electronics

MDPI

Article

Deep Multi-Similarity Hashing with Spatial-Enhanced Learning for Remote Sensing Image Retrieval

Huihui Zhang [1,2], Qibing Qin [1], Meiling Ge [1,*] and Jianyong Huang [3]

[1] School of Computer Engineering, Weifang University, Weifang 261061, China; huihui@wfu.edu.cn (H.Z.); qinbing@wfu.edu.cn (Q.Q.)
[2] Faculty of Computer Science and Technology, Qilu University of Technology (Shandong Academy of Sciences), Jinan 250353, China
[3] Qingdao Education Equipment and Information Technology Center, Qingdao 266022, China; 11170211020@stu.ouc.edu.cn
* Correspondence: gemeiling@wfu.edu.cn

Abstract: Remote sensing image retrieval (RSIR) plays a crucial role in remote sensing applications, focusing on retrieving a collection of items that closely match a specified query image. Due to the advantages of low storage cost and fast search speed, deep hashing has been one of the most active research problems in remote sensing image retrieval. However, remote sensing images contain many content-irrelevant backgrounds or noises, and they often lack the ability to capture essential fine-grained features. In addition, existing hash learning often relies on random sampling or semi-hard negative mining strategies to form training batches, which could be overwhelmed by some redundant pairs that slow down the model convergence and compromise the retrieval performance. To solve these problems effectively, a novel Deep Multi-similarity Hashing with Spatial-enhanced Learning, termed DMsH-SL, is proposed to learn compact yet discriminative binary descriptors for remote sensing image retrieval. Specifically, to suppress interfering information and accurately localize the target location, by introducing a spatial enhancement learning mechanism, the spatial group-enhanced hierarchical network is firstly designed to learn the spatial distribution of different semantic sub-features, capturing the noise-robust semantic embedding representation. Furthermore, to fully explore the similarity relationships of data points in the embedding space, the multi-similarity loss is proposed to construct informative and representative training batches, which is based on pairwise mining and weighting to compute the self-similarity and relative similarity of the image pairs, effectively mitigating the effects of redundant and unbalanced pairs. Experimental results on three benchmark datasets validate the superior performance of our approach.

Keywords: deep hashing; remote sensing image retrieval; fine-grained features; random sampling; spatial-enhanced learning; multi-similarity loss

check for updates

Citation: Zhang, H.; Qin, Q.; Ge, M.; Huang, J. Deep Multi-Similarity Hashing with Spatial-Enhanced Learning for Remote Sensing Image Retrieval. *Electronics* **2024**, *13*, 4520. https://doi.org/10.3390/electronics13224520

Academic Editors: José Carlos Castillo and George A. Papakostas

Received: 14 October 2024
Revised: 2 November 2024
Accepted: 14 November 2024
Published: 18 November 2024

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

1. Introduction

With the rapid advancement of remote sensing technology, there has been an exponential increase in both the volume and quality of remote sensing (RS) data [1,2]. Today, extensive databases of remote sensing images encompass a wealth of information, playing a crucial role in various domains such as environmental conservation [3] and disaster management [4]. However, this surge in the amount of remote sensing imagery presents significant challenges in efficiently and effectively retrieving relevant images from these large-scale datasets. Consequently, the image retrieval of large-scale remote sensing has become a prominent area of focus within the research community [5,6].

As data volume and dimensionality continue to expand, the corresponding storage requirements increase significantly. Hash-based search techniques have recently been introduced in the realm of remote sensing image retrieval to address challenges related to storage and computational efficiency. These techniques have become some of the most

popular and effective solutions for large-scale image retrieval tasks. Hash learning focuses on projecting high-dimensional data into low-dimensional discrete Hamming space via hash projection. The similarity between data is then computed through binary coding and XOR operations on hash codes, which helps to reduce storage needs and enhance retrieval efficiency while preserving the similarity between the original images [7,8]. Hashing methods are generally classified into two primary types: supervised hashing [9] and unsupervised hashing [10]. Unlike supervised hashing, which employs manually labeled data to direct model training and produce more representative and informative binary hash codes [11], unsupervised hashing frameworks automatically learn the distribution and structural features of the data to map similar samples into similar hash code spaces. It should be noted that supervised hashing typically outperforms unsupervised methods due to the use of manual semantic annotations, which offer more precise and meaningful guidance for generating hash codes. In this paper, we focus on supervised hashing learning to produce similarity-preserving and discriminative hash codes, enhancing the efficiency of large-scale image retrieval.

As deep neural networks demonstrate powerful feature learning capabilities in computer vision work, hashing methods based on deep neural networks have gradually become the main method in image retrieval in large-scale remote sensing [12]. Compared to handmade features, current complex deep-hashing techniques have made substantial progress in remote sensing image retrieval. However, these methods still have some limitations [13]. During the deep hashing framework optimization process, constructing representative training batches through random sampling or semi-hard negative sample mining mechanisms may introduce some redundant pairs, which can slow down the model's convergence speed [14,15]. Figure 1 provides a clear example in which the random sampling strategy overlooks the distributional relationships of the original samples, leading to an imbalance in the training batches, with a small number of positive samples and a large number of negative samples. This sampling mechanism causes parameter optimization to be biased toward the majority of semantic categories, hindering the learning of the hashing function. The presence of complex backgrounds can introduce noise and instability into feature extraction, making it difficult to provide robust feature responses to objects. In addition, content-based remote sensing image retrieval methods use deep neural networks to extract semantic features at different levels.

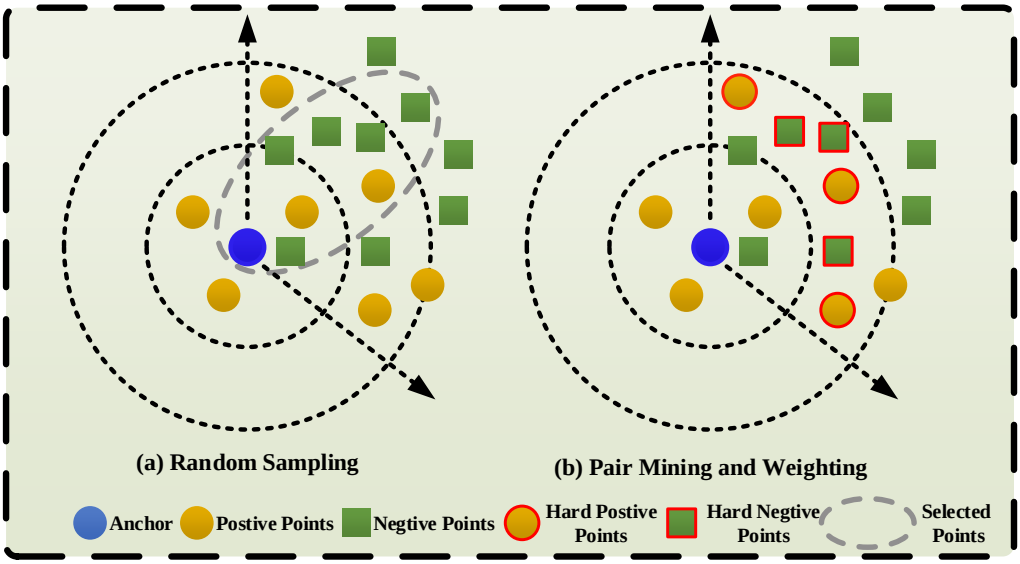


Figure 1. The motivation of the proposed deep multi-similarity hash framework. (**a**) The random sampling strategy ignores the distribution relationship of the original samples, resulting in an imbalanced sample problem in the training batch; that is, it contains a small number of positive samples and a large number of negative samples. (**b**) The pair mining and weighting strategy explores multiple similarity relationships between sample pairs to construct representative training batches.

However, these learned features are often affected by noise, such as background information, which can limit their ability to provide a stronger response for the target. From the reasons presented above, the performance improvement of deep hashing in remote sensing image retrieval is relatively restricted and is far from satisfactory [15,16].

To solve the above problems, we propose a novel unified deep hashing framework called Deep Multi-similarity Hashing with Spatial-enhanced Learning (DMsH-SL) to generate discrete codes with significant discrimination for efficient and accurate remote sensing image retrieval, as illustrated in Figure 2. Specifically, our proposed DMsH-SL framework consists of two main parts. (1) Noise-robust and fine-grained semantic characteristics are obtained by designing the spatial group-enhanced hierarchical network. (2) Hashing codes are generated by jointly optimizing the multi-similarity loss and classification loss. The main contributions of this paper are summarized as follows.

