remote sensing

MDPI

Article

A Semantically Guided Deep Supervised Hashing Model for Multi-Label Remote Sensing Image Retrieval

Bowen Liu [1,2], Shibin Liu [1] and Wei Liu [1,*]

[1] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China; liubowen221@mails.ucas.ac.cn (B.L.)
[2] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: liuwei202614@aircas.ac.cn; Tel.: +86-133-6699-7183

Abstract: With the rapid growth of remote sensing data, efficiently managing and retrieving large-scale remote sensing images has become a significant challenge. Specifically, for multi-label image retrieval, single-scale feature extraction methods often fail to capture the rich and complex information inherent in these images. Additionally, the sheer volume of data creates challenges in retrieval efficiency. Furthermore, leveraging semantic information for more accurate retrieval remains an open issue. In this paper, we propose a multi-label remote sensing image retrieval method based on an improved Swin Transformer, called Semantically Guided Deep Supervised Hashing (SGDSH). The method aims to enhance feature extraction capabilities and improve retrieval precision. By utilizing multi-scale information through an end-to-end learning approach with a multi-scale feature fusion module, SGDSH effectively integrates both shallow and deep features. A classification layer is introduced to assist in training the hash codes, incorporating RS image category information to improve retrieval accuracy. The model is optimized for multi-label retrieval through a novel loss function that combines classification loss, pairwise similarity loss, and hash code quantization loss. Experimental results on three publicly available remote sensing datasets, with varying sizes and label distributions, demonstrate that SGDSH outperforms state-of-the-art multi-label hashing methods in terms of average accuracy and weighted average precision. Moreover, SGDSH returns more relevant images with higher label similarity to query images. These findings confirm the effectiveness of SGDSH for large-scale remote sensing image retrieval tasks and provide new insights for future research on multi-label remote sensing image retrieval.

Keywords: remote sensing image retrieval; deep supervised hash; multi-label; similarity measure

check for updates

Academic Editor: Chiman Kwan

Received: 1 December 2024
Revised: 6 February 2025
Accepted: 26 February 2025
Published: 27 February 2025

Citation: Liu, B.; Liu, S.; Liu, W. A Semantically Guided Deep Supervised Hashing Model for Multi-Label Remote Sensing Image Retrieval. *Remote Sens.* **2025**, *17*, 838. https://doi.org/10.3390/rs17050838

Copyright: © 2025 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, the ability to observe the Earth's surface has reached unprecedented levels, with remote sensing satellites generating vast amounts of data daily. Efficiently organizing, managing, and rapidly querying these large datasets to retrieve relevant image data from remote sensing image databases remains a significant challenge [1,2]. Remote sensing image retrieval (RSIR) has garnered increasing attention from the remote sensing community as a key technique for querying and retrieving similar images from large-scale databases [3].

Early text-based image retrieval (TBIR) methods involved manually assigning keywords to each image, allowing users to retrieve corresponding images based on these

keywords [4]. However, the proliferation of remote sensing data makes it impractical to manually annotate each image. In addition, RS images cover large areas and contain multiple features, and the subjective judgment and cognitive biases of annotators can lead to inconsistencies in image descriptions. As a result, accurate retrieval of remote sensing images based solely on keywords is challenging. Content-Based Image Retrieval (CBIR) eliminates the need for metadata or textual descriptions, overcoming the limitations of TBIR by directly analyzing the visual features of images for retrieval tasks [5]. On the one hand, CBIR can support various image processing tasks, such as scene classification [6], target identification and localization [7,8]. On the other hand, it has a wide range of applications across fields such as ecological monitoring [9], disaster assessment [10], and urban management [11], making it a key area of research in remote sensing image retrieval technology.

Content-based remote sensing image retrieval consists of two main components: feature extraction and similarity metrics [12]. Feature extraction focuses on obtaining representative features of remote sensing images, while similarity metrics calculate the similarity between query image and database images to retrieve similar images. Deep learning has been increasingly applied to remote sensing image retrieval due to its powerful feature extraction capabilities. The advanced features extracted by deep learning methods are more stable and generalizable, and extensive research has been conducted on using pre-trained or fine-tuned models to extract the content features of remote sensing images, yielding promising results [13–18]. However, most of these methods assume that both query image and database image are single-labeled. While this approach may be sufficient for certain remote sensing scenarios, such as a beach or forest with a uniform background, images often contain diverse features (e.g., buildings, roads, vegetation, etc.), making it difficult to fully represent image content with a single label. In contrast, multi-label images can describe the information within remote sensing images in greater detail and provide a deeper understanding of image semantics [19]. Many researchers have focused on multi-label remote sensing image retrieval methods. However, most of these studies emphasize improving the feature extraction capabilities of remote sensing images, which often results in high-dimensional features and the problem of "dimensional catastrophe". Additionally, there is relatively little research on similarity metrics for multi-label remote sensing image retrieval [20]. Some studies have even oversimplified the multi-label retrieval problem by assuming that images are similar if they share a common label.

To address the aforementioned challenges, we propose an end-to-end deep hashing network model incorporating semantic information named Semantically Guided Deep Supervised Hashing (SGDSH). For feature extraction, we enhance the Swin Transformer model to obtain multi-scale feature fusion from remote sensing images and reduce feature dimensionality through hash learning. Additionally, the multi-label semantic information of images is fully utilized to guide the learning of hash codes, thereby improving retrieval accuracy. This approach provides an effective solution for large-scale remote sensing image retrieval tasks. The main contributions of this paper can be summarized as follows:

(1) In this paper, we propose a deep hashing network based on Swin Transformer for extracting image features. We design a loss function that integrates image label loss, image pair similarity loss, and hash code quantization loss.

(2) Leveraging the model's ability to recognize image categories, we propose a category-weighted Hamming distance-based ranking strategy to enhance the similarity metric.

(3) We conduct experiments on three public remote sensing datasets of varying sizes. In addition to evaluating the retrieval accuracy of the proposed method across different

dataset sizes, we also compare it with other state-of-the-art methods to demonstrate its effectiveness and superiority.

The remainder of this paper is organized as follows: Section 2 reviews conventional and deep hashing methods for retrieval tasks. Section 3 details the SGDSH model. Experimental results and discussions are presented in Sections 4 and 5. Finally, Section 6 concludes this paper.

2. Related Work

2.1. Multi-Label Remote Sensing Image Retrieval

Multi-label remote sensing image retrieval associates each remote sensing image with multiple labels. However, labeling remote sensing images with multi-label information is a costly task, resulting in a lack of large-scale multi-label remote sensing datasets, which limits research in this area.

Shao et al. [19] released the DLRSD multi-label dataset based on the UCM dataset, extending the annotation of image libraries across 17 categories. They evaluated the performance of both single-label and multi-label remote sensing image retrieval methods, providing a benchmark for multi-label retrieval. Additionally, a novel multi-labeling method based on Fully Convolutional Networks (FCNs) was proposed, marking the first introduction of deep learning into multi-label retrieval tasks [20]. Chaudhuri et al. [21] developed a multi-label RS dataset based on the UCM dataset and proposed a semi-supervised graph-theoretic approach for remote sensing image retrieval. Cheng [22] introduced a multi-label semantic preservation deep hashing model to improve retrieval efficiency, reduce feature storage, and retain semantic information. Jin et al. [23] proposed an interpretable network model for hash retrieval of cloud images, consisting of a feature learning module and a hash learning module, for efficient content-based multi-label retrieval of satellite cloud images. Despite these advancements, the scarcity of large-scale multi-label remote sensing image databases remains a significant challenge. To address this gap, Qi et al. [24] released the MLRS dataset, a large-scale multi-label remote sensing image dataset designed for multi-label classification and retrieval tasks. It contains 109,161 samples across 46 scenes. The existing research primarily focuses on extracting powerful features from RS images, while there is relatively limited work on similarity metrics. Regarding similarity metric computation, Imbriaco et al. [25] proposed a multi-label loss function and reordering method, investigating the impact of commonly used loss functions and reordering techniques on multi-label retrieval performance. Lu et al. [26] introduced a two-stage hierarchical image retrieval (HIR) approach, which incorporates a semantic segmentation module and supplements the similarity metric with visual similarity and semantic statistical information to enhance retrieval accuracy.

