate detection through multimodal fusion of marine ships, and the interrelationships between modules have not been fully elucidated.

Feature-level fusion can preserve more data information. It first extracts features from the image and then performs fusion. Pedergnana et al. [41] used optical and LiDAR data by extracting extended attribute contours of the two modalities and connecting them with the original modalities. Then, a two-layer DBN network structure was proposed, which first learns the features of the two modalities separately and then connects the features of

the two modalities to learn the second layer. Finally, a support vector machine (SVM) is utilized to evaluate and classify the connected features [42].

However, feature-level fusion requires high computing power and is prone to the curse of dimensionality, and the application of decision-level fusion is also common. To address these issues, a SAR and infrared fusion scheme based on decision-level fusion was introduced [43]. This scheme uses a dual-weighting strategy to measure the confidence of offline sensors and the reliability of online sensors. The structural complexity of decision-level fusion is relatively low and does not require strict temporal synchronization, which performs well in some application scenarios.

3. Methodology

In this section, the details of the proposed CNN-based special HGR correlation pooling fusion framework for multimodal data are introduced. The framework can preserve adequate multimodal information and extract the correlation between modal 1 and modal 2 data so that discriminative information can be learned more directly.

3.1. Problem Definition

Given paired observations from multimodal data $\{(x^{(i)}, y^{(i)}) \mid x^{(i)} \in R^{n1}, y^{(i)} \in R^{n2}, i = 1, \ldots, N\}$, let x and y represent the modal 1 image and modal 2 image with dimensionalities R^{n1} and R^{n2}, respectively. The ith components $x^{(i)} \in x$, $y^{(i)} \in y$ match each other and come from the same region, while the ith component $x^{(i)} \in x$ and the jth component $y^{(j)} \in y$ do not match each other and come from different regions.

3.2. Model Overview

In this paper, to solve the problems of low recognition and classification accuracy and difficulty in effectively integrating multimodal features, an HGR maximal correlation pooling fusion framework is proposed for recognition and classification in multimodal remote sensing data. The overall structure of the framework is shown in Figure 1. In the subsequent subsections, the model will be discussed in detail.

For multimodal image pairs, the framework has two separate feature extraction networks. To reduce the dimensionality of features, following the feature extraction backbone, a 1×1 convolution layer is used, and the modal 1 feature map and modal 2 feature map are obtained separately. Then, a special HGR maximal correlation pooling layer is employed. The HGR pooling handles multimodal pairs (vectors) and only intervenes in the information transmission without changing the value of the transmitted information. The principle is to filter the values of information with different relevant characteristics and transmit them to the corresponding subsequent classification channels. The feature data are processed to obtain positive sample data, zero sample data, and negative sample data for modal 1 and modal 2 features, respectively. Then, the three types of sample data from modal 1 and modal 2 are input into the ResNet50 [44] network to extract feature vectors, and feature level fusion is performed on the corresponding feature vectors to obtain fused positive samples, fused zero samples, and fused negative samples.

Finally, the positive, zero, and negative samples of modal 1/modal 2 images are fused respectively using the recognition and classification network, thereby accomplishing multimodal recognition tasks.

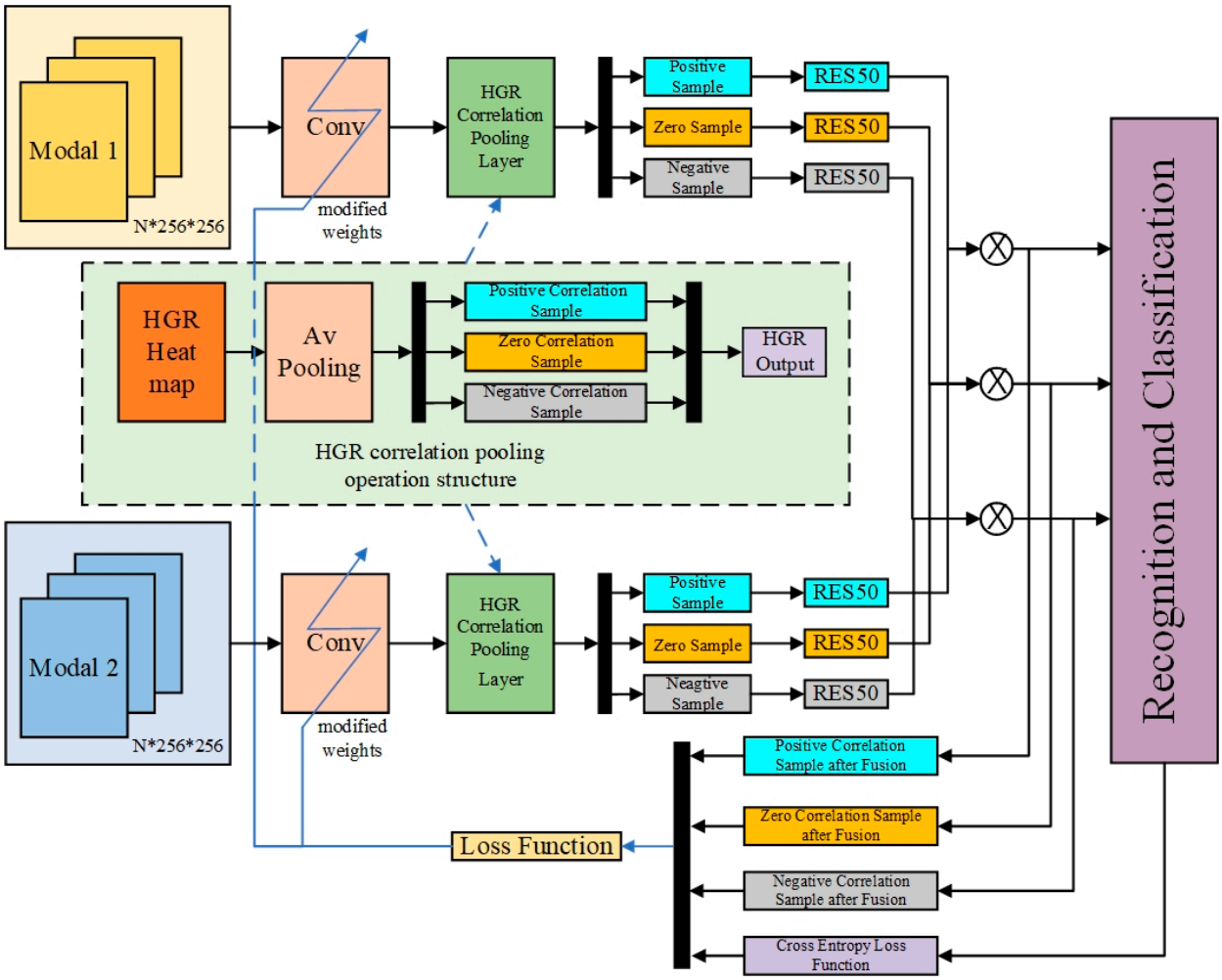


Figure 1. The overall framework of the proposed multimodal data fusion model.

3.3. Heatmap and HGR Correlation Pooling

The input set of multimodal images needs to be pre-aligned to generate heatmaps. X denotes the input data of modal 1 pixel matrix of size $n \times n$, and Y denotes the input data of modal 2 pixel matrix of size $n \times n$. The statistical matrices of modal 1 and modal 2 images are illustrated by empirical distributions and defined as follows:

For each pixel of modal 1 and modal 2 images:

$$X(p_s, x) = (\#of''p_s''in \quad x) = \sum_{i=1}^{n} 1\{x_i = p_s\} \tag{1}$$

$$Y(p_o, y) = (\#of''p_o''in \quad y) = \sum_{i=1}^{n} 1\{y_i = p_o\} \tag{2}$$

where p_s and p_o represent the pixel position in each modal 1 and modal 2 image, respectively, and # represents the number of pixels in modal 1 and modal 2 images. The expressions appear to be defining functions $X(p_s,x)$ and $Y(p_o,y)$, which count the occurrences of specific pixel positions p_s and p_o in two different modal images, respectively.

According to the definition of X and Y, the statistical matrix calculation process is as follows:

$$D_s = F_s(X, Y) \tag{3}$$

where $F_s(\cdot)$ is the statistical matrix calculation function and D_s is the statistical matrix. The calculation of the statistical matrix process is shown in Figure 2.

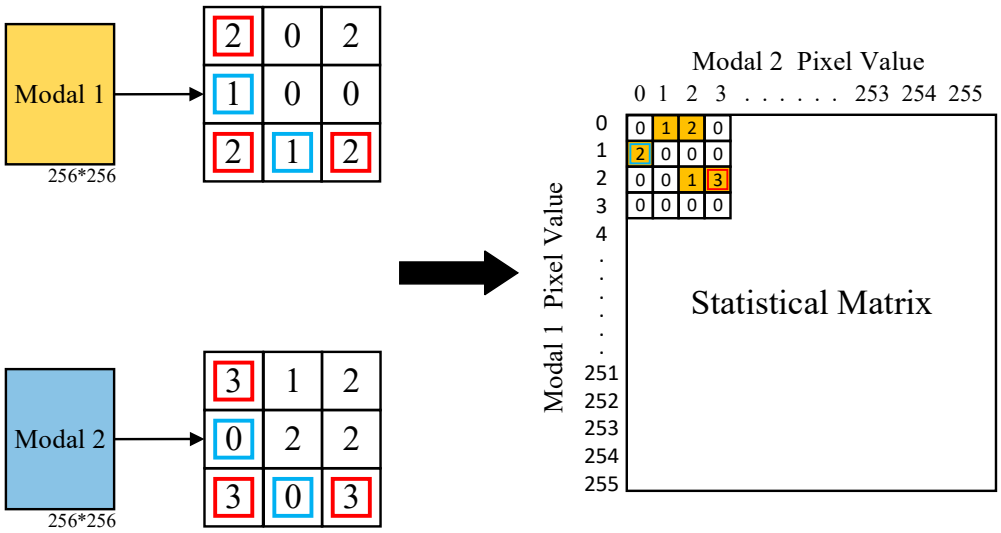


Figure 2. Statistical matrix calculation process.