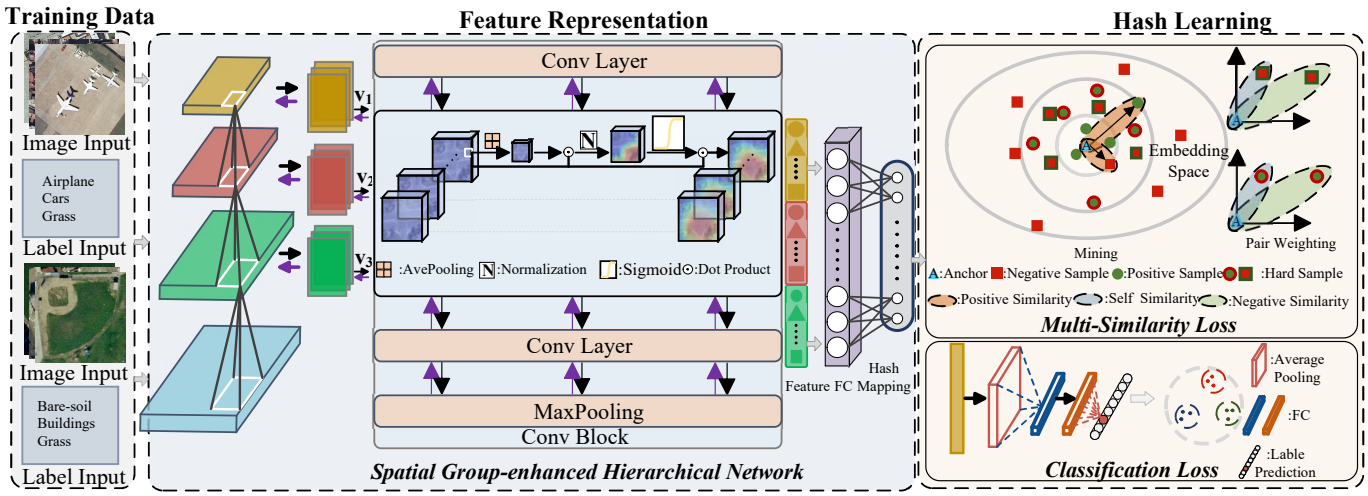


Figure 2. Overview of our proposed DMsH-SL framework, which mainly includes two parts: (1) Feature Representation: A spatial group-enhanced hierarchical network is proposed for the noise-robust and fine-grained semantic representation. (2) Hash Learning: Multi-similarity loss and classification loss are jointly explored to optimize the parameters of the deep hashing framework.

- By introducing spatial-enhanced learning into each feature group mechanism, the spatial group-enhanced hierarchical network is designed to highlight the spatial distribution of different semantic sub-features, which utilize the similarities between the global and local characteristics in each group to learn the attention mask at each position, generating the noise-robust and discriminative representation.
- By employing pair mining and weighting to calculate self-similarity and relative similarity between pairs, the multi-similarity loss is incorporated into the deep hashing to construct the informative and representative training batches, effectively mitigating the effects of redundant and unbalanced pairs.
- Extensive experiments on three widely used benchmark datasets show that our DMsH-SL framework surpasses other state-of-the-art hashing methods for remote sensing image retrieval applications.

The rest of this paper is organized as follows. Section 2 summarizes the related work, and we demonstrate every detail of our approach in Section 3. In Section 4, we design and perform the empirical evaluation of a series of experiments to access our proposed framework, and we conclude the study in Section 5.

2. Related Work

2.1. Remote Sensing Image Retrieval

The primary objective of remote sensing image retrieval is to identify samples that closely match the given queries. Traditional methods for retrieving remote sensing images

rely on hand-crafted features to represent the content of the images. Unfortunately, these hand-made features often do not capture the intricate semantic details of remote sensing images accurately, leading to less effective retrieval outcomes. With significant advances in deep learning within computer vision, combining convolutional neural networks (CNNs) with hashing techniques has become a prevalent approach in remote sensing image retrieval (RSIR). Hashing-based retrieval measures the Hamming distance between binary codes produced by the hashing function. The appeal of hash learning lies in its reduced storage requirements and rapid search capabilities, which has led to considerable interest from researchers in remote sensing. Over the past few years, many related algorithms have been developed. For example, Song et al. reformulate the image retrieval problem as visual and semantic retrieval of images and propose a novel Deep Hashing Convolutional Neural Network (DHCNN) to simultaneously retrieve similar images and classify their semantic labels in a unified framework [17]. To solve the data imbalance problem, Li et al. propose a quantized deep learning to hash (QDLH) framework for the retrieval of large-scale remote sensing images [18]. Liu et al. represent a CBRSIR approach called Feature and Hash (FAH) learning, which consists of a deep feature learning model (DFLM) and an adversarial hash learning model (AHLM) [19]. In addition, Song et al. propose an Asymmetric Hash Code Learning (AHCL) designed to produce binary codes for both queries and database images in an asymmetric manner [20].

2.2. Hash Learning

Hash learning focuses on mapping continuous high-dimensional feature vectors into discrete Hamming spaces using non-linear hash functions while retaining semantic information. Due to their impressive efficiency in terms of speed and storage, hashing techniques have garnered significant attention from researchers around the world. Hash learning methods can typically be categorized into two main types: unsupervised hashing [21,22] and supervised hashing [23,24]. Unsupervised hashing techniques automatically learn the distribution and structural properties of the data and then map similar samples to corresponding hash code spaces based on these properties. However, because of the limited capacity of hash codes, these methods often lack robustness against noise and variations in images. In unsupervised hashing approaches, the hash functions are learned from unlabeled training datasets. Typical unsupervised hashing methods include spectral hashing (SH) [25], iterative quantization (ITQ) [26], and discrete graph hashing (DGH) [27]. On the other hand, supervised hashing methods aim to utilize labeled information to generate hash codes. These methods predominantly encompass techniques such as kernel-based supervised hashing (KSH) [28] and sparse embedding with minimum variance encoding (SELVE) [29], and so on.

As deep learning technology advances rapidly, researchers are integrating it with hashing functions to create end-to-end frameworks for training deep hashing models. Zhao et al. [30] propose a multi-scale contextual deep hashing network and suppress the interference of irrelevant information. Ye et al. [31] propose an end-to-end deep hash remote sensing image retrieval method which uses a global attention mechanism and multi-head self-attention mechanism to construct a multi-scale feature fusion module (MSF) to reduce background interference and enhance the representation ability of image features. To learn the inter-class similarity and intra-class similarity between RS images, Sun et al. [32] present an unsupervised framework based on soft pseudo-labels for content-based RS image retrieval.

Unlike previous studies, by introducing spatial-enhanced learning into each feature group mechanism, the spatial group-enhanced hierarchical network is designed to highlight the spatial distribution of different semantic sub-features, and the multi-similarity loss is incorporated into the deep hashing to construct the informative and representative training batches by employing pair mining and weighting to calculate self-similarity and relative similarity between pairs.

3. Methodology

Given a collection O containing n image points $O = \{o_1, o_2, \ldots, o_n\}$, and $\mathcal{Y} = \{Y_1, Y_2, Y_3, \ldots, Y_n\}$ denotes the corresponding annotated information, of which there are a total of C semantic categories. Each image $o_i \in O$ has multiple semantic labels, where each sample can be represented by a label vector $Y_i = \{y_i^1, y_i^2, \ldots, y_i^C\}$. To reduce storage space and improve retrieval speed, by learning the project function, the obtained hashing could generate compact binary codes $H(O) = \{H(o_1), H(o_2), \ldots, H(o_n)\}$, where $H(o_i) = \{-1, 1\}^K$ and K represents the number of hash bits.

3.1. Spatial Group-Enhanced Hierarchical Network

Due to the complex background information in multi-label remote sensing images, convolutional neural networks struggle to capture semantic targets effectively [16]. Since hierarchical networks can extract image features at different levels, this hierarchical property enables the network to capture complex spatial and semantic information in remote sensing images. To enhance the feature learning ability of the model, we develop a novel hierarchical network based on spatial enhancement to generate discriminative embedding representations by learning the spatial information of different levels of semantic sub-features.

Specifically, the multi-label remote sensing image o_i is fed into the pre-trained ResNet50 backbone network to learn the hierarchical feature representations $V = \{v_1, v_2, v_3\}$, where $v_i \in R^{H \times W \times C}$. To enhance spatial information and pinpoint the key areas, we introduce a spatial-enhanced learning module into the network. As illustrated in Figure 2, we partition the obtained intermediate representation v_i into G groups along the channel direction, and each group has a vector representation $P = \{P_1, P_2, \ldots P_m\}$, where $m = H \times W$, $P_j \in R^{(C/G)}$. Different from other attention mechanisms [33], our approach integrates global statistical and local spatial features to learn the discriminative semantic image features. We also use average pooling $Ave(\cdot)$ to derive the global statistical representation f_g and compute the similarity d_j between global and spatial features via dot product operation.

$$f_g = Ave(P) = \frac{1}{m} \sum_{j=1}^{m} P_j \tag{1}$$

$$d_j = f_g P_j = \|f_g\| \|P_j\| \cos \theta_{f_g, P_j} \tag{2}$$

where θ_{f_g, P_j} is the angle between f_g and P_j. To prevent the deviation problem of different sample coefficients, by introducing the normalization operation, two variables α and β are added to ensure the normalized identity transformation. To obtain the significant features, we use the Sigmoid function to scale them to a certain range. The specific formula is as follows.

$$\begin{aligned} \hat{d}_j &= \frac{d_j - \mu_d}{\sigma_d + \tau}, a_j = \alpha \cdot \hat{d}_j + \beta \\ \hat{P}_j &= P_j \cdot \sigma(a_j) \end{aligned} \tag{3}$$

where $\tau = 1 \times 10^{-5}$, $\alpha = \beta = G$, $\mu_d = (1/m) \sum_i^m d_i$, $\sigma_d^2 = (1/m) \sum_i^m (d_i - \mu_d)^2$, and $\sigma(\cdot)$ is the Sigmoid function. The enhanced feature of v_i is represented as $\hat{P} = \{\hat{P}_1, \hat{P}_2, \ldots, \hat{P}_m\}$. By concatenation operations, the features with different levels are combined to obtain the representative representation $F_i = concat(f_1, f_2, f_3)$.