The development of deep learning techniques can enable multi-label remote sensing image retrieval to effectively improve retrieval accuracy. The proposed method focuses on enhancing the extraction of high-level semantic features and improving the model's understanding of image content. However, there remains a gap in the research on similarity metrics, and the discussion on the storage and retrieval efficiency of large-scale remote sensing image datasets is still limited.

2.2. Hash-Based Remote Sensing Image Retrieval

Image retrieval tasks typically rely on an Approximate Nearest Neighbor (ANN) search [27], which can lead to significant computational burdens in large-scale data retrieval, especially as deep learning methods tend to produce high-dimensional feature vectors. To mitigate this issue, feature hashing is used to reduce storage requirements and improve computational efficiency by mapping high-dimensional feature vectors to

low-dimensional binary codes [28]. In recent years, numerous hash retrieval methods have been proposed, which can generally be classified into two categories: unsupervised hashing and supervised hashing.

Unsupervised hashing methods rely on unlabeled data to generate binary hash codes for images. Li et al. [29] employed random projection to generate an initial estimate of the hash code, which was subsequently reprojected into the original feature space using a linear model to train the projection matrix and generate a final hash code. Sun et al. [30] proposed an unsupervised deep hashing method based on soft pseudo-labeling, which autonomously generates image soft pseudo-labels and local similarity matrices to facilitate similarity learning among remote sensing images. Although unsupervised hashing methods are simple and effective in generating binary codes for large-scale remote sensing image datasets, their accuracy is relatively limited due to the lack of supervisory information.

In comparison, supervised hashing, particularly methods incorporating deep learning, fully leverages the potential of deep networks and hash learning to improve retrieval accuracy. Li et al. [31] proposed an end-to-end Deep Hash Neural Network (DHNN), which uses hashing to transform high-dimensional features into low-dimensional hash codes. Inspired by generative adversarial networks (GANs), Liu et al. [32] used a uniformly distributed truth matrix as input to the discriminator and designed a loss function with multiple constraints to train the generator and produce high-quality hash codes. Song et al. [33] introduced an asymmetric hash learning method to generate hash codes for querying database images in an asymmetric manner. In deep hash learning, to generate hash codes that effectively distinguish remote sensing images with different contents, a metric learning strategy is often employed. The core objective of this strategy is to maximize the distance between different classes while minimizing the variation within the same class, thus enhancing the distinguishing ability of the deep hash model. Roy et al. [34] proposed a hash network based on metric learning, utilizing a pre-trained deep CNN without fine-tuning, focusing solely on learning the hash function. Cheng et al. [22] introduced a pairwise label similarity loss algorithm that fully exploits multi-label information and demonstrated the effectiveness of this hashing method in multi-label remote sensing image retrieval. Wang et al. [13] proposed a ternary ordered cross-entropy hashing method to address the limitation of most hashing algorithms, which focus only on pointwise or pairwise similarity. Zhou et al. [35] introduced a deep global semantic structure-preserving hashing (DGSSH) method based on triplet loss, optimizing triplet loss and asymmetric constraints in hash learning to generate distinguishable hash codes.

Efficient hash code learning and the utilization of loss functions to retain as much semantic information as possible remain critical challenges in deep hash-based image retrieval research.

3. Proposed Method

The overall framework of Semantically Guided Deep Supervised Hashing (SGDSH) is illustrated in Figure 1. In Section 3.1, we provide a detailed description of the model architecture for our multi-label image retrieval model. Section 3.2 focuses on the design of loss function, and, in Section 3.3, we introduce a similarity metric ranking scheme that incorporates label information.

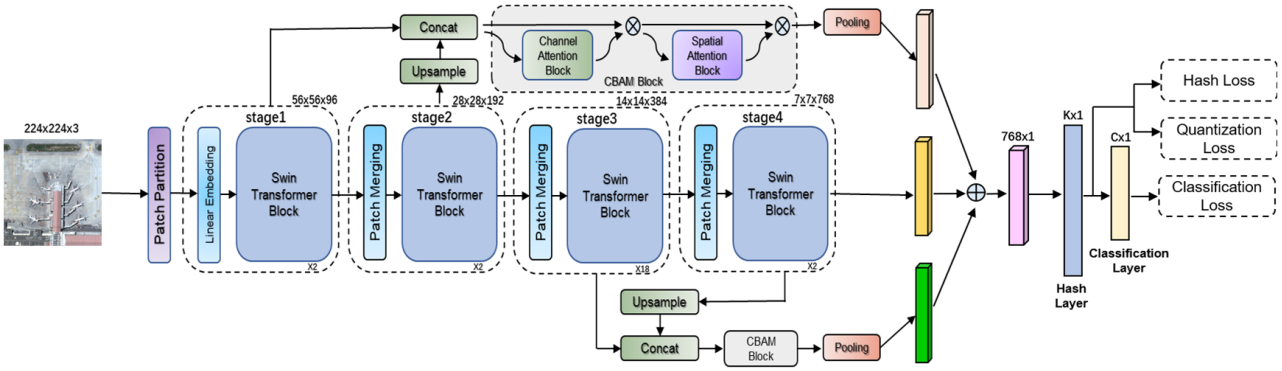


Figure 1. The overall structure of SGDSH. It consists of two main components: deep feature extraction and hash learning. The deep feature extraction module employs an improved Swin Transformer as a backbone for feature extraction. The hash learning network includes a hash layer, followed by a classification layer to assist in hash code training. The hash learning process integrates hash loss, quantization loss, and classification loss to effectively learn the hash codes.

3.1. Overall Model Framework

3.1.1. Feature Extraction Module

In order to extract multi-scale feature information and global contextual dependencies from remote sensing images, we propose a multi-scale feature fusion network based on Swin Transformer [36]. As shown in Figure 2, the Swin Transformer module consists of a W-MSA module, a SW-MSA module, and two MLPs, which capture the dependencies of input sequences through a self-attention mechanism. The self-attention operation is expressed in Equation (1):

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{1}$$

where Q, K, and V denote query, key, and value matrices; $Softmax$ denotes the normalization function; and d denotes the dimension of the K vector.

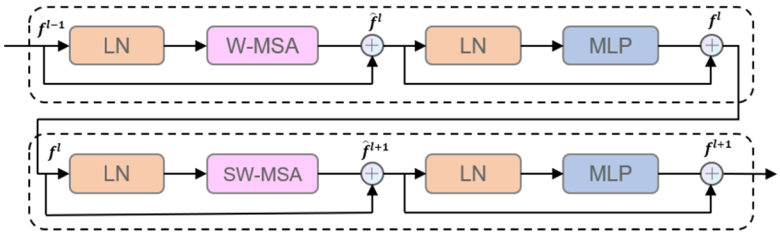


Figure 2. The structure of the Swin Transformer Block.

The image features pass through Swin Transformer module with the following equation:

$$\hat{f}^l = WMSA\left(LN\left(f^{l-1}\right)\right) + f^{l-1} \tag{2}$$

$$f^l = MLP\left(LN\left(\hat{f}^l\right)\right) + f^l \tag{3}$$

$$\hat{f}^{l+1} = SWMSA\left(LN\left(f^l\right)\right) + f^l \tag{4}$$

$$f^{l+1} = MLP\left(LN\left(\hat{f}^{l+1}\right)\right) + f^{l+1} \tag{5}$$

where $\hat{f}^l$ and f^l are output features through the W-MSA module and MLP module in block l; $\hat{f}^{l+1}$ and f^{l+1} are output features through the SW-MSA module and MLP module in block $l+1$.

Given an input remote sensing image I, feature extraction is performed using a pre-trained Swin Transformer model. The image and its generated feature maps exhibit multi-scale properties across different stages of the model's output. The feature maps from the final stage are then flattened to obtain a global feature vector f^g of the image I.

The mathematical expression of f^g is shown in Equation (6):

$$f^g = Swin(I, \theta) \tag{6}$$

where θ denotes the parameters of network.