Then, the pixel-level maximal nonlinear cross-correlation between sets is given by:

$$\rho(x,y) = \max_{\substack{E[f(x)]=0, E[g(y)]=0 \\ Var\{f(x)]=1, Var[g(y)]=1}} E[f(x)g(y)] \tag{4}$$

$$H = f(x) * g(y) \tag{5}$$

where $f(x)$ represents modal 1 in each pixel position, $g(y)$ represents modal 2 in each pixel position, and H represent the nonlinear correlation between the pixel points in modal 1 and modal 2 images.

According to the obtained statistical matrix, the HGR cross-correlation is calculated as follows:

$$F_{HGR}(D_s) = [f(x), g(y)]^T \tag{6}$$

where $F_{HGR}(\cdot)$ is the HGR cross-correlation calculation function, and $f(x)$ and $g(y)$ are projection vectors. Based on the obtained projection vector, the heatmap pixel matrix is calculated as follows:

$$D_{HM} = f(X_{ij}) - g(Y_{ij}), i, j \in \{1, 2, \cdots, n\} \tag{7}$$

where D_{HM} is a heatmap pixel matrix of size $n \times n$. The heatmap calculation process is shown in Figure 3.

Based on the heatmap pixel matrix obtained above, the average pooling is calculated as follows:

$$B = AP(D_{HM}) \tag{8}$$

where $AP(\cdot)$ is the average pooling function, and B is the HGR region cross-correlation matrix of size $(n-2) \times (n-2)$. Furthermore, the calculation of cross-correlation positive, zero, and negative sample matrices can be expressed as:

$$B^+ = \begin{cases} B_{kl}, B_{kl} \geq 0.4 \\ 0, B_{kl} < 0.4 \end{cases} \tag{9}$$

$$B^0 = \begin{cases} B_{kl}, |B_{kl}| < 0.4 \\ 0, |B_{kl}| \geq 0.4 \end{cases} \tag{10}$$

$$B^- = \begin{cases} B_{kl}, B_{kl} \leq 0.4 \\ 0, B_{kl} > 0.4 \end{cases} \tag{11}$$

where $k, l \in \{1, 2, \cdots, n-2\}$, and B^+, B^0, and B^- represent positive, zero, and negative cross-correlation sample matrices, respectively. The calculation of the HGR cross-correlation sample matrix is shown in Figure 4.

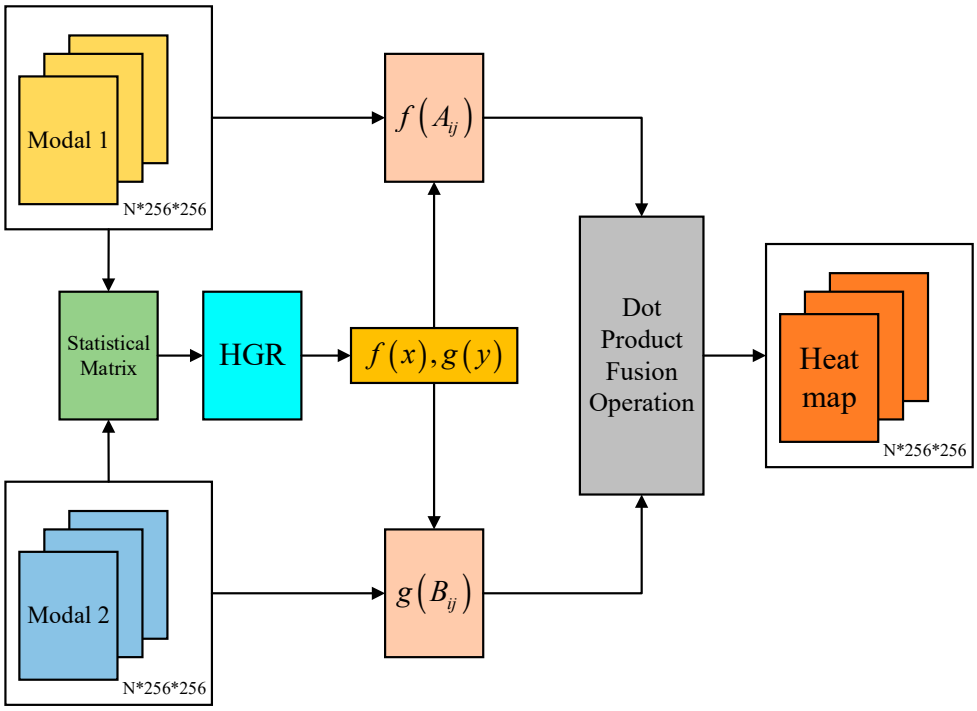


Figure 3. Heatmap calculation process.

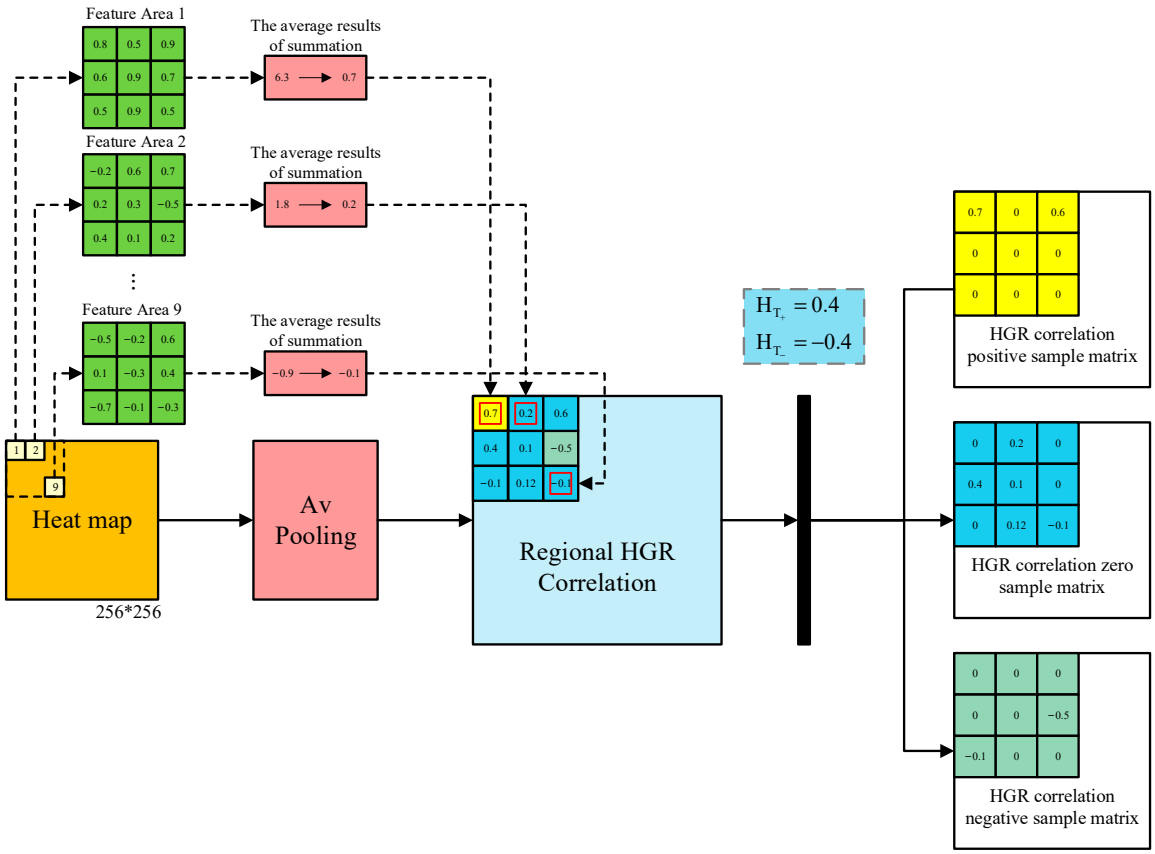


Figure 4. The calculation process of the HGR correlation sample matrix.

Meanwhile, A is defined as the image pixel position matrix, and $\overset{\max}{\otimes}$ is defined as the matrix maximum pooling dot product operation. Based on the cross-correlation positive, zero, and negative sample matrix obtained by the above calculation, the maximum pooling calculation process is given by:

$$\begin{cases} R^{+} = A \overset{\max}{\otimes} B^{+} \\ R^{0} = A \overset{\max}{\otimes} B^{0} \\ R^{-} = A \overset{\max}{\otimes} B^{-} \end{cases} \tag{12}$$

where R^{+}, R^{0}, and R^{-} are HGR cross-correlation positive, zero, and negative sample matrix maximum pooling results, respectively. The HGR cross-correlation pooling process is shown in Figure 5.