3.2. Multi-Similarity Loss

To construct informative and representative training batches, based on pair mining and pair weighting [14], the multi-similarity loss is proposed to explore the semantic correlation between sample pairs, involving self-similarity within samples, the relative similarity of negative samples, and the relative similarity of positive samples.

3.2.1. Pairwise Sampling

The multi-label image pair (o_i, o_j) is fed into the hashing network to generate the corresponding binary code (H_i, H_j), and the similarity matrix $S_{ij} = \langle H(o_i), H(o_j) \rangle \in R^{n \times n}$ could be calculated, where $< \cdot, \cdot >$ represents the dot product operation. Initially, we identified representative data points by calculating the positive similarity between sample pairs and removing those redundant pairs that impair model performance. Specifically, for a given anchor point o_i, we selected a negative sample with a similarity higher than the hardest positive sample (i.e., minimum similarity). According to Equation (4), these chosen negative pairs are employed to construct the negative dataset, denoted as N_i.

$$S_{ij}^{-} > \min_{y_k = y_i} S_{ik} - \omega \tag{4}$$

where ω represents the threshold value. Similarly, based on Equation (5), the positive sample pair P_i could be generated, which exhibits a similarity lower than the hardest negative sample with maximum similarity.

$$S_{ij}^{+} < \min_{y_k \neq y_i} S_{ik} + \omega \tag{5}$$

3.2.2. Pairwise Weighting

The selected sample pairs are weighted by calculating self-similarity and negative sample similarity to further select informative samples, improving the performance of the model. Specifically, on the basis of the cosine similarity of sample pairs, the self-similarity could be defined, as shown below.

$$\Omega_{ij}^{S} = e^{\gamma(\lambda - S_{ij})} \tag{6}$$

where the parameter Ω denotes the self-similarity. The relative similarity of negative samples (o_i, o_j) not only considers self-similarity but also calculates the similarity between other remaining negative sample pairs, which could be formulated as follows.

$$\Omega_{ijk}^{N} = e^{\gamma(S_{ik} - S_{ij})} \tag{7}$$

By introducing the hyperparameter γ, the negative similarity and self-similarity could be obtained. Meanwhile, based on the learned self-similarity and negative similarity, the negative pair $(o_i, o_j) \in N_i$ could be weighted.

$$\begin{aligned} w_{ij}^{-} &= \frac{1}{\Omega_{ij}^{S} - \sum\limits_{k \in N_i} \Omega_{ijk}^{N}} = \frac{1}{e^{\gamma(\lambda - S_{ij})} - \sum\limits_{k \in N_i} e^{\gamma(S_{ik} - S_{ij})}} \\ &= \frac{e^{\gamma(S_{ik} - \lambda)}}{1 + \sum\limits_{k \in N_i} e^{\gamma(S_{ik} - \lambda)}} \end{aligned} \tag{8}$$

Similarly, the positive samples $(o_i, o_j) \in P_i$ could be weighted by computing Equation (9).

$$w_{ij}^{+} = \frac{1}{e^{-\eta(\lambda - S_{ij})} - \sum\limits_{k \in P_i} e^{-\eta(S_{ik} - S_{ij})}} \tag{9}$$

By jointly computing self-similarity and relative similarity, our proposed multi-similarity is formulated as follows.

$$\mathcal{L}_{ms} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \begin{array}{l} \frac{1}{\eta} \log \left[1 + \sum\limits_{k \in P_i} e^{-\eta(S_{ik} - \lambda)} \right] \\ + \frac{1}{\gamma} \log \left[1 + \sum\limits_{k \in N_i} e^{\gamma(S_{ik} - \lambda)} \right] \end{array} \right\} \tag{10}$$

3.3. Out-of-Sample Extension

After optimizing the model, the spatial group-enhanced hierarchical network captures the high-dimensional feature vector of the query sample o_q, and the subsequent hash layer maps the feature representation to a discrete Hamming space to generate a K-bit hash code. Based on the binary code obtained, we can quickly and accurately return data similar to the query sample. The hash code for a specific query sample o_q is obtained by the following formula:

$$H(o_q) = H(F_{o_q}) = sign(W_h^T F_{o_q} + b_h) \tag{11}$$

where W_h and b_h are the weight matrix and the bias coefficient of the hash layer, respectively.

3.4. Classification Loss

Previous research has shown that semantic labels could lead to compact binary representations of raw images [34]. To enhance discriminative power, we incorporate a classification component-based cross-entropy loss into our framework, generating high-quality binary codes. In addition, to improve accuracy, we utilize label smoothing (as shown in Equation (12)) to transform the original one-hot labels into smoother variants.

$$\mathcal{L}_{cla} = -\sum_{i=1}^{n} [y_i(1-\delta) + \delta/C] \log \frac{\exp(\bar{y}_i)}{\sum_{h=1}^{C} \exp(\bar{y}_{ih})} \tag{12}$$

where $y_i \in \{0, 1\}$, δ is a learnable parameter, which is set as $\{0.5, 0.8, 1.0\}$. During hash learning, the model employs deep multi-similarity loss, which considers the semantic relationship between samples, to construct representative training batches. The proposed framework integrates the classification loss to optimize the model parameters, generating high-quality discrete codes. The final optimization problem is formulated below.

$$\mathcal{L}_{all} = \mathcal{L}_{ms} + \xi \cdot \mathcal{L}_{cla} \tag{13}$$

where ξ denotes the weighted parameter, which could balance the multi-similarity item and the classification function.

4. Experiments

To verify the effectiveness of our designed DMsH-SL for multi-label remote sensing image retrieval, we conducted comprehensive and systematic experiments on three commonly used public datasets, including UCMerced, MLRSNet, and DFC15.

4.1. Datasets

UCMerced dataset [35] is a multi-label dataset designed to retrieve remote sensing images. It encompasses 100 distinct categories of samples sourced from the United States Geological Survey (USGS), specifically focusing on urban regions throughout the country. The images are sized at 256 × 256 pixels, each pixel having a spatial resolution of 0.3 m. This dataset is widely used in remote sensing research to examine land use trends and urban growth. In our experiment, we created the training set by randomly choosing 80% of the scene categories, while the remaining images were reserved for testing purposes.

MLRSNet dataset [36] is a multi-label dataset with a high spatial resolution for remote sensing tasks, containing a total of 109,161 images and 60 different semantic labels. Each image is annotated with many labels, ranging between 1 and 13, and features a pixel dimension of 256 × 256 with varying resolutions. To set up the training data, 5000 images are randomly chosen, whereas the query dataset comprises 1000 images. The remaining images are allocated for retrieval.

DFC15 dataset [37] serves as a multi-label benchmark and is derived from the DFC15 semantic segmentation collection. It includes a total of 3342 images distributed across eight categories, each with a resolution of 600 × 600 pixels. For model training, we randomly

choose 80% of the images as the training set. The remaining 20% of the images are set aside for testing.

4.2. Experimental Settings

4.2.1. Baselines

To evaluate the effectiveness of our proposed DMsH-SL method, we performed a comprehensive comparison with several leading deep hashing techniques. These techniques include DPSH [38], ADSH [39], DHCNN [17], OrthoHash [40], HyP^{2}Loss [41], RelaHash [42], HHF [43], and SWTH [44], which are applied to the benchmark datasets mentioned earlier. Unlike traditional hashing methods, these advanced deep hashing approaches leverage deep neural networks to simultaneously learn feature representations and hash projections in a unified end-to-end framework. Furthermore, the parameter configurations for each model are based on the specifications provided in their respective original papers or available open-source implementations.

4.2.2. Implementation Details

All the experiments are performed on the Pytorch framework with an NVIDIA GeForce RTX 3090 GPU. In the DMsH-SL model, we set the training batch to 32 and use the Adam optimizer to optimize the model parameters, where the learning rate is 0.0001, the momentum is 0.9, and the weight decay rate is 0.0005. Based on the parameter tuning results reported in Section 4.4, we set the parameter values involved in the overall objective loss as follows: $\omega = 5.0$, $\lambda = 2.0$, $\eta = 3.0$, $\gamma = 50$ and the weighted parameter $\xi = 0.1$, respectively.