Due to the rich feature content and feature variability at different scales in remote sensing images, it is challenging to comprehensively capture multi-scale information of the targets with only the final output of the network. Additionally, as the number of network layers increases, the receptive field expands, which can reduce the model's ability to detect small targets and increase the risk of information loss. Considering these factors, we consider global features f^g alone cannot fully represent all the information contained in the RS image. To address this, we propose a multi-scale feature fusion module. This module extracts features from each stage of the network and performs multi-scale feature fusion. The specific process is as follows:

$$\left\{f^i\right\}_{i=1}^{4} = Swin(I, \theta) \tag{7}$$

$$f^l = Concat\left(f^1, Upsample\left(f^2\right)\right) \tag{8}$$

$$f^h = Concat\left(f^3, Upsample\left(f^4\right)\right) \tag{9}$$

where $\left\{f^i\right\}_{i=1}^{4}$ denotes image features output from ith stage; $Concat()$ and $Upsample()$ represent concatenation and upsampling operations; f^l denotes low-level image features; and f^h denotes high-level feature generated from higher stages.

We successfully obtain two key features: high-level feature f^h, enriched with global semantic information of the image, and low-level feature f^l, which captures rich local details. To further enhance the representational capability of these features, we introduce the Convolutional Block Attention Module (CBAM) [37]. By employing the CBAM module, we apply a weighting mechanism to the feature map, emphasizing important feature information while reducing the influence of background noise. This allows the model to focus its attention on the most informative areas of the image. The CBAM module can be summarized by the following equation:

$$F_2 = M_c(F_1) \bigotimes F_1 \tag{10}$$

$$F_3 = M_s(F_2) \bigotimes F_2 \tag{11}$$

where F_1 denotes the input feature map; F_2 represents intermediate features generated by channel attention; F_3 refers to output features generated by spatial attention; M_c and M_s denote the channel attention module and spatial attention module. The structure is shown in Figures 3 and 4.

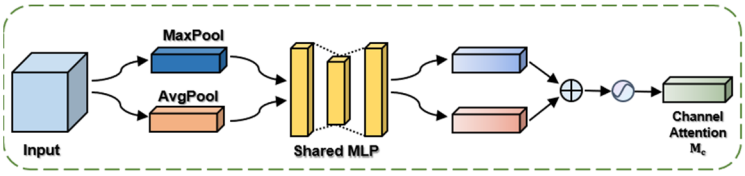


Figure 3. The structure of the channel attention module. The input feature is downsampled using global maximum pooling and global average pooling, resulting in two $1 \times 1 \times C$ feature maps. These are then passed through two fully connected layers, each producing a $1 \times 1 \times C$ feature map. Finally, the two feature maps are summed and passed through a sigmoid activation function, constraining output values to be between 0 and 1.

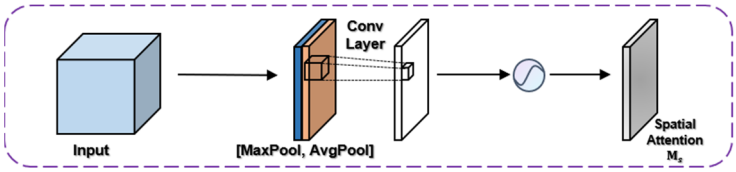


Figure 4. The structure of the spatial attention module. The input feature map is downsampled using global max pooling and global average pooling, producing two feature maps of size $H \times W \times 1$. These two feature maps are then concatenated along the channel dimension, resulting in a feature map of size $H \times W \times 2$. Next, a convolution operation is applied to transform this feature map into a $H \times W \times 1$ map. Finally, a sigmoid activation function is used to constrain the values of the feature map between 0 and 1.

After processing by CBAM module, the updated expressions for the low-level features and high-level features are as follows:

$$F_{low} = CBAM\left(f^{l}\right) \tag{12}$$

$$F_{high} = CBAM\left(f^{h}\right) \tag{13}$$

We obtain the high-level feature F_{high} and low-level feature F_{low} and then perform pooling operations and fusion with the global feature f^{g} to generate the final image features. This feature fusion process is carried out by the Multi-Scale Feature Fusion Module (MSFFM), which is expressed as follows:

$$F = Pool\left(F_{high}\right) + Pool\left(F_{high}\right) + f^{g} \tag{14}$$

With this feature fusion strategy, the model is able to more accurately capture and interpret the complex scenes and fine-grained details within remote sensing images, thereby providing robust and detailed feature representations for remote sensing image retrieval tasks.

3.1.2. Hashing Layer

To reduce the storage space required for image features and enable fast retrieval from large remote sensing databases, we employ a hashing method to compress the high-dimensional image features. After the image passes through feature extraction network to obtain the feature F, it is then processed by a fully connected layer to produce the hash feature $h \in \mathbb{R}^{K}$, where K denotes the length of the hash code. This process is expressed by Equation (15):

$$h = tanh(F) \tag{15}$$

where the tanh activation function maps the feature vector values between -1 and 1.

The classification labels of an image typically capture its overall semantic content, particularly for multi-labeled images. These labels can serve as valuable semantic clues

to assist network training and facilitate the generation of hash codes that contain richer semantic information. To leverage this, we introduce a classification layer after the hash layer, with the number of neurons set to match the predefined number of C categories. The classification result, denoted as $classifier \in \mathbb{H}^C$, is obtained through a fully connected layer. A sigmoid activation function is then applied to predict the probability that the image belongs to each of C categories. This process can be expressed by Equation (16):

$$classifier = \sigma(F_c(h)) \tag{16}$$

where F_c denotes the fully connected layer, and the sigmoid activation function maps category vector values between 0 and 1.

3.2. Loss Function Design

3.2.1. Pairwise Similarity Loss

The primary goal of hash learning is to learn a set of nonlinear hash functions that map data from the original feature space to a binary hash code, while preserving semantic similarity between data points. Specifically, the objective is to make samples from the same category closer in Hamming distance, while pushing samples from different categories further apart. In hash learning of single-label remote sensing images, the similarity between two images is determined by whether their category labels are the same. However, for multi-label remote sensing images, the assumption that two images are similar simply because they share a label does not fully capture the nuanced similarity between images. Inspired by [25,38], we propose a method based on image label similarity to better reflect the true similarity relationship. Our approach calculates the label similarity between pairs of images, dividing the samples into positive and negative pairs. This label similarity is then used to weight the loss function, contributing more effectively to the learning process. The label similarity is computed as follows:

$$S_{ij} = \frac{{L_i}^T \cdot L_j}{\|L_i\| \cdot \|L_j\|} \tag{17}$$

where L_i and L_j denote the label vectors of image i and image j, $\| \ \|$ and denotes the L2 paradigm operation.

According to the definition in [25], two images are considered similar if the similarity of their labels $S_{ij} > 0.4$. Otherwise, images are deemed dissimilar, and the distance between their hash codes should be maximized. This can be expressed as follows:

$$I(i,j) = \begin{cases} 1, & S_{ij} \geq 0.4 \\ 0, & S_{ij} < 0.4 \end{cases} \tag{18}$$

In addition, the Hamming distance between the hash codes of two images is computed, with label similarity serving as a weighted loss term. This ensures that image pairs with a higher number of shared labels contribute more heavily to the loss function. This design encourages the model to better capture the semantic similarity between images and learn more accurate hash codes. Specifically, our loss function is defined as follows:

$$L_{cont} = I(i,j) \cdot \left(Dis(h_i, h_j) - \frac{1}{2}(1 - S_{ij}) \right)^2 + (1 - I(i,j)) \cdot \left(max(margin - Dis(h_i, h_j), 0) \right)^2 \cdot (1 - S_{ij}) \tag{19}$$

where $Dis(h_i, h_j)$ denotes the Hamming distance between hash codes extracted from image i and image j. *margin* is a hyperparameter indicating the minimum distance between pairs of negative samples.

The Hamming distance between image pairs is calculated as follows:

$$Dis(h_i, h_j) = \frac{{h_i}^T \cdot h_j}{\|h_i\| \cdot \|h_j\|} \tag{20}$$

According to the proposed contrast loss function, the trained model minimizes the difference between the hash codes of remote sensing image pairs with more common labels, indicating greater similarity between them. Conversely, for dissimilar images with lower label similarity, the corresponding hash codes are pushed further apart, ensuring proper differentiation.