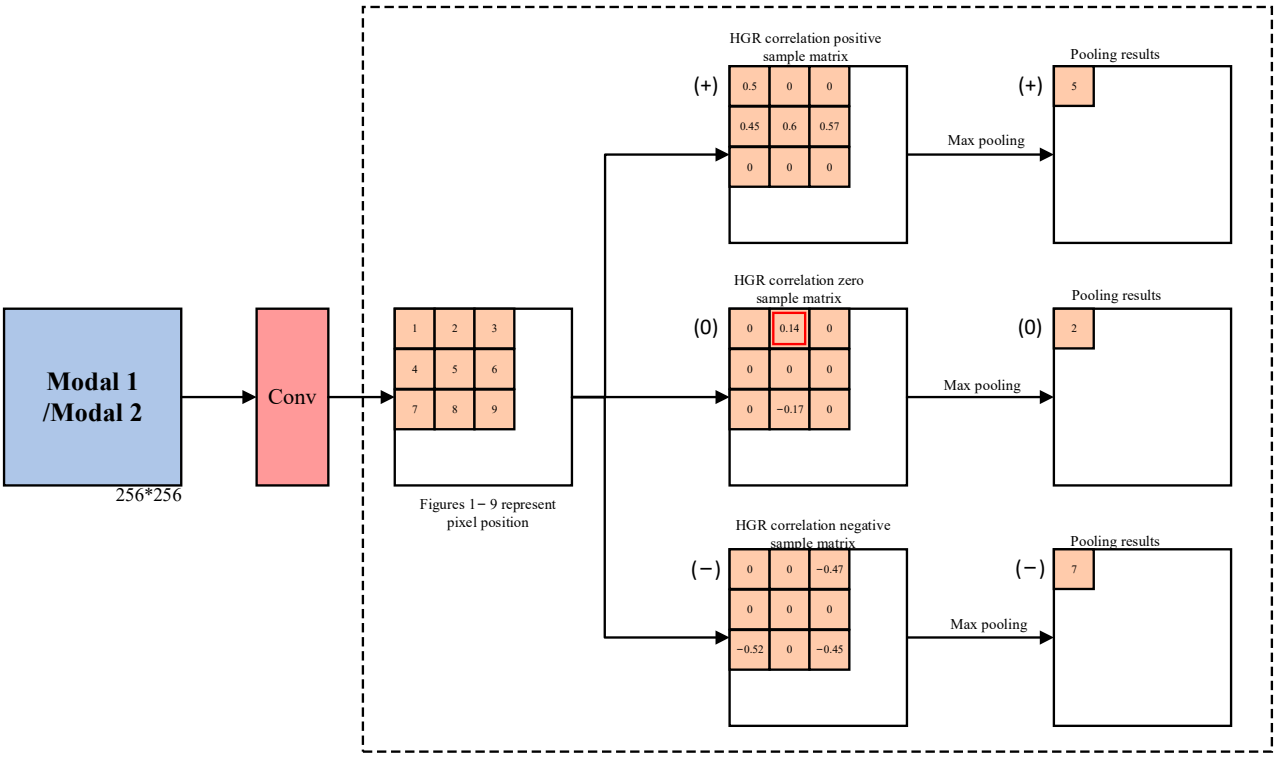


Figure 5. HGR correlation pooling process.

Finally, for the special HGR maximum correlation pooling layer, the key point is to obtain the transfer position based on the generated correlation matrix and transfer the values corresponding to the intermediate matrix in the pooling process. Due to dimensionality reduction, instead of relying on the common maximum value, only one value is passed on for 3×3, and correlation is used for transfer.

As shown in Figure 6, the corresponding modal 1/modal 2 features are transmitted through the special HGR pooling layer, which intervenes in the information transmission based on the HGR maximum correlation matrix instead of normal pooling methods without changing the value of the transmitted information. Meanwhile, the modal 1/modal 2 feature maps are divided into positive samples, zero samples, and negative samples for modal 1 and modal 2 data, respectively.

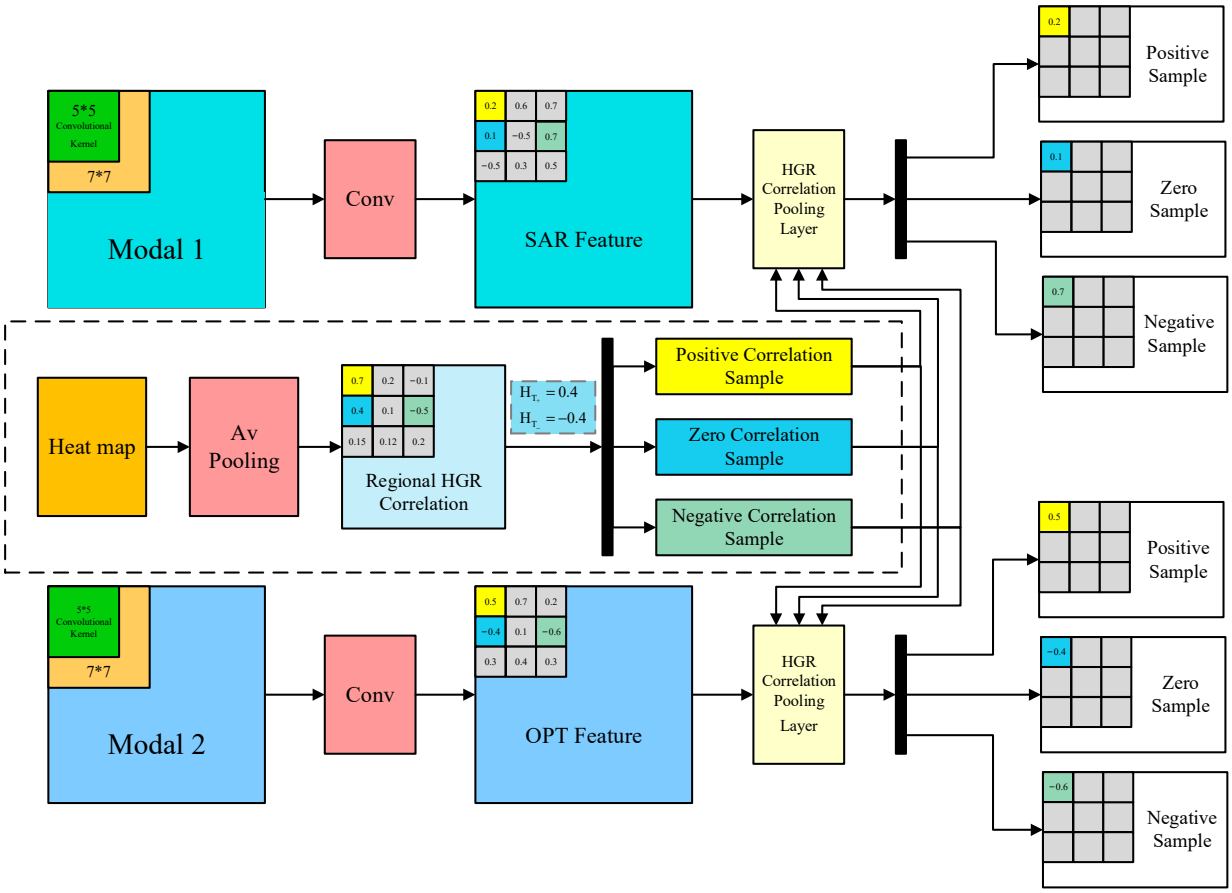


Figure 6. Sample division.

3.4. Learning Objective

The cross-entropy loss and the Soft-HGR loss with modified weights are taken as the loss for the whole network. Thus, the learning objective of the framework can be represented as:

$$L = L_{ce} + \alpha L_{Soft-HGR} \tag{13}$$

where L_{ce} represents the cross-entropy loss and $L_{soft\text{-}HGR}$ denotes the Soft-HGR loss. α is the penalty parameter to balance the cross-entropy loss L_{ce} and the Soft-HGR loss $L_{soft\text{-}HGR}$, which are designer-defined parameters ranging from 0 to 1.

The cross-entropy loss L_{ce} is used to measure the difference between the predicted result $\hat{y}_i$ and the ground-truth label y_i, and can be expressed as follows:

$$L_{ce} = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i)\right] \tag{14}$$

The Soft-HGR loss $L_{soft\text{-}HGR}$ [36] is utilized to maximize the correlation between multi-modal images, and can be represented as follows:

$$L_{Soft-HGR} = \alpha_1 L_{positive} + \alpha_2 L_{zero} + (1 - \alpha_1 - \alpha_2) L_{negative} \tag{15}$$

where α_1 and α_2 are weighting factors, which are designer-defined parameters ranging from 0 to 1. $L_{positive}$, L_{zero}, and $L_{negative}$ represent the loss for positive samples, zero samples, and negative samples, and their definitions are given below:

$$L_{positive} = - \sum_{\substack{s.t., \mathbb{E}(\mathbf{f}_P) = 0, cov(\mathbf{f}_P) = \mathbf{I} \\ \mathbb{E}(\mathbf{g}_P) = 0, cov(\mathbf{g}_P) = \mathbf{I}}} \mathbb{E}\left(\mathbf{f}_P^T(X)\mathbf{g}_P(Y)\right) - \frac{1}{2} tr(cov(\mathbf{f}_P(X))cov(\mathbf{g}_P(Y))) \quad (16)$$

$$L_{zero} = - \sum_{\substack{s.t., \mathbb{E}(\mathbf{f}_Z) = 0, cov(\mathbf{f}_Z) = \mathbf{I} \\ \mathbb{E}(\mathbf{g}_Z) = 0, cov(\mathbf{g}_Z) = \mathbf{I}}} \mathbb{E}\left(\mathbf{f}_Z^T(X)\mathbf{g}_Z(Y)\right) - \frac{1}{2} tr(cov(\mathbf{f}_Z(X))cov(\mathbf{g}_Z(Y))) \quad (17)$$

$$L_{negative} = - \sum_{\substack{s.t., \mathbb{E}(\mathbf{f}_N) = 0, cov(\mathbf{f}_N) = \mathbf{I} \\ \mathbb{E}(\mathbf{g}_N) = 0, cov(\mathbf{g}_N) = \mathbf{I}}} \mathbb{E}\left(\mathbf{f}_N^T(X)\mathbf{g}_N(Y)\right) - \frac{1}{2} tr(cov(\mathbf{f}_N(X))cov(\mathbf{g}_N(Y))) \quad (18)$$

where $f_p(X)$ and $g_p(Y)$ represent a pair of positive samples. Similarly, $f_Z(X)$ and $g_Z(Y)$ represent a pair of zero samples, and $f_N(X)$ and $g_N(Y)$ represent a pair of negative samples. As a supplement, the expectations and covariance are approximated through the sample mean and sample covariance.

4. Experiments and Analysis

4.1. Dataset

The ship recognition experiments were conducted on the QXS-SROPT dataset, and the land cover classification experiments were conducted on the Houston 2013 and MUUFL datasets to verify the effectiveness of the proposed model and test the improvement in remote sensing data recognition and classification when maximizing the utilization of multimodal information.

QXS-SAROPT [45] contains 20,000 pairs of optical and SAR images collected from Google Earth remote sensing optical images and GaoFen-3 high-resolution spotlight images. The size of each image is 256 × 256 to fit the neural network with a resolution of 1 m. This dataset covers San Diego, Shanghai, and Qingdao.