4.2.3. Evaluation Protocols

To thoroughly assess the retrieval performance, we employ four standard evaluation metrics: mean average precision (mAP), precision–recall (PR) curve, TopK precision curve and Hamming rank precision within Hamming radius 2 (P@H$\leq$2). Specifically, mAP is a widely used metric that measures retrieval effectiveness by calculating the average precision (AP) for query samples. The PR curve illustrates the trade-off between precision and recall with a higher model performance reflected in a larger area under the curve. The TopK precision curve shows the proportion of samples matching the queries among the TopK retrieved results, where K is set to 1000. Lastly, the P@H$\leq$2 curve indicates precision within a Hamming radius of 2.

4.3. Experimental Results

To thoroughly assess the performance of our proposed framework, we compute mAP scores for various baselines alongside our DMsH-SL model across the UCMerced, MLRSNet, and DFC15 datasets. The results are detailed in Table 1, Table 2, and Table 3, respectively. Furthermore, we include visualizations to depict the retrieval efficacy of the deep hashing methods, such as the precision–recall curve, TopK recall curve, and Hamming precision within a Hamming radius of 2. These visualizations are illustrated in Figure 3, Figure 4, Figure 5, and Figure 6, respectively. Note that, as described in Section 3, the total number of hash bits is represented as a K value that directly affects the retrieval speed. A smaller K value means faster calculation and comparison of each hash value, which speeds up the retrieval. However, too few hash bits may lead to information loss and reduce the accuracy of retrieval. Increasing the K value usually improves the accuracy of retrieval because more hash bits can better retain information about data features. However, too many hash bits may lead to the "curse of dimensionality" and affect retrieval efficiency. In our study, based on the results shown in Tables 1–3, the bigger K values achieved better performance in most cases.

Tables 1–3 compare the mean average precision (mAP) results of various methods on the UCMerced, MLRSNet, and DFC15 datasets. According to the experimental results, it can be observed that compared to other methods, the DMsH-SL deep hashing model proposed in this paper has achieved ideal performance in most cases. Specifically, for

the UCMerced dataset, our DMsH-SL framework achieves superior performance, which achieves 2.04% improvement over the best baseline SWTH with five different hash lengths. For the DFC15 dataset, our proposed framework achieves an increase of 4% at most. For the MLRSNet dataset, HHF's performance is slightly better than our model in the case of 48-bit hash coding. This may be because the hashing-guided hinge function proposed by HHF can effectively alleviate the conflict between metric learning and quantization learning. However, in the UCMerced and DFC15 datasets, the performance of the model is 5.51% and 2.89% higher than that of HHF, respectively. This is mainly because the feature extraction module based on the spatial enhancement learning module generates discriminative feature representations for multi-label remote sensing images to guide hash learning; in addition, by fully mining the semantic relationship between sample pairs, the designed deep multi-similarity function constructs representative training samples to optimize the training of the model.

Table 1. Comparison of mAP results for different baselines with 16 bits, 32 bits, 48 bits, 64 bits, and 128 bits on the UCMerced dataset.

Method	UCMerced Dataset				
	16 bits	**32 bits**	**48 bits**	**64 bits**	**128 bits**
DPSH	0.9142	0.9209	0.9065	0.9092	0.9124
ADSH	0.8799	0.8874	0.8786	0.8790	0.8849
DHCNN	0.8946	0.9088	0.9123	0.9339	0.9385
OrthoHash	0.8070	0.8390	0.8490	0.8415	0.8445
HyP^{2}Loss	0.9264	0.9395	0.9590	0.9612	0.9736
RelaHash	0.8261	0.8204	0.8291	0.8431	0.8374
HHF	0.8814	0.9071	0.9044	0.9048	0.8962
SWTH	0.9223	0.9460	0.9531	0.9612	0.9592
DMsH-SL	**0.9571**	**0.9622**	**0.9697**	**0.9724**	**0.9826**

Table 2. Comparison of mAP results for different baselines with 16 bits, 32 bits, 48 bits, 64 bits, and 128 bits on the MLRSNet dataset.

Method	MLRSNet Dataset				
	16 bits	**32 bits**	**48 bits**	**64 bits**	**128 bits**
DPSH	0.9232	0.9491	0.9463	0.9360	0.9436
ADSH	0.9430	0.9074	0.8787	0.9083	0.9070
DHCNN	0.9473	0.9491	0.9063	0.9172	0.9169
OrthoHash	0.9234	0.9278	0.9370	0.9405	0.9462
RelaHash	0.9335	0.9421	0.9484	0.9467	0.9524
HHF	0.9350	0.9639	**0.9723**	0.9667	0.9526
SWTH	0.8503	0.8554	0.8653	0.8662	0.8599
DMsH-SL	**0.9487**	**0.9642**	0.9681	**0.9711**	**0.9762**

To visually assess retrieval performance, we draw the PR curve, TopK precision curve, and P@H$\leq$2 curve with different hash lengths. Specifically, Figure 3 shows the PR and TopK precision curves for the UCMerced dataset across different hash lengths. Figure 4 presents these curves for the MLRSNet dataset with varying hash lengths, while Figure 5 illustrates the TopK precision curves for the DFC15 dataset. Our comparative results indicate that the proposed DMsH-SL framework performs well in most scenarios, underscoring its effectiveness. Additionally, Figure 6 displays the P@H$\leq$2 values for our DMsH-SL method and several baseline approaches across three benchmark datasets. Through the analysis of experimental results, the performance of the deep hashing model proposed in this paper is better than other comparative methods. This is because the multi-similarity loss function enables the DMsH-SL framework to obtain the global optimal solution and improve the identification ability of generated binary codes. At the same time,

by learning the spatial distribution information of features at different levels through the spatial group-enhanced mechanism, the designed hierarchical feature extraction network can effectively reduce noise interference in the image and better respond to key semantic targets in the image. Therefore, the P@H≤2 performance of the DMsH-SL model has been significantly improved.

Table 3. Comparison of mAP results for different baselines with 16 bits, 32 bits, 48 bits, 64 bits, and 128 bits on the DFC15 dataset.

Method	DFC15 Dataset				
	16 bits	**32 bits**	**48 bits**	**64 bits**	**128 bits**
DPSH	0.9447	0.9363	0.9197	0.9190	0.9536
ADSH	0.9584	0.9584	0.9686	0.9586	0.9586
DHCNN	0.9239	0.9521	0.9455	0.9527	0.9585
OrthoHash	0.9647	0.9557	0.9548	0.9589	0.9625
HyP^{2}Loss	0.9622	0.9677	0.9673	0.9635	0.9670
RelaHash	0.9563	0.9606	0.9617	0.9610	0.9609
HHF	0.9573	0.9603	0.9707	0.9697	0.9760
SWTH	0.9301	0.9475	0.9553	0.9564	0.9626
DMsH-SL	**0.9908**	**0.9892**	**0.9937**	**0.9941**	**0.9966**

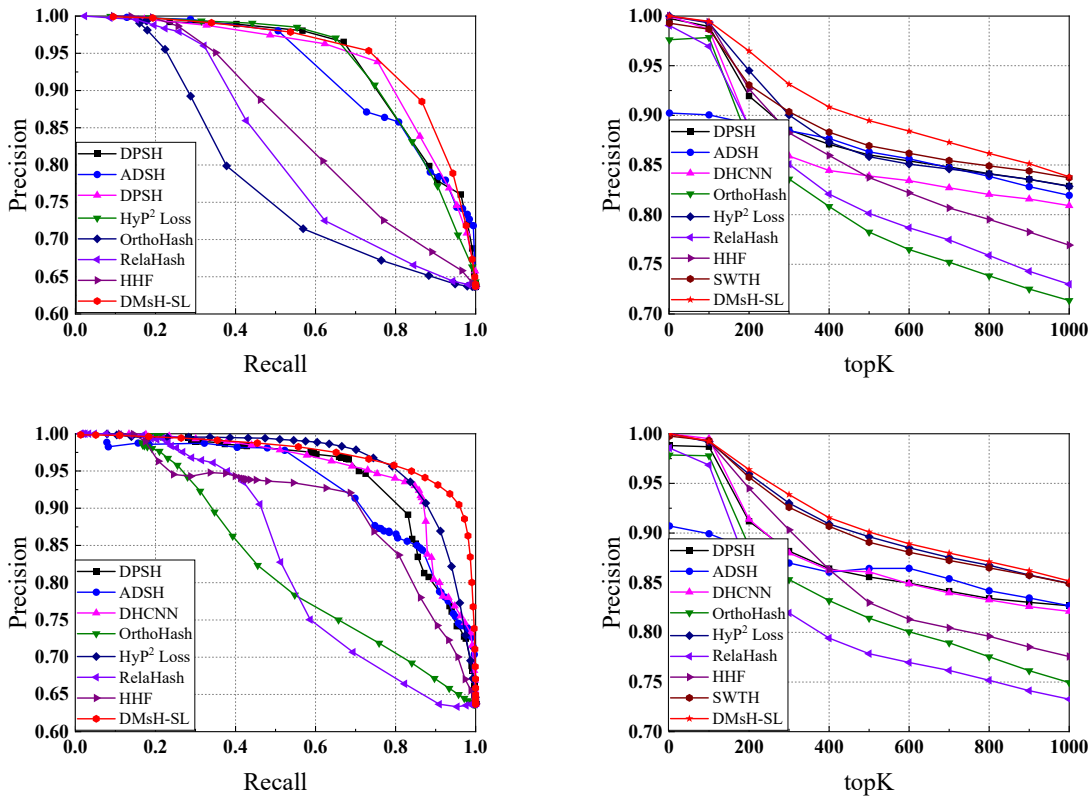


Figure 3. Results of precision–recall curves and TopK precision curves on UCMerced dataset with respect to 16 bits and 48 bits.