3.2.2. Quantization Loss

Ideally, the hash code **b** should be compact and consist of discretized 0 s and 1 s. However, the discrete nature of the hash code can cause the gradient to vanish during model training, hindering backpropagation and parameter updates. To address this issue, we use the floating-point vector **u** output from the hash layer to approximate the hash code **b** during model training. Additionally, we introduce a quantization loss function that encourages each element in **u** to move closer to either -1 or 1. The quantization loss function is defined as follows:

$$L_{quan} = \frac{1}{N}\sum_{i=1}^{N}||u_i| - 1| \tag{21}$$

3.2.3. Classification Loss

For multi-label classification problems, the cross-entropy loss function is the most commonly used objective function. In a multi-category prediction scenario, a sigmoid activation function is used instead of a softmax function, transforming the problem into a set of binary classification tasks for each category. The objective function is then constructed by applying cross-entropy loss to the binomial distribution of each category label. Thus, the classification loss function is defined as follows:

$$L_{cls} = -(y \cdot log(\hat{p}) + (1 - y) \cdot log(1 - \hat{p})) \tag{22}$$

where y represents the actual probability of the image belonging to that class, taking values of 1 or 0, and $\hat{p}$ denotes the predicted probability of the model for the image belonging to that class.

For the retrieval task of multi-label remote sensing image datasets, this study aims not only to ensure the accuracy of retrieval results but also to prioritize returning images that share more labels with the query image. To achieve this, we propose a semantic-guided deep supervised hashing method that fully leverages image semantic information. This approach efficiently integrates deep hash learning with the multi-label classification task by designing an objective function that simultaneously considers both the multi-label semantic information of remote sensing images and the similarity features between image pairs. Specifically, the objective function of this paper is defined as follows:

$$\begin{aligned}\mathcal{L}_{total} &= \alpha \cdot \mathcal{L}_{cls} + \beta \cdot \mathcal{L}_{cont} + \gamma \cdot \mathcal{L}_{quan} \\ &= -\alpha \cdot (y \cdot log(\hat{p}) + (1 - y) \cdot log(1 - \hat{p})) + \beta \\ &\cdot \Bigg(I(i,j) \cdot \Big(Dis(h_i, h_j) - \tfrac{1}{2}(1 - S_{ij}) \Big)^2 + (1 - I(i,j)) \cdot \big(max\big(margin - Dis(h_i, h_j), 0\big) \big)^2 \\ &\cdot \big(1 - S_{ij}\big) \Bigg) + \gamma \cdot \tfrac{1}{N} \sum_{i=1}^{N} ||u_i| - 1|\end{aligned} \tag{23}$$

where $\mathcal{L}_{cls}$ is multi-label classification loss; $\mathcal{L}_{cont}$ is pairwise similarity loss; $\mathcal{L}_{quan}$ is quantization loss; and α, β, and γ are sensitivity parameters that control three parts of loss, respectively.

3.3. Similarity Metric Search

A set of query images is defined as $Q = \{q_i\}_{i=1}^{N}$, and corresponding hash codes $B_q = \left\{b_i^q\right\}_{i=1}^{N}$ are obtained for each image after passing through a deep hash network. During the retrieval phase, the similarity between the hash code of the query image and all images in the dataset is measured by calculating the Hamming distance between them. The results of the Hamming distance calculations are then sorted to determine the set of images most similar to the query image. The Hamming distance between hash codes is computed as follows:

$$Dis_H(b_i, b_j) = \sum b_i \oplus b_j \tag{24}$$

In Hamming distance-based retrieval, the quality of the retrieval results heavily depends on the quality of learned hash codes. However, the lack of an intuitive understanding of how these hash codes represent image content makes the results from Hamming distance ranking difficult to interpret. Additionally, the retrieval database often contains a large volume of data, and, since Hamming distance measures the sum of dissimilar bits between two hash codes, it can produce many results with identical Hamming distances. This creates the need for finer ordering of images that have similar Hamming distances. Furthermore, differences between images are not only reflected in the feature vector values but also in the overall category differences. Yet, most methods do not fully leverage the learned category information during retrieval.

To address these challenges, this paper proposes an Improved Similarity Measurement Strategy (ISMS), which aims to better utilize the category information learned by the model and reduce the distance between images belonging to same category. Specifically, after the query image passes through the network, we obtain its hash code b_i as well as the predicted category label vector L_i. In the similarity calculation, the label similarity between the query image's category label and the image to be retrieved is computed using the Jaccard coefficient as a distance weight. The Jaccard coefficient is calculated using the following formula:

$$Jaccard(L_i, L_j) = \frac{|L_i \cap L_j|}{|L_i \cup L_j|} \tag{25}$$

The Hamming distance between two hash codes is finally weighted as the final similarity result, and the similarity metric expression is defined as follows:

$$Dis(q_i,\ q_j) = Dis_H(b_i, b_j) \cdot e^{1 - Jaccard(L_i, L_j)} \tag{26}$$

4. Experiments and Analysis

To evaluate the effectiveness of proposed method, several experiments are conducted in this section. First, the datasets and experimental environments are introduced, followed by the implementation details of the various experiments. Finally, the results are presented, analyzed, and discussed.

4.1. Datasets and Evaluation Indicators

4.1.1. Datasets

In this paper, we utilized three widely used publicly available multi-labeled remote sensing datasets: the UC Merced multi-labeled dataset [21], AID multi-labeled dataset [39], and MLRS dataset [24].

The UCM-multi dataset is labeled with 17 categories based on the UC Merced remote sensing image dataset [40], which contains 21 geographic categories, with 100 remote sensing images per category, totaling 2100 images. The image size is 256 × 256 pixels, and the spatial resolution is 30 cm/pixel.

The AID-multi dataset is an enhanced version of the AID dataset [41], containing 3000 aerial images selected from 30 scenes in the AID dataset. Multiple object labels are manually assigned to each image through visual inspection. The image resolution for different object categories ranges from 0.5 m/pixel to 8 m/pixel, and each image has a size of 600 × 600 pixels, making it more challenging compared to the UCM-multi dataset.

The MLRS dataset consists of 109,161 labeled RGB images obtained from Google Earth, annotated into 46 broad categories. The number of images varies across categories, ranging from 1500 to 3000 images per category. Each image is assigned a subset of 60 pre-defined category labels, with the number of labels per image ranging from 1 to 13. The spatial resolution of the images ranges from 0.1 m to 10 m, making it ideal for data-driven deep learning research.

4.1.2. Evaluation Indicators

In our experiments, we used four commonly used retrieval evaluation metrics to assess the performance of multi-label remote sensing image retrieval model: mean average precision (MAP), normalized discounted cumulative gain (NDCG), average cumulative gain (ACG), and weighted average precision (WAP).

MAP measures mean of average precision (AP) across all query images, where AP represents average precision for each query. MAP is calculated as follows:

$$MAP = \frac{1}{Q}\sum_{i=1}^{Q} AP_i \tag{27}$$

where Q denotes the number of query images. Average precision (AP) is defined as follows:

$$AP_i = \frac{1}{R(i,n)}\sum_{j=1}^{n} \frac{I(i,j) \times R(i,j)}{j} \tag{28}$$

where $I(i,j)$ is an indicative function; if label similarity between ith query image and jth image is greater than 0.4, then $I(i,j) = 1$; otherwise, it is zero.

NDCG is well suited for multi-label retrieval problems because the metric measures the relevance of the top-ranked results with the following formula:

$$NDCG@n = \frac{DCG@n}{IDCG@n} \tag{29}$$

where DCG is calculated as follows:

$$DCG@n = \sum_{i=1}^{n} \frac{2^{J(q,\,i)} - 1}{log_2(1+i)} \tag{30}$$

where $J(q,\ i)$ denotes the number of public labels of the query image q and ith retrieved image.

$IDCG@n$ denotes $DCG@n$ in the ideal state, calculated as follows:

$$IDCG@n = \sum_{i=1}^{n} \frac{2^{J(q,\,i)_r} - 1}{log_2(1+i)} \tag{31}$$

where $J(q,i)_r$ denotes the number of common labels of the query image q and ith retrieved image in the ideal state.