The HSI-LiDAR Houston2013 dataset [46], provided for IEEE GRSS DFC2013, consists of imagery collected by the ITRES CASI-1500 imaging sensor. This imagery encompasses the University of Houston campus and its adjacent rural areas in Texas, USA. This dataset is widely used for land cover classification.

The HSI-LiDAR MUUFL dataset was constructed over the campus of the University of Southern Mississippi using the Reflective Optics System Imaging Spectrometer (ROSIS) sensor [47,48]. This dataset contains HSI and LiDAR data and is widely used for land cover classification.

4.2. Data Preprocessing and Experimental Setup

Since the image pairs in the QXS-SROPT dataset do not contain labels and are not aligned, in this study, manual alignment operations were performed on 131 image pairs. The proposed model learns the corresponding pixel correlation from SAR and optical image pairs, and calculates the maximum pixel-level HGR cross-correlation between SAR-optical datasets and the corresponding projection vectors $f(x)$, $g(y)$ to generate the heatmap and HGR cross-correlation pooling matrix modules. The pixel-level HGR has a certain potential for improving network multimodal information learning.

In the Houston 2013 dataset [46], the HSI image contains 349 × 1905 pixels and features 144 spectral channels at a spectral resolution of 10 nm, spanning a range from 364 to 1046 nm. Meanwhile, LiDAR data for a single band provide elevation information for the same image area. The study scene encompasses 15 distinct land cover and land use categories. This dataset contains 2832 training samples and 12,197 test samples, as listed in Table 1 [49].

Table 1. The Houston2013 dataset with 2832 training samples and 12,197 testing samples.

No	Class Name	Training Set	Testing Set
1	Healthy grass	198	1053
2	Stressed grass	190	1064
3	Synthetic grass	192	505
4	Trees	188	1056
5	Soil	186	1056
6	Water	182	143
7	Residential	196	1072
8	Commercial	191	1053
9	Road	193	1059
10	Highway	191	1036
11	Railway	181	1054
12	Parking lot 1	192	1041
13	Parking lot 2	184	285
14	Tennis court	181	247
15	Running track	187	473
	Total	2832	12,197
	Percentage	18.84%	81.16%

In the MUUFL dataset [48], the HSI image contains 325 × 220 pixels, covering 72 spectral bands. The LiDAR imagery incorporates elevation data across two grids. Due to noise considerations, 8 initial and final bands were discarded, and 64 bands remained. The data encompass 11 urban land cover classes, comprising 53,687 ground truth pixels. Table 2 presents the distribution of 5% samples randomly extracted from each category.

Table 2. The MUUFL Gulfport dataset with 2669 training samples and 51,018 testing samples.

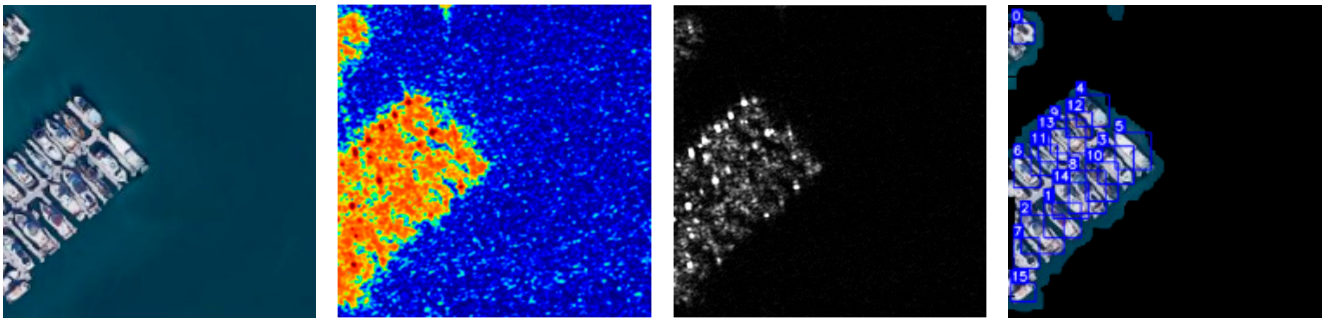
No	Class Name	Training Set	Testing Set
1	Trees	1166	22,080
2	Grass-Pure	222	4048
3	Grass-Groundsurface	356	6526
4	Dirt-and-Sand	86	1740
5	Road-Materials	315	6372
6	Water	30	436
7	Buildings'-Shadow	93	2140
8	Buildings	302	5938
9	Sidewalk	74	1311
10	Yellow-Curb	9	174
11	ClothPanels	16	253
	Total	2669	51,018
	Percentage	4.97%	95.03%

The proposed model was trained using the Lion optimizer for 500 epochs and a batch size of 32 with an initial learning rate of 0.0001. After 30 epochs, the learning rate gradually decreased by $1 \times 10^{-0.01}$ times in each epoch. All experiments were conducted on a computer equipped with an Intel(R) Xeon(R) Gold 6133 CPU @ 2.50 GHz and an NVIDIA GeForce RTX3090 GPU (NVIDIA, Santa Clara, CA, USA) with 24 G memory, 64-bit Ubuntu 20.04 operating system, CUDA 12.2, and cuDNN 8.8. The source code was implemented using PyTorch 2.1.1 and Python 3.9.16.

4.3. Ship Recognition Experiment

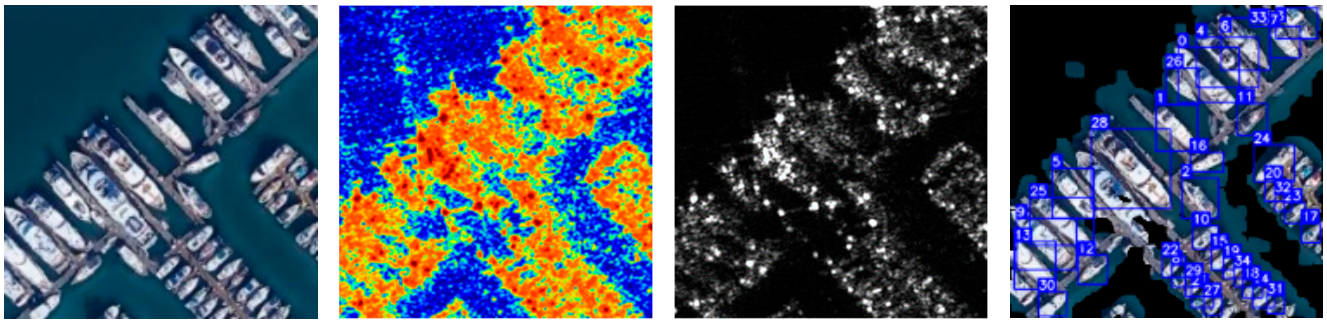
To verify the performance of the proposed HGRPool method, a series of experiments were conducted on the QXS-SAROPT dataset (100 training samples with 4358 instances, 31 testing samples with 1385 instances) to perform ship feature recognition on SAR–optical image pairs. The results of three experiments are illustrated in Figures 7–9, where (a) illustrates

the optical image, (b) shows the heatmap, (c) displays the SAR image, and (d) depicts the ship recognition results. The results from these figures demonstrate that the HGR-Pool method effectively identifies different ships, achieving commendable recognition performance and effectively distinguishing water bodies.



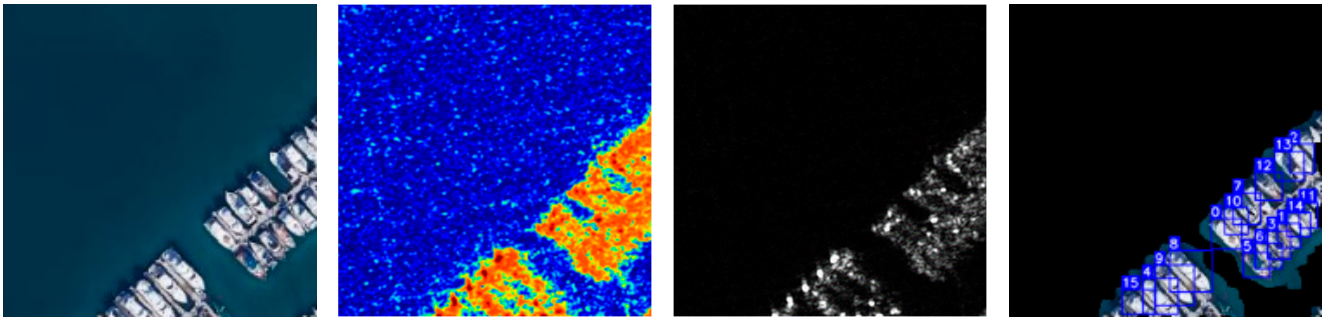
(**a**) (**b**) (**c**) (**d**)

Figure 7. Ship recognition experiment 1: (**a**) optical image; (**b**) heatmap; (**c**) SAR image; (**d**) ship recognition results.



(**a**) (**b**) (**c**) (**d**)

Figure 8. Ship recognition experiment 2: (**a**) optical image; (**b**) heatmap; (**c**) SAR image; (**d**) ship recognition results.



(**a**) (**b**) (**c**) (**d**)

Figure 9. Ship recognition experiment 3: (**a**) optical image; (**b**) heatmap; (**c**) SAR image; (**d**) ship recognition results.

The proposed HGRPool method was compared with the BNN method proposed by Bao et al. [50]. Table 3 lists the values of four commonly used indicators, namely precision (P), recall (R), F1-score (F1), and accuracy (Acc) for the recognition results. Meanwhile, under the same experimental conditions, comparative experiments were conducted with other existing methods, including MoCo-BNN [51], CCR-Net [2], and MFT [52]. The experimental data are presented in Table 3. The precision, recall, F1-score, and accuracy of the proposed method reached 0.908, 0.988, 0.946, and 0.898, respectively. The proposed

method achieved better results than existing methods, especially in terms of accuracy. The results suggest that the proposed method has greater recognition stability and accuracy and higher localization accuracy.

Table 3. Ship identification experimental results (best results are bolded).