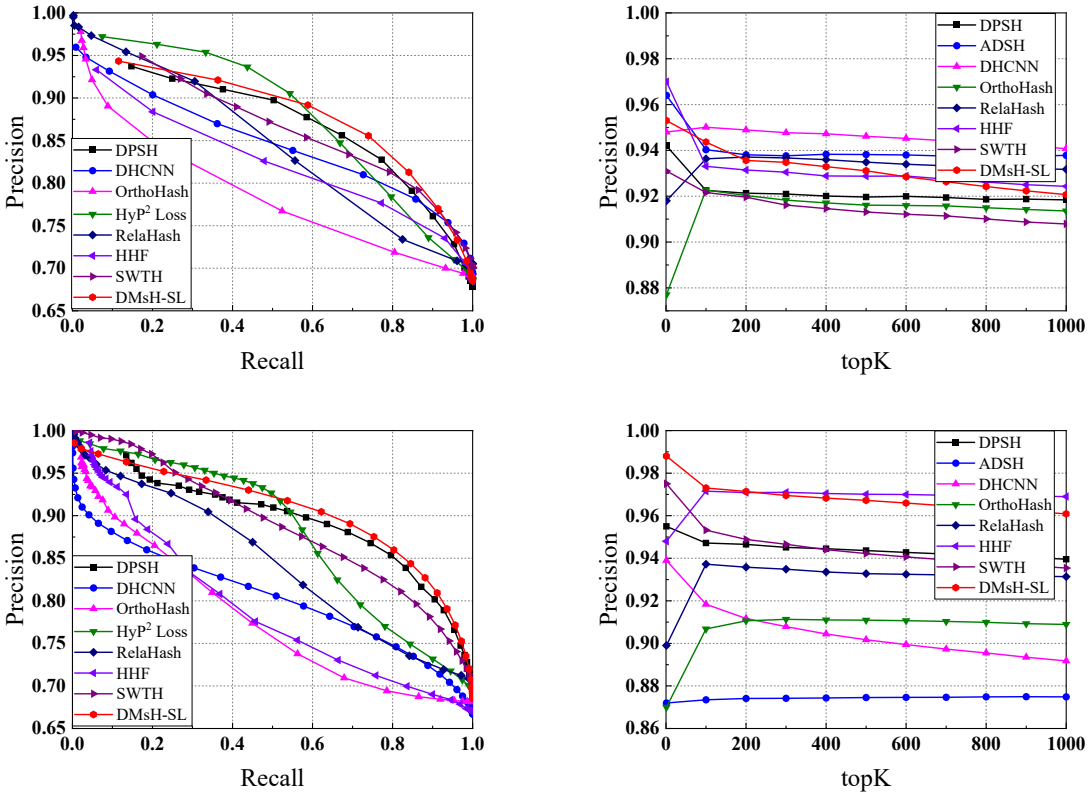


Figure 4. Results of precision–recall curves and TopK precision curves on MLRSNet dataset with respect to 16 bits and 48 bits.

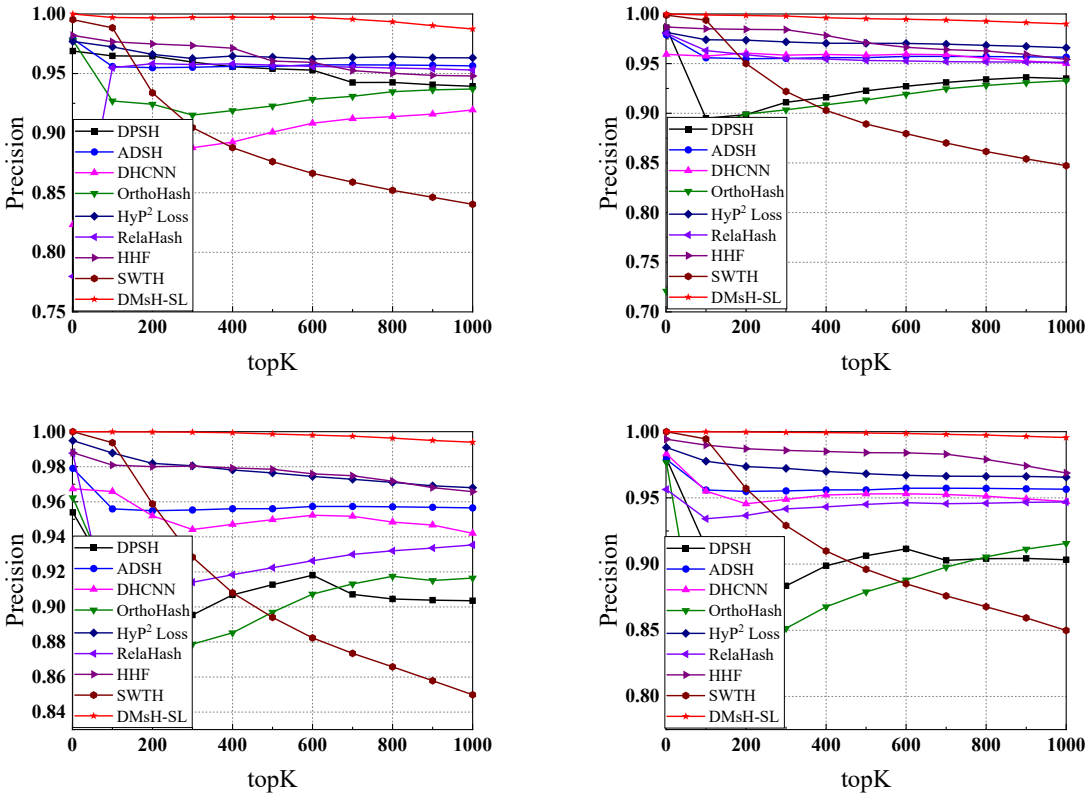


Figure 5. Results of TopK precision curves on DFC15 dataset with respect to 16 bits, 32 bits, 48 bits, and 64 bits.

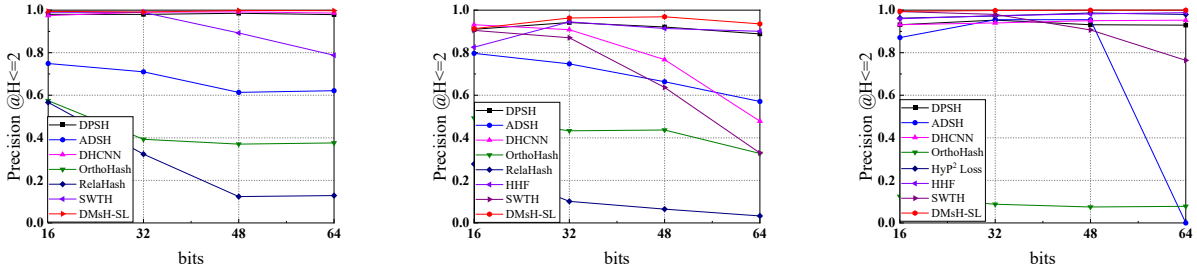


Figure 6. P@H$\leq$2 curves on UCMerced, MLRSNet, and DFC15 datasets.

4.4. Parameter Sensitivity Analysis

Parameters are also important factors that affect the performance of the model. In this section, parameter sensitivity experiments were conducted in three datasets for the loss function parameters, and the results are shown in Figure 7. We study the impact of a certain parameter on the model by fixing the other three parameters. In the stage of rough mining of samples, ω represents the threshold range of similarity. We set $\omega = \{1.0, 3.0, 5.0, 7.0, 9.0\}$ through the analysis of experimental data; when it increases to 5.0, the retrieval performance of the model on the UCMerced dataset reaches the best. Even on MLRSNet and DFC15, the mAP value does not reach the maximum value, but to maintain the consistency of parameters, we set $\omega = 5.0$ on the three datasets. To select representative training samples, the multi-similarity loss performs a weighted operation on the roughly mined sample pairs. In this process, three parameters are set to $\lambda = \{1.0, 2.0, 3.0, 4.0, 5.0\}$, $\eta = \{1.0, 2.0, 3.0, 4.0, 5.0\}$, and $\gamma = \{10, 20, 30, 40, 50\}$, respectively. Specifically, when the parameters η and γ continue to increase, the spatially enhanced deep multi-similarity hashing framework is also in a stable state. When $\eta = 3.0$ and $\lambda = 2.0$, the model obtains the best results. As for λ, as the parameter value increases on the UCMerced and DFC15 datasets, the performance of the framework continues to decline or stabilize. When $\lambda = 2.0$, the retrieval performance of the deep hashing model achieves the global optimal solution. Although the performance of MLRSNet does not reach the best state, it is close to the highest mAP value. Therefore, we set $\lambda = 2.0$ on the three datasets.