ACG describes the average similarity between query image and first n retrieved images, calculated as follows:

$$ACG@n = \frac{1}{n}\sum_{i=1}^{n} \frac{J(q,i)}{c} \tag{32}$$

where c denotes the number of labels for retrieving image q.

WAP is a variant of MAP which can be computed based on ACG and is more accurate for evaluating the retrieval accuracy of the model. The WAP of first n images obtained from retrieval is calculated as follows:

$$WAP@n = \frac{1}{Q}\sum_{i=1}^{Q} \frac{1}{R(n)} \sum_{j=1}^{n} I(i,j)ACG@j \tag{33}$$

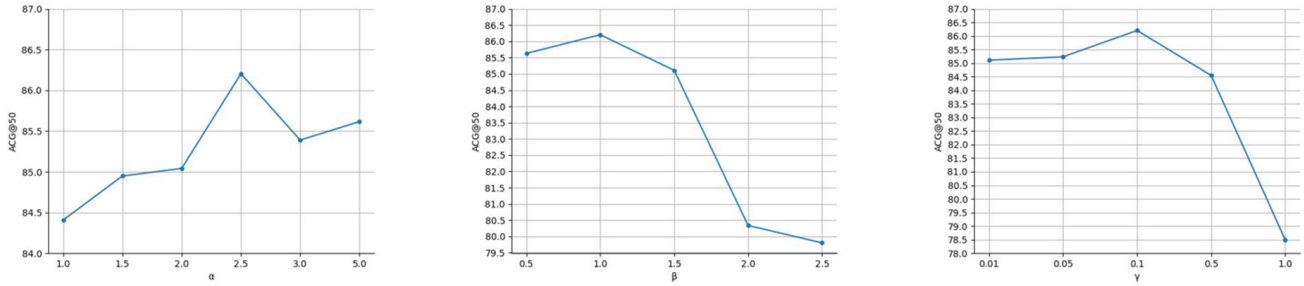
4.2. Experimental Settings

To validate the effectiveness of the proposed method, we conducted experiments on three publicly available multi-labeled remote sensing image datasets. In each experiment, the dataset was split into two parts: 80% of images were used to construct the training set, while the remaining 20% were used as the validation set. In the retrieval phase, all images in UCM-multi and AID-multi datasets were used to form the retrieval database, with each image serving as a query instance. For the MLRS dataset, the validation set was used as the retrieval database, and 10% of images were randomly selected to serve as query examples.

The experiments in this paper were conducted on the PyTorch platform, with the software environment consisting of Ubuntu 5.4, CUDA-11.4, PyTorch-1.10.0, and Python 3.7, and accelerated by NVIDIA A100. During the model training, all the input images were resized to 224 × 224. A learning rate decay strategy was applied with a decay rate of 0.05 and an initial learning rate of 0.0001. The batch size was set to 128, and AdamW was used for optimization. The number of epochs was set to 200 for both the UCM-multi and AID-multi datasets, and 150 for the MLRS dataset.

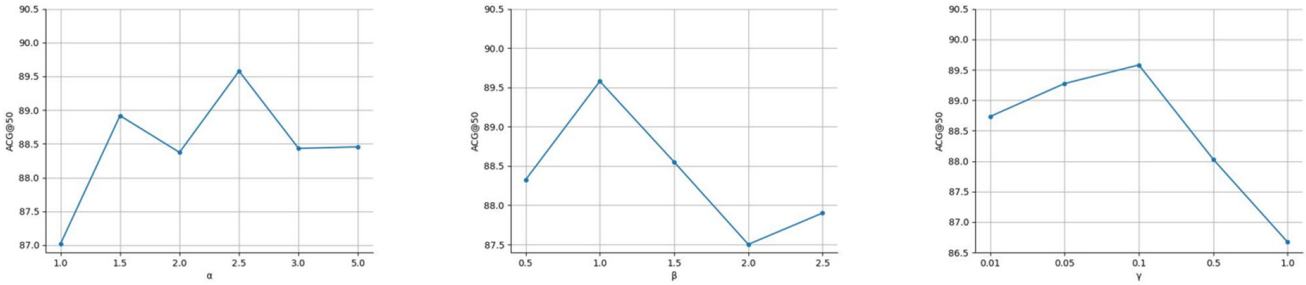
4.3. Parameter Analysis

As shown in Equation (23), our loss function consists of three components: classification loss, pairwise similarity loss, and quantization loss, each associated with penalty coefficients α, β, and γ, respectively. To deeply analyze the impact of these parameters on retrieval performance, we conducted experiments on the UCM-multi, AID-multi, and MLRS datasets. In these experiments, the hash code length was set to 128 bits, and ACG@50 was used as the evaluation metric for retrieval results. During analysis, when the effect of a particular parameter is evaluated, the other two parameters are kept fixed. Figure 5a,b,c, Figure 6a,b,c, and Figure 7a,b,c show the variation in retrieval results as the values of α, β, and γ change on these three datasets, respectively.



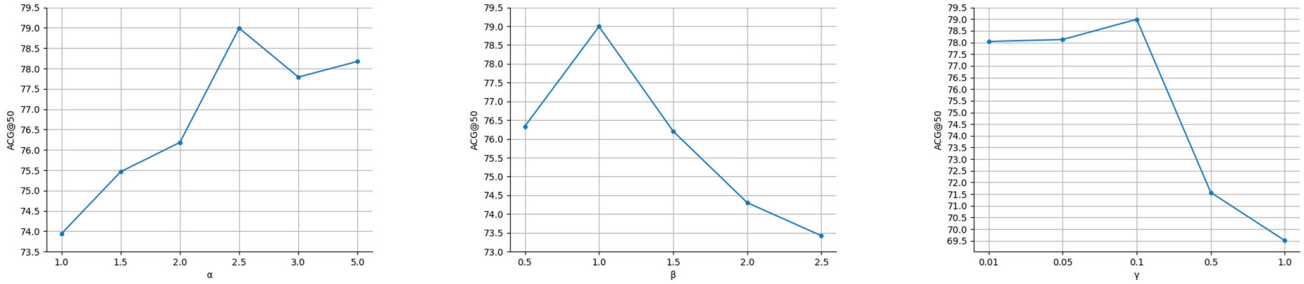
(**a**) (**b**) (**c**)

Figure 5. Effects of ACG@50 on UCM-multi dataset with different sensitivity coefficients. (**a**) Effect of α on ACG. (**b**) Effect of β on ACG. (**c**) Effect of γ on ACG.



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

Figure 6. Effects of ACG@50 on AID-multi with different sensitivity coefficients. (**a**) Effect of α on ACG. (**b**) Effect of β on ACG. (**c**) Effect of γ on ACG.



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

Figure 7. Effects of ACG@50 on MLRS with different sensitivity coefficients. (**a**) Effect of α on ACG. (**b**) Effect of β on ACG. (**c**) Effect of γ on ACG.

We first determined the range of variation for penalty coefficient α, fixing β and γ at 1.0 and 0.1, respectively, based on empirical observation. By analyzing the trends in Figures 5a, 6a, and 7a, we observed that model achieves optimal retrieval performance when α is set to 2.5 for all three datasets. Additionally, when α and γ are set to 2.5 and 0.1, respectively, the best retrieval results are obtained when β is set to 1.0. As β increases, the model's retrieval accuracy begins to decrease, particularly on the MLRS and UCM-multi dataset. For sensitivity analysis of quantization loss, we tested a set of empirical values for γ: {0.01, 0.05, 0.1, 0.1, 1.0}, and the retrieval accuracy reaches its peak when γ is set to 0.1 in the three datasets, as shown in Figures 5c, 6c, and 7c. Further increases in the weight of γ lead to a substantial decline in accuracy. This suggests that the weight of quantization loss should be carefully tuned to avoid over-emphasizing the quantization effect on hash codes.