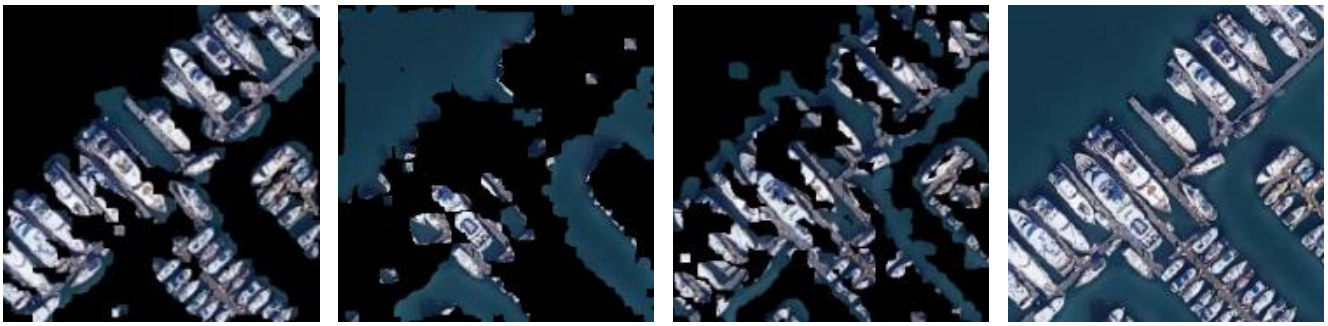
Model	Accuracy (Acc)	Precision (P)	Recall (R)	F1-Score
NP-BNN + ResNet50	0.831	0.750	0.990	0.853
NP-BNN + Darknet53	0.826	0.761	0.980	0.857
IP-BNN + ResNet50	0.829	0.748	0.993	0.853
IP-BNN + Darknet53	0.828	0.746	0.995	0.853
MoCo-BNN + ResNet50	0.873	0.808	0.995	0.892
MoCo-BNN + Darknet53	0.871	0.809	**0.997**	0.893
CCR-Net	0.854	0.883	0.963	0.894
MFT	0.876	0.892	0.980	0.934
HGRPool (ours)	**0.898**	**0.908**	0.988	**0.946**

4.4. Information Traceability Experiment

To demonstrate information integrity throughout the processing stages, an information traceability experiment was conducted using images in three distinct forms: positive, negative, and zero. The experiment aimed to retrieve the original images based on these three forms. This part of the experiment involved six sets of tests, three with optical images and three with SAR images. The results are illustrated in Figures 10–15. In these figures, (a) represents the positive image, (b) the negative image, (c) the zero image, and (d) the target traceability result image. From Figures 10–15, it can be observed that the traced images are consistent with the original images, with no information loss. This indicates that the proposed algorithm maintains information integrity throughout the processing stages, ensuring that no information is lost.

(**a**) (**b**) (**c**) (**d**)

Figure 10. Optical information traceability experiment 1: (**a**) optical positive image; (**b**) optical negative image; (**c**) optical zero image; (**d**) optical original image.

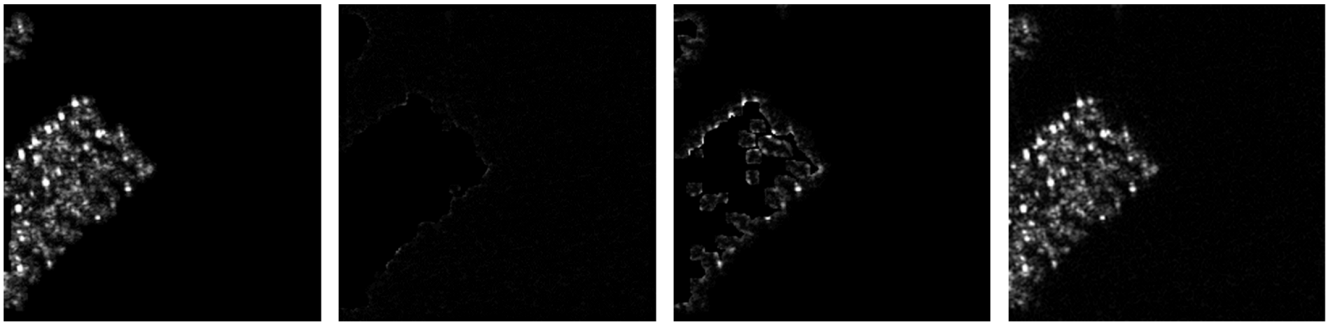


(**a**) (**b**) (**c**) (**d**)

Figure 11. Optical information traceability experiment 2: (**a**) optical positive image; (**b**) optical negative image; (**c**) optical zero image; (**d**) optical original image.

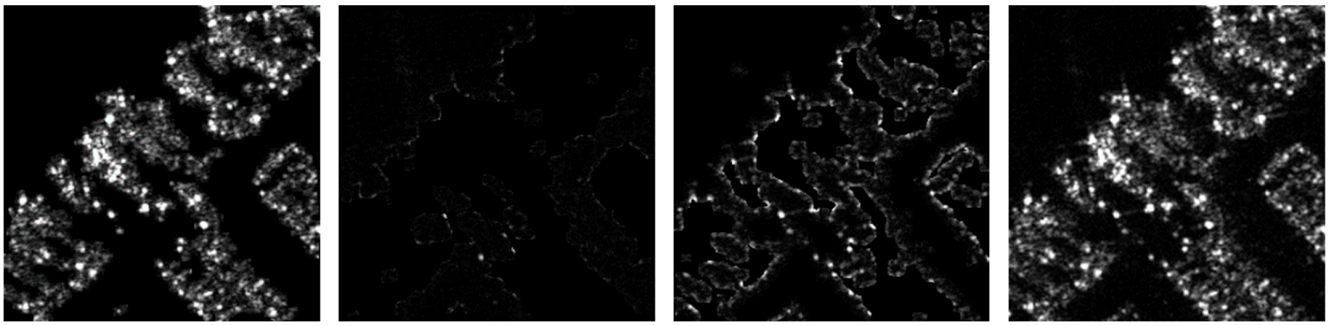
(**a**) (**b**) (**c**) (**d**)

Figure 12. Optical information traceability experiment 3: (**a**) optical positive image; (**b**) optical negative image; (**c**) optical zero image; (**d**) optical original image.



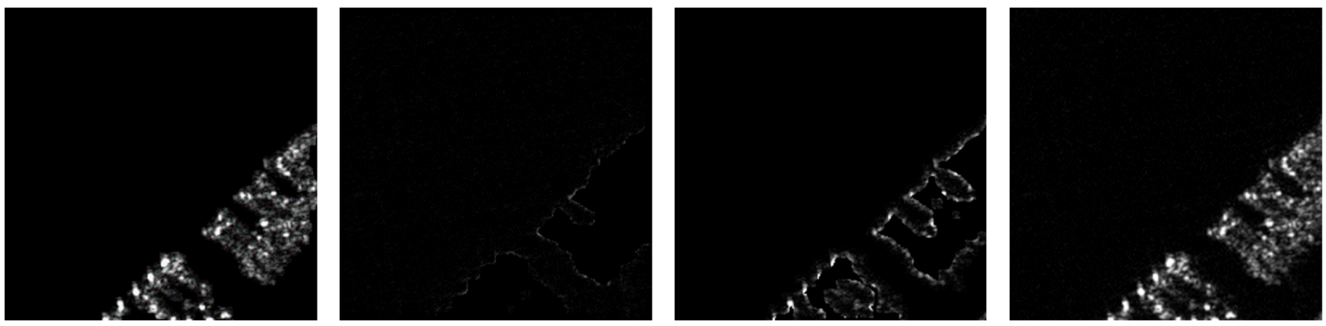
(**a**) (**b**) (**c**) (**d**)

Figure 13. SAR information traceability experiment 1: (**a**) SAR positive image; (**b**) SAR negative image; (**c**) SAR zero image; (**d**) SAR original image.



(**a**) (**b**) (**c**) (**d**)

Figure 14. SAR information traceability experiment 2: (**a**) SAR positive image; (**b**) SAR negative image; (**c**) SAR zero image; (**d**) SAR original image.



(**a**) (**b**) (**c**) (**d**)

Figure 15. SAR information traceability experiment 3: (**a**) SAR positive image; (**b**) SAR negative image; (**c**) SAR zero image; (**d**) SAR original image.

4.5. Land Cover Classification Experiment on the Houston 2013 Dataset

To validate the generalizability of the method proposed in this paper, land cover classification experiments were conducted on the Houston 2013 dataset. Our method was compared with traditional machine learning algorithms and state-of-the-art methods in the field of deep learning, including CCF [53], CoSpace [54], Co-CNN [55], FusAT-Net [56], ViT [57], S2FL [49], Spectral-Former [58], CCR-Net [2], MFT [52], and DIMNet [59]. The specific results are shown in Figure 16, where (a) displays the DSM of LiDAR data, (b) shows the heatmap, (c) represents the three-band color composite for HSI spectral information, (d) shows the train ground-truth map, (e) shows the test ground-truth map, and (f) illustrates the classification results, with good contrast post-reconstruction. The values of three universal indicators, namely overall accuracy (OA), class accuracy (AA), and Kappa coefficient, are presented in Table 4 for comparison, where the top outcomes are highlighted in bold. It is evident that our method outperforms the others in terms of OA (92.23%), AA (93.55%), and Kappa coefficient (0.9157). It surpasses other methods in eight categories (stressed grass, synthetic grass, water, residential, road, parking lot 1, tennis court, and running track), especially achieving the highest accuracy of 100% in the four categories of synthetic grass, tennis court, water, and running track. Even in the remaining seven categories, our method provides commendable results. Therefore, statistically, our method exhibits superior performance compared to all other models. This suggests that our method is general and universally applicable, and is thus a reliable model.