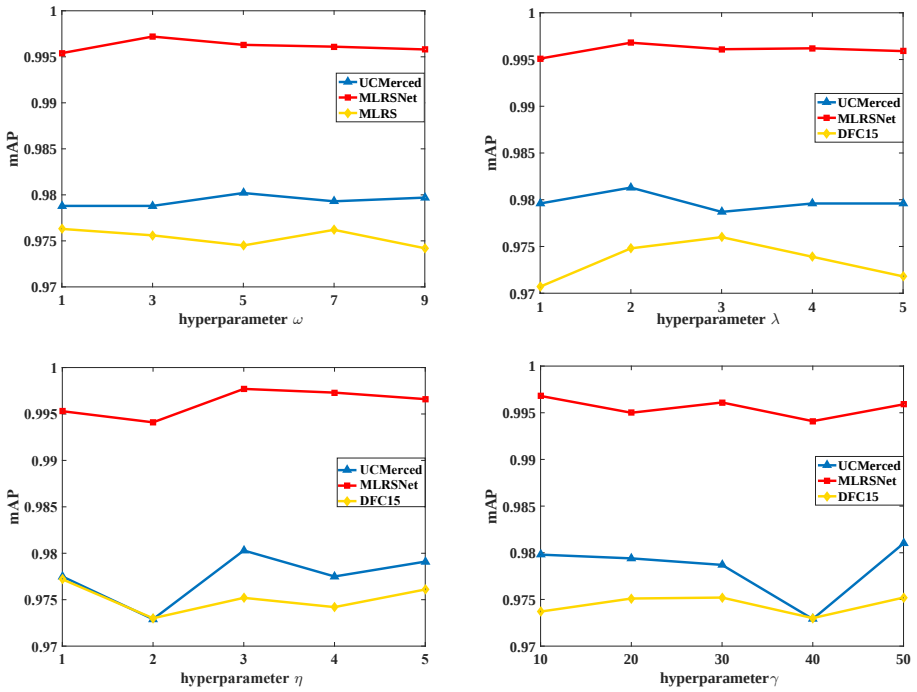


Figure 7. mAP results of different t and τ for DItSH on UCMerced and DFC15 datasets with respect to 32 bits and 48 bits.

4.5. Ablation Study

To verify the effectiveness of each module, we performed some ablation experiments on the UCMerced dataset and the MLRSNet dataset, as illustrated in Tables 4 and 5. Specifically, DMsH-SL-*C* indicates that the spatially enhanced deep hashing framework only uses the label smoothing strategy to generate hash codes. DMsH-SL-*D* means that the supervised hashing algorithm is applied from DSH [45]. The DMsH-SL-*N* framework uses Proxy-NCA [46] as the objective function. In particular, DMsH-SL-*D* overlooks relative similarities between sample pairs, while DMsH-SL-*N* focuses solely on pairwise similarities and neglects self-similarity between samples. In contrast, our proposed approach leverages the multi-similarity loss function to construct the informative training batches, improving the model performance. The comparative analysis in Tables 4 and 5 demonstrates the excellence of our DMsH-SL framework on the UCMerced dataset and MLRSNet dataset.

Table 4. Comparison of mAP results for variant comparisons of our DMsH-SL on the UCMerced dataset.

Method	16 bits	48 bits	64 bits
DMsH-SL-*C*	0.7666	0.7861	0.7812
DMsH-SL-*D*	0.8429	0.8495	0.8798
DMsH-SL-*N*	0.8771	0.9322	0.9356
DMsH-SL	**0.9571**	**0.9697**	**0.9724**

Table 5. Comparison of mAP results for variant comparisons of our DMsH-SL on the MLRSNet dataset.

Method	16 bits	48 bits	64 bits
DMsH-SL-*C*	0.7946	0.8818	0.9060
DMsH-SL-*D*	0.8401	0.8707	0.8642
DMsH-SL-*N*	0.9360	0.9567	0.9629
DMsH-SL	**0.9387**	**0.9681**	**0.9711**

4.6. Visualization

4.6.1. Grad-Cam Visualization

Remote sensing images have complex background information and other noise interference. Generally speaking, deep neural networks often use deep semantic features to represent the content of multi-label remote sensing images, which makes it impossible to accurately locate the target regions in remote sensing images and reduces the feature learning ability of the model. Unlike the above methods, the mechanism of spatial group enhancement captures the key area information in the image by learning sub-features at different levels, which is conducive to generating discriminative embedding representations. To better illustrate the superiority of the DMsH-SL framework, this section randomly visualizes the attention maps of several sample data on UCMerced, MLRSNet, and DFC15, as shown in Figure 8. Obviously, the feature extraction framework based on a spatial gropup-enhanced learning module can effectively suppress noise interference and accurately locate semantic targets in images, improving the discrimination ability of visual semantic features.

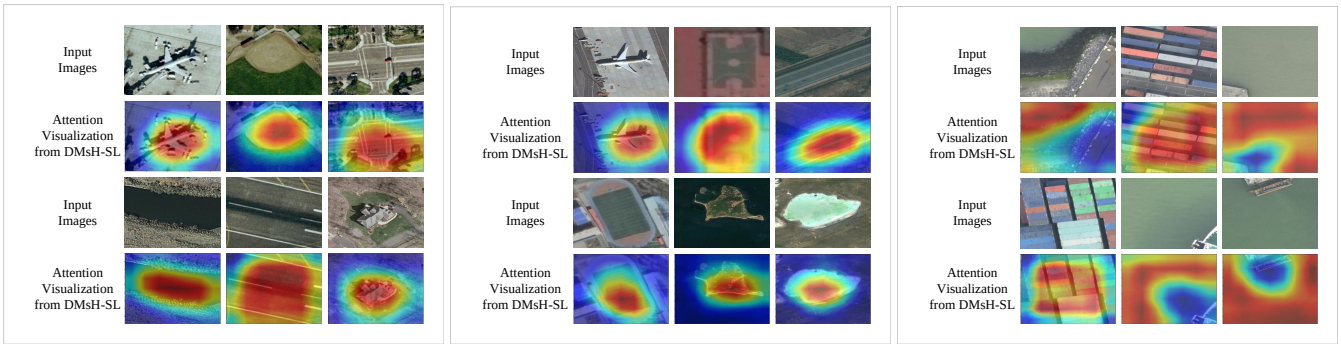
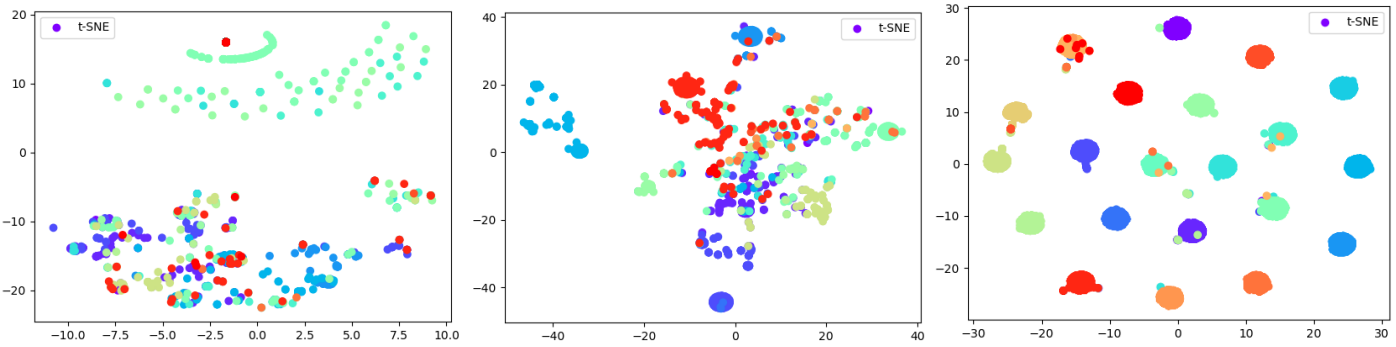


Figure 8. Some visual examples of the semantic features from attention-aware augmentation module on UCMerced, MLRSNet, and DFC15 datasets.