Based on the above analyses, we determined that penalty coefficients α, β, and γ should be set to 2.5, 1.0, and 0.1, respectively. These parameter values strike a balance for the model, ensuring high classification accuracy of hash codes while maintaining an

appropriate trade-off between pairwise similarity loss and quantization loss. This optimal configuration enables the model to achieve the best retrieval performance for the task of multi-label remote sensing image retrieval.

4.4. Ablation Experiment

After determining the optimal weights of each component of the loss function, we conducted a series of ablation experiments on three datasets. These experiments aim to not only validate the contribution of each module to overall accuracy in different datasets but also to assess model's retrieval performance when handling a large-scale remote sensing dataset (the MLRS dataset being retrieved in the ablation experiments includes more than 20,000 images). The proposed model consists of two core modules: a feature extraction module and a hash learning module. In the feature extraction module, we incorporate a fusion strategy that combines shallow and deep features to enhance feature representation. In the hash learning module, we leverage image category information learned by the model and integrate it into the similarity metric to improve retrieval performance.

We progressively integrated these modules into the baseline model and constructed four configurations for comparison: (i) Extracting image features using the Swin Transformer base model and relying on hash codes in the similarity metric; (ii) Using the Swin Transformer base model and weighting the Hamming distance by incorporating category labeling information; (iii) Using the improved Swin Transformer model with Hamming distance; (iv) Extracting image features using the improved Swin Transformer model and weighting Hamming distance by combining category label information.

We evaluated the performance using NDCG, ACG, and WAP metrics, focusing on the top 10 and top 50 retrieved images. Final experimental results, presented in Tables 1–3, illustrate the quantitative impact of different model components on retrieval performance across the three datasets.

Table 1. Retrieve results of the ablation experiments on the UCM-multi dataset. The best values are highlighted in bold.

Method			NDCG@10	ACG@10	WAP@10	NDCG@50	ACG@50	WAP@50
Swin	MSFFM	ISMS						
√			99.186	93.365	93.102	98.926	84.607	83.796
√		√	99.411	94.105	93.826	99.144	86.161	85.291
√	√		99.359	93.435	93.168	98.995	84.747	83.919
√	√	√	**99.615**	**94.148**	**93.924**	**99.172**	**86.203**	**85.438**

Table 2. Retrieve results of ablation experiments on AID-multi dataset. The best values are highlighted in bold.

Method			NDCG@10	ACG@10	WAP@10	NDCG@50	ACG@50	WAP@50
Swin	MSFFM	ISMS						
√			98.793	92.561	92.126	98.696	88.299	87.496
√		√	98.875	92.734	92.394	98.864	88.863	88.224
√	√		99.272	93.143	93.143	98.917	89.269	88.748
√	√	√	**99.351**	**93.388**	**93.178**	**99.074**	**89.579**	**89.042**

By analyzing the experimental results, we find that the proposed method achieves satisfactory performance in the retrieval task, with all components of the model contributing positively to the results. Specifically, feature extraction capability is enhanced by the feature fusion module compared to the base Swin Transformer model. For the top 10 retrieved

images, the accuracy of NDCG, ACG, and WAP improved by 0.2%, 0.1%, and 0.1% on the UCM-multi dataset; 0.5%, 0.6%, and 0.9% on the AID-multi dataset; and 0.2%, 0.4%, and 0.5% on the MLRS dataset. For the top 50 retrieved images, these metrics showed improvements of 0.1%, 0.1%, and 0.1% on the UCM-multi dataset; 0.2%, 0.8%, and 1.0% on the AID-multi dataset; and 0.2%, 0.6%, and 0.8% on the MLRS dataset.

Table 3. Retrieve results of ablation experiments on MLRS dataset. The best values are highlighted in bold.

Method			NDCG@10	ACG@10	WAP@10	NDCG@50	ACG@50	WAP@50
Swin	**MSFFM**	**ISMS**						
√			95.196	78.775	76.031	94.376	75.894	72.079
√		√	96.242	81.244	79.242	95.268	78.389	75.304
√	√		95.357	78.931	76.358	94.512	76.431	72.829
√	√	√	**96.413**	**81.824**	**79.975**	**95.471**	**78.992**	**76.194**

Furthermore, by incorporating model-predicted image category information into the similarity metric, retrieval accuracy improves even more significantly, especially on the MLRS dataset. For the top 10 retrieved images, the accuracy increased by 0.2%, 0.7%, and 0.7% on the UCM-multi dataset; 0.1%, 0.2%, and 0.2% on the AID-multi dataset; and 1.1%, 2.7%, and 3.4% on the MLRS dataset. For the top 50 retrieved images, the accuracy improved by 0.2%, 1.5%, and 1.5% on the UCM-multi dataset; 0.2%, 0.4%, and 0.5% on the AID-multi dataset; and 1.0%, 2.5%, and 3.3% on the MLRS dataset. These results demonstrate that the inclusion of inter-image category information can effectively enhance ACG, with the improvement being particularly evident in large-scale data retrieval, indicating that retrieved images share more common labels with the query image, thereby contributing to a substantial increase in WAP.

The experimental results demonstrate that the multi-feature fusion module proposed in this paper enhances the feature extraction capability of remote sensing images compared to the basic Swin Transformer model. Additionally, by incorporating category label information into the similarity metric and weighting the Hamming distance according to label similarity, the number of common labels between the query image and retrieved images increases. This ensures that retrieved images share more similar visual content with the query image, thereby improving retrieval accuracy.

4.5. Comparative Experiments

To comprehensively evaluate effectiveness of proposed method, we compare it with several existing multi-label retrieval methods, including HyP [42], DMHR [22], IDHN [43], and FAH [16], all of which use pairwise similarity to learn the relationships between image pairs. The comparisons are conducted on the UCM-multi and AID-multi datasets, using three different hash code lengths to assess the performance of retrieving the top 10 and top 50 most similar images.

Experimental results for UCM-multi dataset are presented in Tables 4 and 5. The data clearly demonstrate that the SGDSH model outperforms comparison methods across all hash code lengths. Specifically, when hash code length is set to 96 or 128, our method achieves superior results. For the task of retrieving the top 10 most similar images, MAP reaches 99.623%, outperforming HyP by 0.237%, DMHR by 3.491%, IDHN by 10.236%, and FAH by 28.916%. In the task of retrieving the top 50 images, MAP reaches 98.688%, which is 0.347% higher than HyP, 8.57% higher than DMHR, 16.947% higher than IDHN, and 32.657% higher than FAH. Figures 8 and 9 present the performance results of each method across different hash code lengths on the UCM-multi dataset.

Table 4. UCM-multi dataset retrieval of the top 10 images comparison experiment results. The best values are highlighted in bold.

Method	Bits	MAP	NDCG@10	ACG@10	WAP@10
Our method	64	98.824	98.803	91.793	91.455
	96	99.493	99.521	93.749	93.384
	128	**99.623**	**99.615**	**94.148**	**93.924**
HyP	64	98.307	98.021	91.171	90.749
	96	99.436	99.115	91.665	91.327
	128	99.386	99.094	91.475	91.097
DMHR	64	95.177	96.872	76.141	73.447
	96	95.639	97.137	77.528	75.095
	128	96.132	97.288	78.177	76.019
IDHN	64	89.422	94.283	69.584	63.919
	96	90.175	94.685	70.198	64.879
	128	89.387	94.321	68.435	62.769
FAH	64	59.077	85.744	53.523	40.468
	96	62.442	80.749	49.085	38.453
	128	70.707	88.715	56.549	45.516

Table 5. UCM-multi dataset retrieval of the top 50 images comparison experiment results. The best values are highlighted in bold.