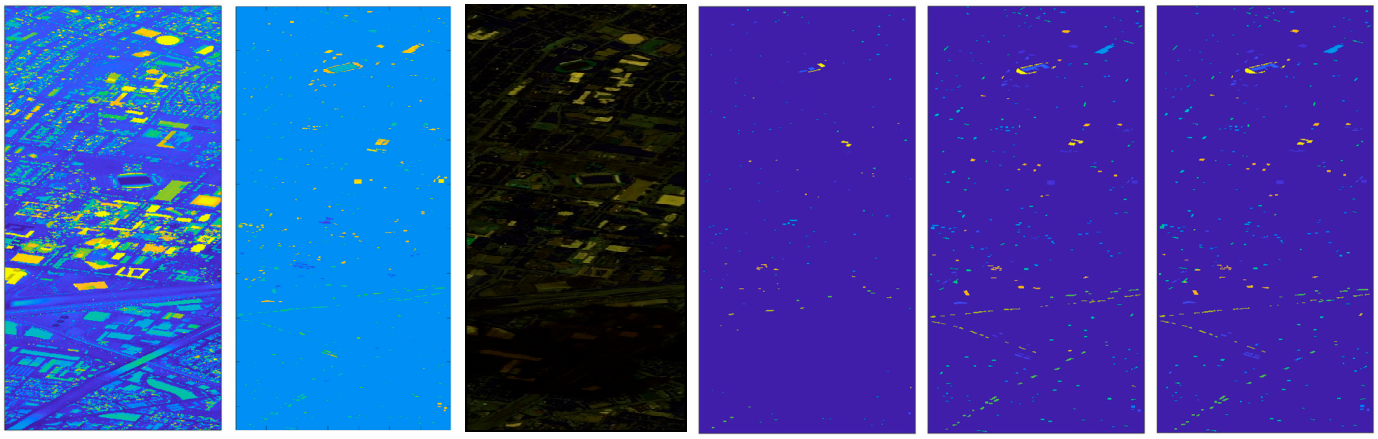


Figure 16. Houston 2013 dataset: (**a**) DSM obtained from LiDAR; (**b**) heatmap; (**c**) three-band color composite for HSI images (bands 32, 64, 128); (**d**) train ground-truth map; (**e**) test ground-truth map; (**f**) classification map.

Table 4. Comparison of various methods on the Houston 2013 dataset (best results are bolded).

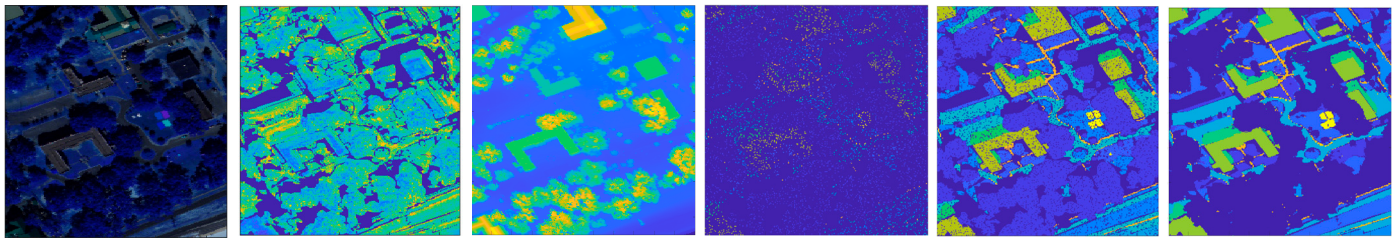
Class	CCF	CoSpace	Co-CNN	FusAt-Net	ViT	S2FL	Spectral-Former	CCR-Net	MFT	DIMNet	HGRPool (Ours)
OA (%)	83.46	82.14	87.23	88.69	85.05	85.07	86.14	88.15	89.15	91.47	**92.23**
AA (%)	85.95	84.54	88.82	90.29	86.83	86.11	87.48	89.82	90.56	92.48	**93.55**
Kappa coefficient	0.8214	0.8062	0.8619	0.8772	0.8384	0.8378	0.8497	0.8719	0.8822	0.9077	**0.9157**
Healthy grass	83.10	81.96	83.1	**96.87**	82.59	80.06	83.48	83	82.72	83.00	83.00
Stressed grass	83.93	83.27	84.87	82.42	82.33	84.49	95.58	84.87	85.09	84.68	**98.87**
Synthetic grass	**100.00**	**100.00**	99.8	**100.00**	97.43	98.02	99.60	**100.00**	98.55	99.01	**100.00**
Trees	92.42	94.22	92.42	91.95	92.93	87.31	**99.15**	92.14	95.99	91.38	98.58
Soil	98.77	99.34	99.24	97.92	99.84	**100.00**	97.44	99.81	99.78	99.62	92.90
Water	**99.30**	**99.30**	95.8	90.91	84.15	83.22	95.10	95.8	97.20	95.10	**100.00**
Residential	84.42	81.44	95.24	92.91	87.84	73.32	88.99	95.34	86.32	92.91	**98.50**
Commercial	52.90	66.1	81.86	**89.46**	79.93	74.84	73.31	81.39	81.16	87.27	79.58
Road	76.02	69.97	85.08	82.06	82.94	78.38	71.86	84.14	87.76	88.01	**88.22**
Highway	67.18	48.94	61.1	66.60	52.93	83.30	87.93	63.22	74.71	**93.82**	86.50

Table 4. *Cont.*

Class	CCF	CoSpace	Co-CNN	FusAt-Net	ViT	S2FL	Spectral-Former	CCR-Net	MFT	DIMNet	HGRPool (Ours)
Railway	84.44	88.61	83.87	80.36	80.99	81.69	80.36	90.32	**93.71**	88.80	91.46
Parking lot 1	92.80	88.57	91.26	92.41	91.07	95.10	70.70	93.08	96.16	96.54	**97.50**
Parking lot 2	76.49	68.07	88.77	**92.63**	87.84	72.63	71.23	88.42	92.51	90.53	90.88
Tennis court	99.60	**100.00**	91.09	**100.00**	**100.00**	**100.00**	98.79	96.36	**100.00**	96.76	**100.00**
Running track	97.89	98.31	98.73	97.89	**99.65**	99.37	98.73	99.37	86.82	99.79	**100.00**

4.6. Land Cover Classification Experiment on the MUUFL Dataset

To validate the generalizability of the proposed method, the land cover classification experiments were conducted on the HSI-LiDAR MUUFL dataset. Our method was compared with traditional machine learning algorithms and state-of-the-art methods in the field of deep learning, including CCF, CoSpace, Co-CNN, FusAT-Net, ViT [57], S2FL, Spectral-Former, CCR-Net, and MFT. The specific results are illustrated in Figure 17, where (a) displays the three-band color composite for HSI spectral information, (b) shows the heatmap, (c) represents the LiDAR image, (d) shows the train ground-truth map, (e) shows the test ground-truth map, and (f) illustrates the classification results. The values of three universal indicators, namely OA, AA, and Kappa coefficient, are presented in Table 5, with the top outcomes being highlighted in bold. It is evident that our method outperforms the others in terms of OA (94.99%), AA (88.13%), and Kappa coefficient (0.9339). Our method surpasses other methods in five categories (grass-pure, water, buildings'-shadow, buildings, and sidewalk). Even in the remaining six categories, our method obtains commendable results. Therefore, statistically, our method exhibits superior performance compared to all other models. This shows that our method is general and universally applicable, and is thus a reliable model.



(**a**) (**b**) (**c**) (**d**) (**e**) (**f**)

Figure 17. MUUFL dataset: (**a**) three-band color composite for HSI images (bands 16, 32, 64); (**b**) heatmap (**c**) LiDAR image; (**d**) train ground-truth map; (**e**) test ground-truth map; (**f**) classification map.

Table 5. Comparison of various methods on the HSI-LiDAR MUUFL dataset (best results are bolded).

Class	CCF	CoSpace	Co-CNN	FusAt-Net	ViT	S2FL	Spectral-Former	CCR-Net	MFT	HGRPool (Ours)
OA(%)	88.22	87.55	90.93	91.48	92.15	72.49	88.25	90.39	94.34	**94.99**
AA(%)	71.76	71.63	77.18	78.58	78.50	79.23	68.47	76.31	81.48	**88.13**
Kappa	0.8441	0.8353	0.8822	0.8865	0.8956	0.6581	0.8440	0.8603	0.9251	**0.9339**
Trees	96.50	95.89	**98.90**	98.10	97.85	72.40	97.30	96.78	97.90	97.98
Grass-Pure	77.17	66.65	78.60	71.66	76.06	75.97	69.35	83.99	92.11	**92.45**
Grass-Groundsurface	74.80	85.24	90.66	87.65	87.58	54.72	78.48	84.16	**91.80**	89.86
Dirt-and-Sand	91.94	68.45	90.60	86.42	92.05	82.20	82.63	**93.05**	91.59	91.81
Road-Materials	93.45	94.52	**96.90**	95.09	94.73	71.26	87.91	91.37	95.60	95.13
Water	95.05	96.10	75.98	90.73	82.02	94.42	58.77	81.88	88.19	**99.28**
Buildings'-Shadow	79.82	84.91	73.54	74.27	87.11	77.34	85.87	76.54	90.27	**93.25**
Buildings	98.21	91.19	96.66	97.55	97.60	86.19	95.60	94.58	97.26	**97.83**
Sidewalk	0.52	9.69	64.93	60.44	57.83	59.21	53.52	43.02	61.35	**78.14**
Yellow-Curb	0.00	0.00	19.47	09.39	31.99	**98.91**	08.43	00.00	17.43	46.25
ClothPanels	81.89	95.26	62.76	93.02	58.72	**98.88**	35.29	94.70	72.79	87.45

5. Discussion

5.1. Ablation Experiment

The ablation experiments were conducted on the QXS-SROPT dataset, the Houston 2013 dataset, and the MUUFL dataset to evaluate the proposed HGR correlation pooling fusion framework.