4.6.2. t-SNE Visualization

To further illustrate that the deep hashing framework proposed in this paper can effectively improve the discriminability of hash codes, by using the t-SNE technology, we convert the high-dimensional features of multi-label images captured by the network into two-dimensional space. The experimental results are shown in Figure 9. Since multi-label remote sensing data have multiple semantic attributes, it is difficult for us to visualize the embedded representation of all samples. For convenience, we only visualize the hash codes of single-label images of 21 categories in the MLRSNet dataset. As can be seen from Figure 9, compared with the advanced RelaHash and HyP^{2}Loss methods, the DMsH-SL model proposed in this paper can have obvious boundaries between different categories, which further illustrates that the deep multi-similarity hashing framework based on the spatial-enhanced learning module can generate discriminative binary codes, improving the retrieval accuracy of the model.



(**a**) RelaHash (**b**) HyP^{2}Loss (**c**) DMsH-SL

Figure 9. t-SNE visualization of the 16-bit binary codes from RelaHash, HyP2Loss, and DMsH-SL on the MLRSNet dataset.

4.6.3. Top-10 Retrieval Results

To verify the accuracy of the retrieval of the DMsH-SL deep hashing framework proposed in this paper, for a specific query sample, Figure 10 shows the top 10 retrieval results on the UCMerced and DFC15 datasets. Among them, the red box represents data that are extremely dissimilar to the query sample; the blue box represents data that are partially similar to the query sample; and the green box represents data that are extremely similar to the query sample. Compared with state-of-the-art deep hashing, including HyP^{2}Loss OrthoHash, RelaHash, and SWTH, our proposed DMsH-SL framework introduces a spatial group-enhancement mechanism to accurately locate key semantic targets in remote sensing images to generate discriminative visual features. In addition, the multi-similarity loss function fully explores the semantic correlation between sample pairs to construct representative training batches to improve model performance, and its joint label smoothing strategy is used to enhance the discriminability of compact binary encoding. The experimental results show that on the UCMerced dataset, compared with other methods, the proposed DMsH-SL framework returns samples that are partially similar or extremely similar, while on the DFC15 dataset, the retrieval results returned by our deep hashing model are all data images that are extremely similar to the query samples. The above experimental results show that the DMsH-SL framework proposed in this paper has good retrieval performance.

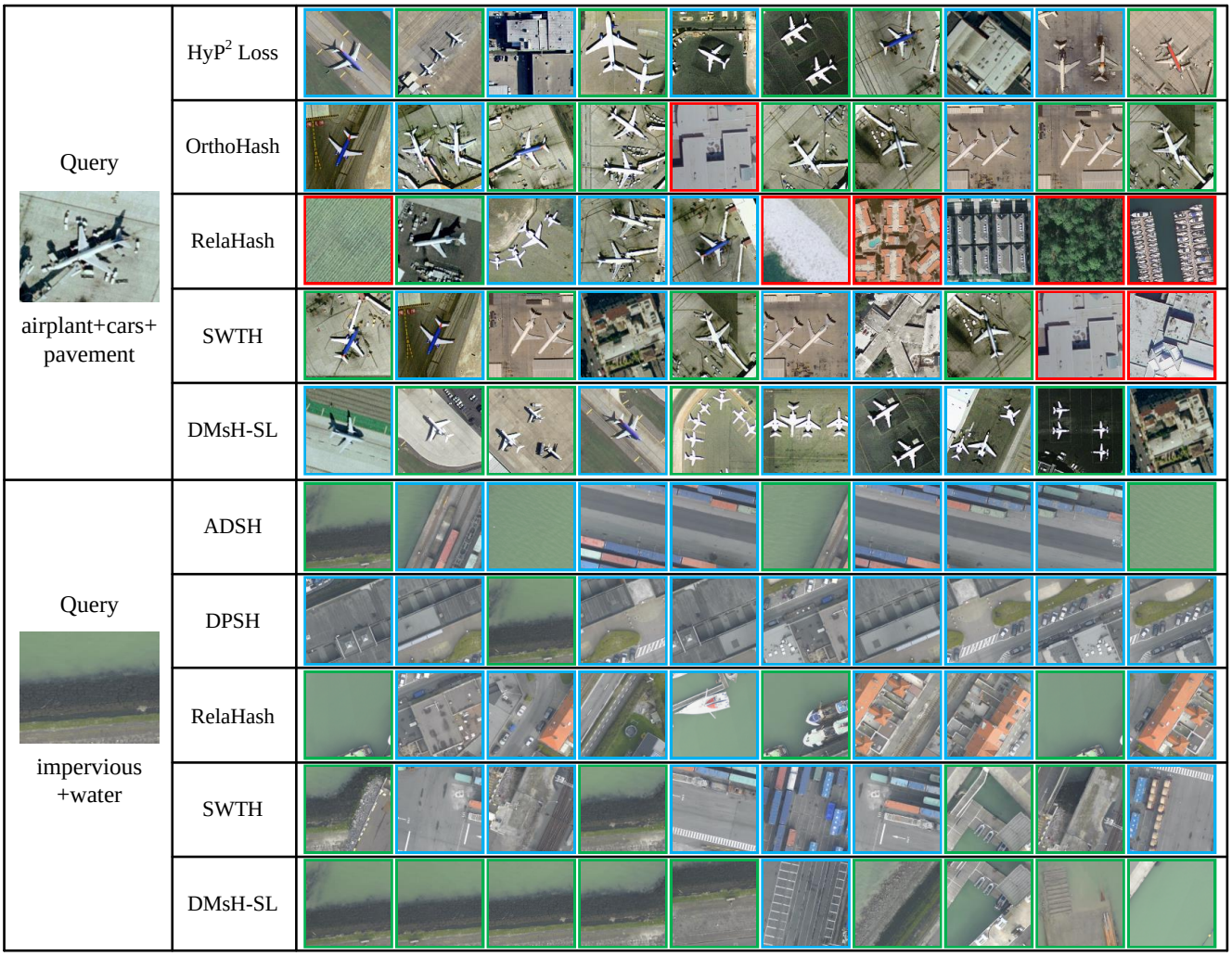


Figure 10. Top-10 ranking results of the DItSH and several baseline methods on UCMerced and DFC15 datasets with respect to 64-bit binary codes. The green boxes mean the retrieved images are completely similar to the query data, the blue boxes represent that the samples share at least one label with the queries, which are called partially similar samples, and the red box denotes that the retrieved samples are dissimilar to the query points.

5. Conclusions and Future Work

In this paper, we propose the novel supervised deep hashing framework, named DMsH-SL, to enhance the accuracy of multi-label remote sensing image retrieval. By integrating a spatial-enhanced mechanism, we accurately highlight the discriminative targets in remote sensing images, mitigating background noise and generating discriminative visual semantic features. Furthermore, by introducing self-similarity, positive relative similarity, and negative relative similarity during model optimization, the multi-similarity loss is incorporated into the deep hashing mainstream to relieve the negative effects of redundant and unbalanced pairs. Extensive experiments on UCMerced, MLRSNet, and DFC15 datasets validate the effectiveness and advantages of the DMsH-SL framework.

Our proposed DMsH-SL framework currently focuses on the scenario of supervised remote sensing image retrieval. However, with the rapid growth of multimedia data from different media types, traditional unimodal retrieval methods are no longer suitable for real-world scenarios. In future work, we aim to extend the spatial-enhanced mechanism to cross-modal remote sensing image retrieval and relieve the negative effects of redundant

and unbalanced pairs. In addition, exploring advanced hashing algorithms that could enhance data retrieval efficiency and accuracy in spatial learning contexts and investigating alternative spatial learning models that could provide additional insights or complementary perspectives are also future research directions.

Author Contributions: Conceptualization, H.Z.; Formal analysis, H.Z. and M.G.; Methodology, H.Z., Q.Q. and M.G.; Writing—original draft, H.Z., Q.Q. and M.G.; Writing—review and editing, H.Z., Q.Q., M.G. and J.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the Shandong Provincial Natural Science Foundation (No. ZR2021MF026) and Weifang Science and Technology Development Plan (No. 2024GX025, No. 2022RKX012).

Data Availability Statement: Experimental results on three benchmark datasets validate the superior performance of our approach, and the source codes are available at https://github.com/QinLab-WFU/DMsH-SL (accessed on 14 November 2024).