Method	Bits	MAP	NDCG@50	ACG@50	WAP@50
Our method	64	96.904	98.235	83.837	82.618
	96	**98.688**	99.137	86.038	85.292
	128	98.642	**99.172**	**86.203**	**85.438**
HyP	64	95.955	98.141	83.136	82.003
	96	98.341	98.354	84.022	83.099
	128	98.232	98.311	83.746	82.753
DMHR	64	88.551	94.334	69.411	63.995
	96	89.286	94.708	70.115	65.038
	128	90.118	94.843	70.551	65.874
IDHN	64	81.382	92.109	64.774	55.959
	96	81.741	92.336	64.221	55.585
	128	80.661	91.731	61.902	52.905
FAH	64	60.427	87.772	53.069	38.115
	96	58.309	82.098	46.024	31.893
	128	66.031	86.811	52.985	39.609

Experimental results on the AID-multi dataset, shown in Tables 6 and 7, demonstrate a performance similar to that observed on the UCM-multi dataset, with the SGDSH method outperforming all other methods when hash code length is set to 128 bits. As the dataset size increases, particularly with the variation in spatial resolution across different images, the gap between the SGDSH method and the other methods becomes more pronounced. Notably, the HyP method remains competitive, maintaining a similar mean average precision (mAP) to SGDSH on the AID-multi dataset. Specifically, for the task of retrieving the top 10 most similar images, SGDSH achieves a MAP of 99.566%, surpassing the other methods by 0.492% over HyP, 2.459% over DMHR, 20.522% over IDHN, and 21.733% over FAH. In the task of retrieving the top 50 most similar images, SGDSH reaches a MAP of 98.932%, which is 0.888% higher than HyP, 4.938% higher than DMHR, 20.626% higher than FAH, and 21.094% higher than IDHN. In terms of ACG, SGDSH shows an even more

significant superiority on this dataset compared to the UCM-multi dataset. For the top 10 and top 50 image retrieval tasks, the ACG results of SGDSH are higher than HyP by up to 7.853% and 7.62%, higher than DMHR by up to 17.707% and 17.2%, higher than FAH by up to 37.163% and 32.762%, and higher than IDHN by up to 39.862% and 35.802%, respectively. These results indicate that the SGDSH method is particularly effective in generating distinguishable hash codes for datasets with varying spatial resolutions. Figures 10 and 11 present the performance results of each method across different hash code lengths on the AID-multi dataset.

Table 6. AID-multi dataset retrieval of the top 10 images comparison experiment results. The best values are highlighted in bold.

Method	Bits	MAP	NDCG@10	ACG@10	WAP@10
Our method	64	98.331	98.564	91.389	90.826
	96	99.345	99.225	92.993	92.618
	128	**99.566**	**99.351**	**93.388**	**93.178**
HyP	64	97.824	96.843	83.826	83.173
	96	98.994	97.729	84.501	83.939
	128	99.074	97.964	85.535	85.015
DMHR	64	96.308	94.806	73.831	71.831
	96	96.719	95.056	73.152	71.389
	128	97.107	95.752	75.681	74.149
FAH	64	77.833	88.801	55.980	49.736
	96	77.744	92.524	56.225	50.283
	128	75.811	91.753	54.326	48.291
IDHN	64	75.079	85.881	52.895	46.549
	96	76.844	85.763	52.045	46.788
	128	79.044	85.949	53.526	49.044

Table 7. AID-multi dataset retrieval of the top 50 images comparison experiment results. The best values are highlighted in bold.

Method	Bits	MAP	NDCG@50	ACG@50	WAP@50
Our method	64	97.282	97.933	87.127	85.887
	96	98.489	98.951	88.646	87.956
	128	**98.932**	**99.074**	**89.579**	**89.042**
HyP	64	96.634	95.951	80.098	78.881
	96	97.841	96.503	80.773	79.705
	128	98.044	96.805	81.959	80.972
DMHR	64	92.555	93.458	70.652	67.085
	96	93.386	93.344	70.704	67.669
	128	93.994	93.974	72.379	69.611
IDHN	64	74.465	85.722	55.569	47.388
	96	78.306	84.545	56.817	50.866
	128	76.349	84.036	55.327	49.088
FAH	64	75.074	87.462	52.652	47.075
	96	75.671	87.415	52.883	47.311
	128	77.838	87.289	53.777	48.715

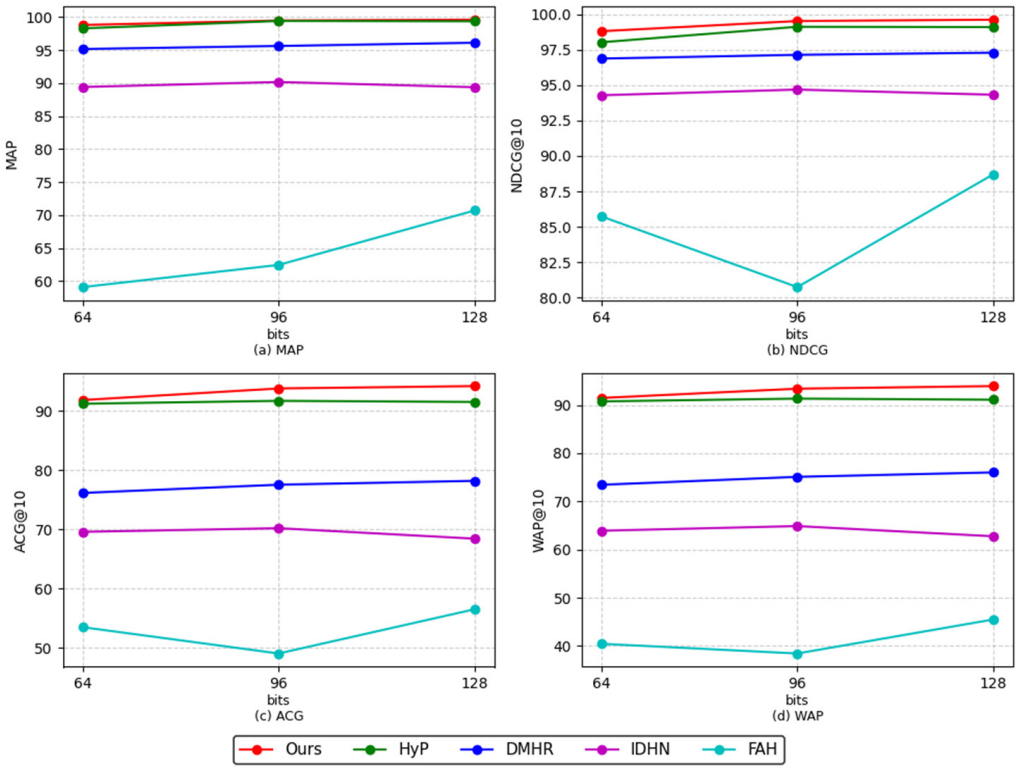


Figure 8. Performance curves of MAP, NDCG, ACG, and WAP for different hash code lengths on top 10 retrieved images from UCM-multi dataset.

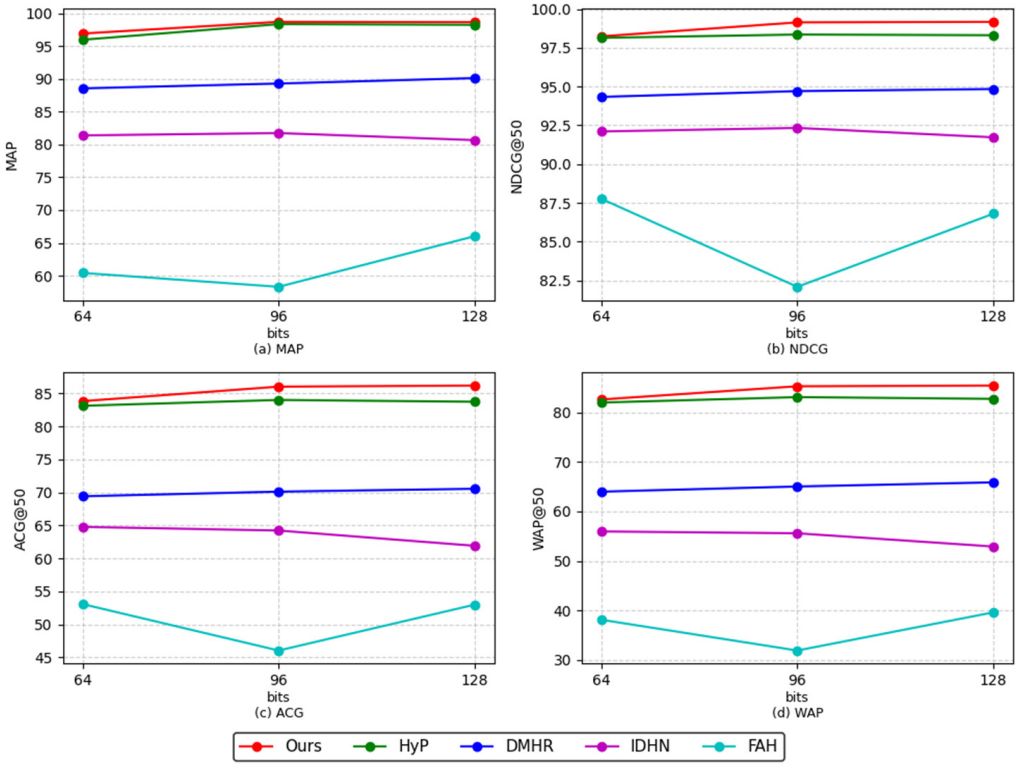


Figure 9. Performance curves of MAP, NDCG, ACG, and WAP for different hash code lengths on top 50 retrieved images from UCM-multi dataset.