In the ablation experiments on QXS-SROPT datasets, the performance was observed when partially using the HGRPool, i.e., using the HGRPool for positive and negative samples or positive and zero samples, as well as when completely omitting it. The results of ablation experiments on the QXS-SROPT dataset are presented in Table 6. The proposed model demonstrates optimal performance in ship recognition experiments on the QXS-SROPT dataset when it incorporates all components, i.e., when fully utilizing the HGRPool. Meanwhile, there is a notable decline in ship recognition accuracy as the HGRPool component is partially employed or entirely excluded.

Table 6. Ablation study by removing different modules on the QXS-SROPT dataset (best results are bolded).

Methods	Accuracy (Acc)	Precision (P)	Recall (R)	F1-Score
Without HGRPool	0.722	0.803	0.877	0.838
Partially using HGRPool (positive/zero sample)	0.789	0.834	0.937	0.883
Partially using HGRPool (positive /negative sample)	0.810	0.849	0.947	0.895
HGRPool	**0.898**	**0.908**	**0.988**	**0.946**

In the ablation experiments on the Houston 2013 and MUUFL datasets, the performance was observed when partially using HGRPool and completely omitting it. The results of ablation experiments on the Houston 2013 and MUUFL datasets are shown in Table 7. Similarly, the proposed model demonstrates optimal performance in land cover classification experiments on the Houston 2013 and MUUFL datasets when it incorporates all components, i.e., when fully utilizing the HGRPool. Due to partial use or complete exclusion of the HGRPool component, there is a significant decrease in land cover classification accuracy.

Table 7. Ablation study by removing different modules on the Houston 2013 and MUUFL datasets (best results are bolded).

Methods	Houston 2013 Dataset			MUUFL Dataset		
	OA (%)	AA (%)	Kappa	OA (%)	AA (%)	Kappa
Without HGRPool	89.64	90.26	0.8851	92.72	80.94	0.9040
Partially using HGRPool (positive/zero sample)	90.20	91.05	0.8937	93.24	84.41	0.9106
Partially using HGRPool (positive/negative sample)	90.81	91.46	0.9013	93.79	85.07	0.9180
HGRPool	**92.23**	**93.55**	**0.9157**	**94.99**	**88.13**	**0.9339**

5.2. Analyzing the Effect of Experiments

The comparative experimental results confirm the precision and accuracy of our method. Compared with various advanced matching networks, this method not only achieves accurate and stable matching in ship recognition but also has particularly obvious advantages in land cover classification. By integrating a feature fusion method with an HGR correlation algorithm to separate information based on intrinsic correlation into different classification channels while maintaining information integrity, this model achieves information separation and maximizes the utilization of multimodal data, thereby improving the precision and accuracy of target recognition and classification.

From Table 6, it can be seen that the proposed model demonstrates optimal performance (with precision, recall, F1-score, and accuracy of 0.898, 0.908, 0.988, and 0.946,

respectively) in ship recognition experiments on the QXS-SROPT dataset when it incorporated all components, i.e., fully using the HGRPool. There is a notable decline in ship recognition accuracy when the HGRPool component is partially used or entirely excluded. When partially using HGRPool (positive/negative sample), only recall (R) is still as high as 0.947. The results of accuracy, precision, and F1-score drop to 0.810, 0.849, and 0.895, respectively. The result of partially using HGRPool (positive/zero sample) is slightly worse than that of partially using HGRPool (positive/negative sample). Additionally, the result corresponding to without HGRPool is the worst, with accuracy, precision, recall, and F1-score being only 0.722, 0.803, 0.877, and 0.838, respectively.

Meanwhile, it can be deduced from Table 7 that there is a notable decline in land cover classification accuracy as the HGRPool component is partially employed or entirely excluded. The result of partially using HGRPool (positive/negative sample) in OA, AA, and Kappa drop to 90.81%, 91.46%, and 0.9013 on the Houston 2013 dataset and 93.79%, 85.07%, and 0.9180 on the MUUFL dataset. The result of partially using HGRPool (positive/zero sample) is slightly worse than that of partially using HGRPool (positive/negative sample). The result corresponding to without HGRPool is the worst, with OA, AA, and Kappa being only 89.64%, 90.26%, and 0.8851 on the Houston 2013 dataset and 92.72%, 80.94%, and 0.9040 on the MUUFL dataset. It can be concluded that the proposed HGR correlation pool fusion framework is effective and helps to improve accuracy.

6. Conclusions

The fusion of multimodal images has always been a research hotspot in the field of remote sensing. To address the issues of low recognition and classification accuracy and difficulty in integrating multimodal features in existing remote sensing data recognition and classification methods, this paper proposes a multimodal remote sensing data recognition and classification model based on a heatmap and HGR cross-correlation pooling fusion operation. Then, an HGR cross-correlation pooling fusion algorithm is proposed by combining the feature fusion method with the HGR cross-correlation algorithm. The model first calculates the statistical matrix through multimodal image pairs, extracts multimodal image features using convolutional layers, and then computes the heatmap from these features. Subsequently, by performing HGR cross-correlation pooling operations, the model can separate information with intrinsic relevance into respective classification channels, achieving dimensionality reduction of multimodal image features. In this approach, less feature data are used to represent the image area information of multimodal images while maintaining the original image information, thereby avoiding the problem of feature dimension explosion. Finally, point multiplication fusion is performed on the dimensionality-reduced feature samples, which are then input into the recognition and classification network for training to achieve recognition and classification of remote sensing data. This method maximizes the utilization of multimodal information, enhances the feature learning capability of multimodal images, improves the performance of specific interpretation tasks related to multimodal image fusion, and achieves classification and efficient utilization of information with different degrees of relevance. By conducting ship recognition experiments on the QXS-SROPT dataset and land cover classification experiments on the Houston 2013 and MUUFL datasets, it was fully verified that the proposed method outperforms other state-of-the-art remote sensing data recognition and classification methods. In future research, efforts will be made to further enhance recognition and classification accuracy and expand the application scope of this method to encompass more complex scenes and additional modalities. Further investigation will also be carried out of adaptive tuning of the parameters to achieve the best recognition and classification effects.

Author Contributions: Author Contributions: Conceptualization, H.Z. and S.-L.H.; methodology, H.Z., S.-L.H. and E.E.K.; validation, H.Z. and S.-L.H.; investigation, H.Z., S.-L.H. and E.E.K.; data curation, H.Z. and S.-L.H.; writing—original draft, H.Z.; writing—review and editing, S.-L.H. and E.E.K.; supervision, S.-L.H. and E.E.K.; funding acquisition, S.-L.H. and E.E.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Key R&D Program of China under Grant 2021 YFA0715202, Shenzhen Key Laboratory of Ubiquitous Data Enabling (Grant No. ZDSYS20220527171406015) and the Shenzhen Science and Technology Program under Grant KQTD20170810150821146 and Grant JCYJ20220530143002005.

Data Availability Statement: The Houston 2013 dataset used in this study is available at https://hyperspectral.ee.uh.edu/?page_id=1075 (accessed on 28 August 2023); the MUUFL dataset is available from https://github.com/GatorSense/MUUFLGulfport/ (accessed on 19 October 2023); the QXS-SAROPT dataset under open access license CCBY is available at https://github.com/yaoxu008/QXS-SAROPT (accessed on 27 June 2023).

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

References

1. Ghamisi, P.; Rasti, B.; Yokoya, N.; Wang, Q.; Hofle, B.; Bruzzone, L.; Bovolo, F.; Chi, M.; Anders, K.; Gloaguen, R.; et al. Multisource and multitem-poral data fusion in remote sensing: A comprehensive review of the state of the art. *IEEE Geosci. Remote Sens. Mag.* **2019**, *7*, 6–39. [CrossRef]
2. Wu, X.; Hong, D.F.; Chanussot, J. Convolutional Neural Networks for Multimodal Remote Sensing Data Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–10. [CrossRef]
3. Hong, D.F.; Gao, L.R.; Yokoya, N.; Yao, J.; Chanussot, J.; Du, Q.; Zhang, B. More Diverse Means Better: Multimodal Deep Learning Meets Remote-Sensing Imagery Classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 4340–4354. [CrossRef]
4. Li, X.; Lu, G.; Yan, J.; Zhang, Z. A survey of dimensional emotion prediction by multimodal cues. *Acta Autom. Sin.* **2018**, *44*, 2142–2159.
5. Wang, C.; Li, Z.; Sarpong, B. Multimodal adaptive identity-recognition algorithm fused with gait perception. *Big Data Min. Anal.* **2021**, *4*, 10. [CrossRef]
6. Zhou, W.J.; Jin, J.H.; Lei, J.S.; Hwang, J.N. CEGFNet: Common Extraction and Gate Fusion Network for Scene Parsing of Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–10. [CrossRef]
7. Asghar, M.; Khan, M.; Fawad; Amin, Y.; Rizwan, M.; Rahman, M.; Mirjavadi, S. EEG-Based multi-modal emotion recognition using bag of deep features: An optimal feature selection approach. *Sensors* **2019**, *19*, 5218. [CrossRef] [PubMed]
8. Yang, R.; Wang, S.; Sun, Y.Z.; Zhang, H.; Liao, Y.; Gu, Y.; Hou, B.; Jiao, L.C. Multimodal Fusion Remote Sensing Image–Audio Retrieval. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 6220–6235. [CrossRef]
9. Li, H.; Wu, X. DenseFuse: A fusion approach to infrared and visible images. *IEEE Trans. Image Process.* **2019**, *28*, 2614–2623. [CrossRef]
10. Yang, B.; Zhong, J.; Li, Y.; Chen, Z. Multi-Focus image fusion and super-resolution with convolutional neural network. *Int. J. Wavelets Multiresolution Inf. Process.* **2017**, *15*, 1750037. [CrossRef]
11. Zhang, X. Benchmarking and comparing multi-exposure image fusion algorithms. *Inf. Fusion.* **2021**, *74*, 111–131. [CrossRef]
12. Song, X.; Wu, X.; Li, H. MSDNet for medical image fusion. In Proceedings of the International Conference on Image and Graphic, Nanjing, China, 22–24 September 2019; pp. 278–288.
13. Cao, X.F.; Gao, S.; Chen, L.C.; Wang, Y. Ship recognition method combined with image segmentation and deep learning feature extraction in video surveillance. *Multimedia Tools Appl.* **2020**, *79*, 9177–9192. [CrossRef]
14. Wang, C.; Pei, J.; Luo, S.; Huo, W.; Huang, Y.; Zhang, Y.; Yang, J. SAR ship target recognition via multiscale feature attention and adaptive-weighed classifier. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1–5. [CrossRef]
15. Zhang, Z.L.; Zhang, T.; Liu, Z.Y.; Zhang, P.J.; Tu, S.S.; Li, Y.J.; Waqas, M. Fine-Grained ship image recognition based on BCNN with inception and AM-Softmax. *CMC-Comput. Mater. Contin.* **2022**, *73*, 1527–1539.
16. Han, Y.Q.; Yang, X.Y.; Pu, T.; Peng, Z.M. Fine-Grained recognition for oriented ship against complex scenes in optical remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–18. [CrossRef]
17. Liu, J.; Chen, H.; Wang, Y. Multi-Source remote sensing image fusion for ship target detection and recognition. *Remote Sens.* **2021**, *13*, 4852. [CrossRef]
18. Cheng, G.; Xie, X.; Han, J.; Guo, L.; Xia, G.S. Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 3735–3756. [CrossRef]
19. Zhu, X.; Tuia, D.; Mou, L.; Xia, G.; Zhan, L.; Xu, F.; Fraundorfer, F. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36. [CrossRef]
20. Tsagkatakis, G.; Aidini, A.; Fotiadou, K.; Giannopoulos, M.; Pentari, A.; Tsakalides, P. Survey of deep-learning approaches for remote sensing observation enhancement. *Sensors* **2019**, *19*, 3929. [CrossRef]
21. Gargees, R.S.; Scott, G.J. Deep Feature Clustering for Remote Sensing Imagery Land Cover Analysis. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 1386–1390. [CrossRef]
22. Tan, C.; Ewe, H.; Chuah, H. Agricultural crop-type classification of multi-polarization SAR images using a hybrid entropy decomposition and support vector machine technique. *Int. J. Remote Sens.* **2011**, *32*, 7057–7071. [CrossRef]
23. Xia, J.; Yokoya, N.; Iwasaki, A. Hyperspectral image classification with canonical correlation forests. *IEEE Trans. Geosci. Remote Sens.* **2016**, *55*, 421–431. [CrossRef]