Conflicts of Interest: Author Jianyong Huang was employed by the company Qingdao Education Equipment and Information Technology Center. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Li, Y.; Ma, J.; Zhang, Y. Image retrieval from remote sensing big data: A survey. *Inf. Fusion* **2021**, *67*, 94–115. [CrossRef]
2. Ye, Y.; Tang, T.; Zhu, B.; Yang, C.; Li, B.; Hao, S. A multiscale framework with unsupervised learning for remote sensing image registration. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5622215. [CrossRef]
3. Li, J.; Pei, Y.; Zhao, S.; Xiao, R.; Sang, X.; Zhang, C. A review of remote sensing for environmental monitoring in China. *Remote Sens.* **2020**, *12*, 1130. [CrossRef]
4. Kucharczyk, M.; Hugenholtz, C.H. Remote sensing of natural hazard-related disasters with small drones: Global trends, biases, and research opportunities. *Remote Sens. Environ.* **2021**, *264*, 112577. [CrossRef]
5. Ma, Y.; Chen, S.; Ermon, S.; Lobell, D.B. Transfer learning in environmental remote sensing. *Remote Sens. Environ.* **2024**, *301*, 113924. [CrossRef]
6. Wang, S.; Han, W.; Huang, X.; Zhang, X.; Wang, L.; Li, J. Trustworthy remote sensing interpretation: Concepts, technologies, and applications. *ISPRS J. Photogramm. Remote Sens.* **2024**, *209*, 150–172. [CrossRef]
7. Jing, J.; Liu, S.; Wang, G.; Zhang, W.; Sun, C. Recent advances on image edge detection: A comprehensive review. *Neurocomputing* **2022**, *503*, 259–271. [CrossRef]
8. Dubey, S.R. A decade survey of content based image retrieval using deep learning. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 2687–2704. [CrossRef]
9. Chen, H.; Zhu, L.; Zhu, X. Deep Class-guided Hashing for Multi-label Cross-modal Retrieval. *arXiv* **2024**, arXiv:2410.15387.
10. Meng, L.; Zhang, Q.; Yang, R.; Huang, Y. Unsupervised Deep Hashing with Dynamic Pseudo-Multi-Labels for Image Retrieval. *IEEE Signal Process. Lett.* **2024**, *31*, 909–913. [CrossRef]
11. Zhu, L.; Zheng, C.; Guan, W.; Li, J.; Yang, Y.; Shen, H.T. Multi-modal Hashing for Efficient Multimedia Retrieval: A Survey. *IEEE Trans. Knowl. Data Eng.* **2023**, *36*, 239–260. [CrossRef]
12. Hu, H.; Xie, L.; Hong, R.; Tian, Q. Creating something from nothing: Unsupervised knowledge distillation for cross-modal hashing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3123–3132.
13. Zhang, L.; Zhang, L. Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities. *IEEE Geosci. Remote Sens. Mag.* **2022**, *10*, 270–294. [CrossRef]
14. Wang, X.; Han, X.; Huang, W.; Dong, D.; Scott, M.R. Multi-similarity loss with general pair weighting for deep metric learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5022–5030.
15. Zhan, J.; Liu, S.; Mo, Z.; Zhu, Y. Multi-similarity semantic correctional hashing for cross modal retrieval. In Proceedings of the 2020 IEEE International Conference on Multimedia and Expo (ICME), London, UK, 6–10 July 2020; pp. 1–6.
16. Li, X.; Hu, X.; Yang, J. Spatial group-wise enhance: Improving semantic feature learning in convolutional networks. *arXiv* **2019**, arXiv:1905.09646.
17. Song, W.; Li, S.; Benediktsson, J.A. Deep hashing learning for visual and semantic retrieval of remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 9661–9672. [CrossRef]
18. Li, P.; Han, L.; Tao, X.; Zhang, X.; Grecos, C.; Plaza, A.; Ren, P. Hashing nets for hashing: A quantized deep learning to hash framework for remote sensing image retrieval. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 7331–7345. [CrossRef]

19. Liu, C.; Ma, J.; Tang, X.; Liu, F.; Zhang, X.; Jiao, L. Deep hash learning for remote sensing image retrieval. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 3420–3443. [CrossRef]
20. Song, W.; Gao, Z.; Dian, R.; Ghamisi, P.; Zhang, Y.; Benediktsson, J.A. Asymmetric hash code learning for remote sensing image retrieval. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5617514. [CrossRef]
21. Chen, Y.; Wang, F.; Lu, L.; Xiong, S. Unsupervised Transformer Balanced Hashing for Multispectral Remote Sensing Image Retrieval. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 7089–7099. [CrossRef]
22. Fernandez-Beltran, R.; Demir, B.; Pla, F.; Plaza, A. Unsupervised remote sensing image retrieval using probabilistic latent semantic hashing. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 256–260. [CrossRef]
23. Zhu, L.; Lu, X.; Cheng, Z.; Li, J.; Zhang, H. Deep collaborative multi-view hashing for large-scale image search. *IEEE Trans. Image Process.* **2020**, *29*, 4643–4655. [CrossRef]
24. Song, G.; Huang, K.; Su, H.; Song, F.; Yang, M. Deep Ranking Distribution Preserving Hashing for Robust Multi-Label Cross-modal Retrieval. *IEEE Trans. Multimed.* **2024**, *26*, 7027–7042. [CrossRef]
25. Weiss, Y.; Torralba, A.; Fergus, R. Spectral hashing. *Adv. Neural Inf. Process. Syst.* **2008**, *21*.
26. Gong, Y.; Lazebnik, S.; Gordo, A.; Perronnin, F. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 2916–2929. [CrossRef] [PubMed]
27. Kulis, B.; Darrell, T. Learning to hash with binary reconstructive embeddings. *Adv. Neural Inf. Process. Syst.* **2009**, *22*.
28. Liu, W.; Wang, J.; Ji, R.; Jiang, Y.G.; Chang, S.F. Supervised hashing with kernels. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2074–2081.
29. Zhu, X.; Zhang, L.; Huang, Z. A sparse embedding and least variance encoding approach to hashing. *IEEE Trans. Image Process.* **2014**, *23*, 3737–3750. [CrossRef]
30. Zhao, D.; Chen, Y.; Xiong, S. Multi-scale context deep hashing for remote sensing image retrieval. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 7163–7172. [CrossRef]
31. Ye, F.; Wu, K.; Zhang, R.; Wang, M.; Meng, X.; Li, D. Multi-Scale Feature Fusion Based on PVTv2 for Deep Hash Remote Sensing Image Retrieval. *Remote Sens.* **2023**, *15*, 4729. [CrossRef]
32. Sun, Y.; Ye, Y.; Li, X.; Feng, S.; Zhang, B.; Kang, J.; Dai, K. Unsupervised deep hashing through learning soft pseudo label for remote sensing image retrieval. *Knowl. Based Syst.* **2022**, *239*, 107807. [CrossRef]
33. Wang, H.; Zhou, Z.; Zong, H.; Miao, L. Wide-context attention network for remote sensing image retrieval. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 2082–2086. [CrossRef]
34. Yang, H.F.; Lin, K.; Chen, C.S. Supervised learning of semantics-preserving hash via deep convolutional neural networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 437–451. [CrossRef]
35. Yang, Y.; Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010; pp. 270–279.
36. Qi, X.; Zhu, P.; Wang, Y.; Zhang, L.; Peng, J.; Wu, M.; Chen, J.; Zhao, X.; Zang, N.; Mathiopoulos, P.T. MLRSNet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding. *ISPRS J. Photogramm. Remote Sens.* **2020**, *169*, 337–350. [CrossRef]
37. Hua, Y.; Mou, L.; Zhu, X.X. Recurrently exploring class-wise attention in a hybrid convolutional and bidirectional LSTM network for multi-label aerial image classification. *ISPRS J. Photogramm. Remote Sens.* **2019**, *149*, 188–199. [CrossRef] [PubMed]
38. Li, W.J.; Wang, S.; Kang, W.C. Feature learning based deep supervised hashing with pairwise labels. *arXiv* **2015**, arXiv:1511.03855.
39. Jiang, Q.Y.; Li, W.J. Asymmetric deep supervised hashing. In Proceedings of the PAAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
40. Hoe, J.T.; Ng, K.W.; Zhang, T.; Chan, C.S.; Song, Y.Z.; Xiang, T. One loss for all: Deep hashing with a single cosine similarity based learning objective. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 24286–24298.
41. Xu, C.; Chai, Z.; Xu, Z.; Yuan, C.; Fan, Y.; Wang, J. Hyp2 loss: Beyond hypersphere metric space for multi-label image retrieval. In Proceedings of the 30th ACM International Conference on Multimedia, Lisbon, Portugal, 10–14 October 2022; pp. 3173–3184.
42. Doan, K.D.; Yang, P.; Li, P. One loss for quantization: Deep hashing with discrete wasserstein distributional matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 9447–9457.
43. Xu, C.; Chai, Z.; Xu, Z.; Li, H.; Zuo, Q.; Yang, L.; Yuan, C. HHF: Hashing-guided hinge function for deep hashing retrieval. *IEEE Trans. Multimed.* **2022**, *25*, 7428–7440. [CrossRef]
44. Peng, L.; Qian, J.; Wang, C.; Liu, B.; Dong, Y. Swin transformer-based supervised hashing. *Appl. Intell.* **2023**, *53*, 17548–17560. [CrossRef]
45. Liu, H.; Wang, R.; Shan, S.; Chen, X. Deep supervised hashing for fast image retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2064–2072.
46. Movshovitz-Attias, Y.; Toshev, A.; Leung, T.K.; Ioffe, S.; Singh, S. No fuss distance metric learning using proxies. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 360–368.

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