24. Jafarzadeh, H.; Mahdianpari, M.; Gill, E.; Moham-madimanesh, F.; Homayouni, S. Bagging and boosting ensemble classifiers for classification of multispectral, hyperspectral and PolSAR data: A comparative evaluation. *Remote Sens.* **2021**, *13*, 4405. [CrossRef]
25. Li, X.; Lei, L.; Zhang, C.G.; Kuang, G.Y. Multimodal Semantic Consistency-Based Fusion Architecture Search for Land Cover Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
26. Cao, J.; Liu, K.; Zhuo, L.; Liu, L.; Zhu, Y.; Peng, L. Combining UAV-based hyperspectral and LiDAR data for mangrove species classification using the rotation forest algorithm. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *102*, 102414. [CrossRef]
27. Yu, K.; Zheng, X.; Fang, B.; An, P.; Huang, X.; Luo, W.; Ding, J.F.; Wang, Z.; Ma, J. Multimodal Urban Remote Sensing Image Registration Via Roadcross Triangular Feature. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 4441–4451. [CrossRef]
28. Li, W.; Gao, Y.H.; Zhang, M.M.; Tao, R.; Du, Q. Asymmetric feature fusion network for hyperspectral and SAR image classification. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *34*, 8057–8070. [CrossRef]
29. Schmitt, M.; Zhu, X. Data fusion and remote sensing: An ever-growing relationship. *IEEE Geosci. Remote Sens. Mag.* **2016**, *4*, 6–23. [CrossRef]
30. Zhang, Z.; Vosselman, G.; Gerke, M.; Persello, C.; Tuia, D.; Yang, M. Detecting Building Changes between Airborne Laser Scanning and Photogrammetric Data. *Remote Sens.* **2019**, *11*, 2417. [CrossRef]
31. Schmitt, M.; Tupin, F.; Zhu, X. Fusion of SAR and optical remote sensing data–challenges and recent trends. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 23–28 July 2017.
32. Kulkarni, S.; Rege, P. Pixel level fusion recognition for SAR and optical images: A review. *Inf. Fusion.* **2020**, *59*, 13–29. [CrossRef]
33. Rényi, A. On measures of dependence. *Acta Math. Hung.* **1959**, *3*, 441–451. [CrossRef]
34. Huang, S.; Xu, X. On the sample complexity of HGR maximal correlation functions for large datasets. *IEEE Trans. Inf. Theory* **2021**, *67*, 1951–1980. [CrossRef]
35. Liang, Y.; Ma, F.; Li, Y.; Huang, S. Person recognition with HGR maximal correlation on multimodal data. In Proceedings of the 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 25 August 2021.
36. Wang, L.; Wu, J.; Huang, S.; Zheng, L.; Xu, X.; Zhang, L.; Huang, J. An efficient approach to informative feature extraction from multimodal data. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 5281–5288.
37. Ma, F.; Li, Y.; Ni, S.; Huang, S.; Zhang, L. Data augmentation for audio-visual emotion recognition with an efficient multimodal conditional GAN. *Appl. Sci.-Basel.* **2022**, *12*, 527.
38. Pande, S.; Banerjee, B. Self-Supervision assisted multimodal remote sensing image classification with coupled self-looping convolution networks. *Neural Netw.* **2023**, *164*, 1–20. [CrossRef] [PubMed]
39. Poria, S.; Cambria, E.; Bajpai, R.; Hussain, A. A review of affective computing: From unimodal analysis to multimodal fusion. *Inf. Fusion.* **2017**, *37*, 98–125. [CrossRef]
40. Pan, J.; He, Z.; Li, Z.; Liang, Y.; Qiu, L. A review of multimodal emotion recognition. *CAAI Trans. Int. Syst.* **2020**, *15*, 633–645.
41. Pedergnana, M.; Marpu, P.; Dalla, M.; Benediktsson, J.; Bruzzone, L. Classification of remote sensing optical and LiDAR data using extended attribute profiles. *IEEE J. Sel. Top. Signal Process.* **2012**, *6*, 856–865. [CrossRef]
42. Kim, Y.; Lee, H.; Provost, E. Deep learning for robust feature generation in audiovisual emotion recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26 May 2013.
43. Kim, S.; Song, W.; Kim, S. Double weight-based SAR and infrared sensor fusion for automatic ground target recognition with deep learning. *Remote Sens.* **2018**, *10*, 72. [CrossRef]
44. 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, Las Vegas, NV, USA, 26 January 2016.
45. Huang, M.; Xu, Y.; Qian, L.; Shi, W.; Zhang, Y.; Bao, W.; Wang, N.; Liu, X.; Xiang, X. The QXS-SAROPT dataset for deep learning in SAR-optical data fusion. *arXiv* **2021**, arXiv:2103.08259. [CrossRef]
46. Debes, C.; Merentitis, A.; Heremans, R.; Hahn, J.; Frangiadakis, N.; Van Kasteren, T.; Liao, W.; Bellens, R.; Pizurica, A.; Gautama, S. Hyperspectral and LiDAR data fusion: Outcome of the 2013 GRSS data fusion contest. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 2405–2418. [CrossRef]
47. Gader, P.; Zare, A.; Close, R.; Aitken, J.; Tuell, G. *MUUFL Gulfport Hyperspectral and Lidar Airborne Data Set*; University of Florida: Gainesville, FL, USA, 2013.
48. Du, X.; Zare, A. *Technical Report: Scene Label Ground Truth Map for MUUFL Gulfport Data Set*; University of Florida: Gainesville, FL, USA, 2017.
49. Hong, D.; Hu, J.; Yao, J.; Chanussot, J.; Zhu, 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] [PubMed]
50. Bao, W.; Huang, M.; Zhang, Y.; Xu, Y.; Liu, X.; Xiang, X. Boosting ship detection in SAR images with complementary pretraining techniques. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 8941–8954. [CrossRef]
51. Qian, L.; Liu, X.; Huang, M.; Xiang, X. Self-Supervised pre-training with bridge neural network for SAR-optical matching. *Remote Sens.* **2022**, *14*, 2749. [CrossRef]
52. Roy, S.K.; Deria, A.; Hong, D.; Rasti, B.; Plaza, A.; Chanussot, J. Multimodal fusion transformer for remote sensing image classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–20. [CrossRef]
53. Franco, A.; Oliveira, L. Convolutional covariance features: Conception, integration and performance in person re-identification. *Pattern Recognit.* **2017**, *61*, 593–609. [CrossRef]

54. Hong, D.; Yokoya, N.; Chanussot, J.; Zhu, X. CoSpace: Common subspace learning from hyperspectral-multispectral correspondences. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 4349–4359. [CrossRef]
55. Hang, R.; Li, Z.; Ghamisi, P.; Hong, D.; Xia, G.; Liu, Q. Classification of hyperspectral and LiDAR data using coupled CNNs. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 4939–4950. [CrossRef]
56. Mohla, S.; Pande, S.; Banerjee, B.; Chaudhuri, S. FusAtNet: Dual attention based SpectroSpatial multimodal fusion network for hyperspectral and lidar classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA, 14 June 2020.
57. Khan, A.; Raufu, Z.; Sohail, A.; Khan, A.R.; Asif, A.; Farooq, U. A survey of the vision transformers and their CNN-transformer based variants. *Artif. Intell. Rev.* **2023**, *56*, 2917–2970. [CrossRef]
58. Hong, D.; Han, Z.; Yao, J.; Gao, L.; Zhang, B.; Plaza, A.; Chanussot, J. SpectralFormer: Rethinking hyperspectral image classification with transformers. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
59. Xu, G.; Jiang, X.; Zhou, Y.; Li, S.; Liu, X.; Lin, P. Robust land cover classification with multimodal knowledge distillation. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–16. [CrossRef]