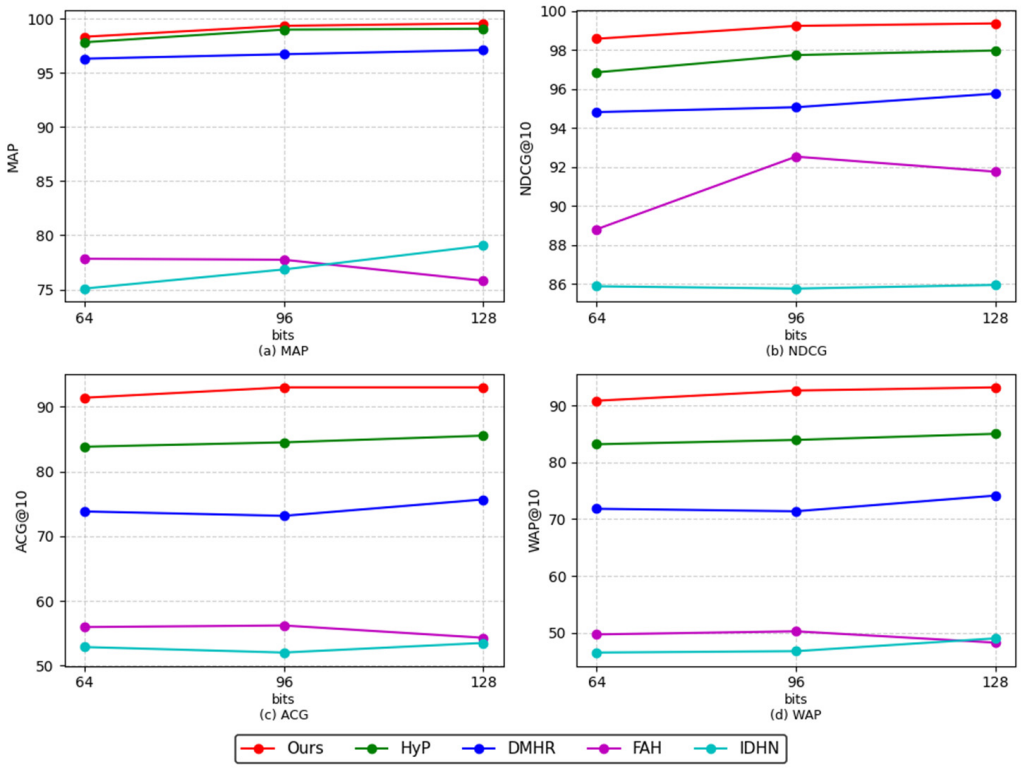


Figure 10. Performance curves of MAP, NDCG, ACG, and WAP for different hash code lengths on top 10 retrieved images from AID-multi dataset.

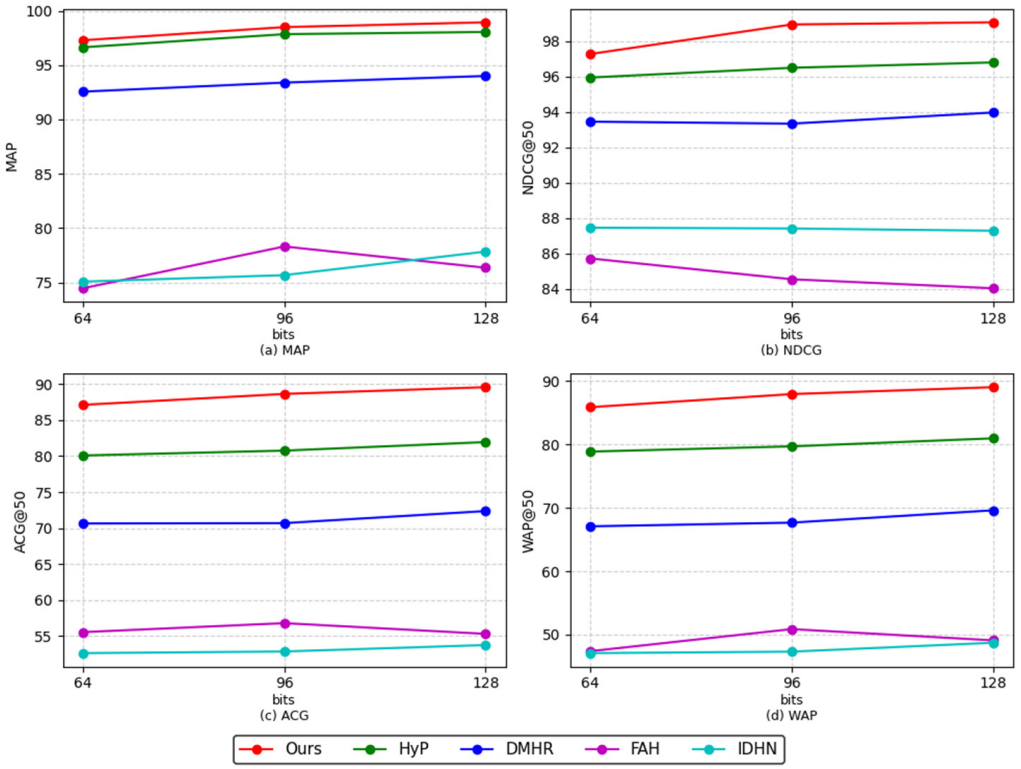


Figure 11. Performance curves of MAP, NDCG, ACG, and WAP for different hash code lengths on top 50 retrieved images from AID-multi dataset.

In the previous comparison experiments, our method with the HyP method showed a very significant superior performance compared to the other comparison methods. In addition, the experimental results of the MAP metrics of our method and HyP on both

datasets are very close to each other. To further validate the effectiveness of the proposed method, we compared it with HyP on the MLRS dataset. In contrast to the UCM-multi dataset (2100 images) and AID-multi dataset (3000 images), the MLRS dataset contains a significantly larger volume of data (109,161 images) and a greater number of categories (60 predefined types), making it more representative of actual applications. The retrieval results on the MLRS dataset are presented in Tables 8 and 9, and Figures 12 and 13 present the performance results of our method and the HyP method.

Table 8. MLRS dataset retrieval of the top 10 images comparison experiment results. The best values are highlighted in bold.

Method	Bits	MAP	NDCG@10	ACG@10	WAP@10
Our method	64	95.081	95.312	79.667	77.287
	96	96.473	96.069	81.555	79.739
	128	**96.583**	**96.413**	**81.824**	**79.975**
HyP	64	89.984	93.298	72.348	67.462
	96	95.045	95.004	77.529	74.711
	128	95.449	95.546	77.775	75.253

Table 9. MLRS dataset retrieval of the top 50 images comparison experiment results. The best values are highlighted in bold.

Method	Bits	MAP	NDCG@50	ACG@50	WAP@50
Our method	64	92.621	94.249	76.851	73.415
	96	94.471	95.187	78.498	75.716
	128	**94.516**	**95.471**	**78.992**	**76.194**
HyP	64	87.827	92.824	70.995	65.156
	96	92.091	94.196	74.636	70.599
	128	92.174	94.431	74.751	70.836

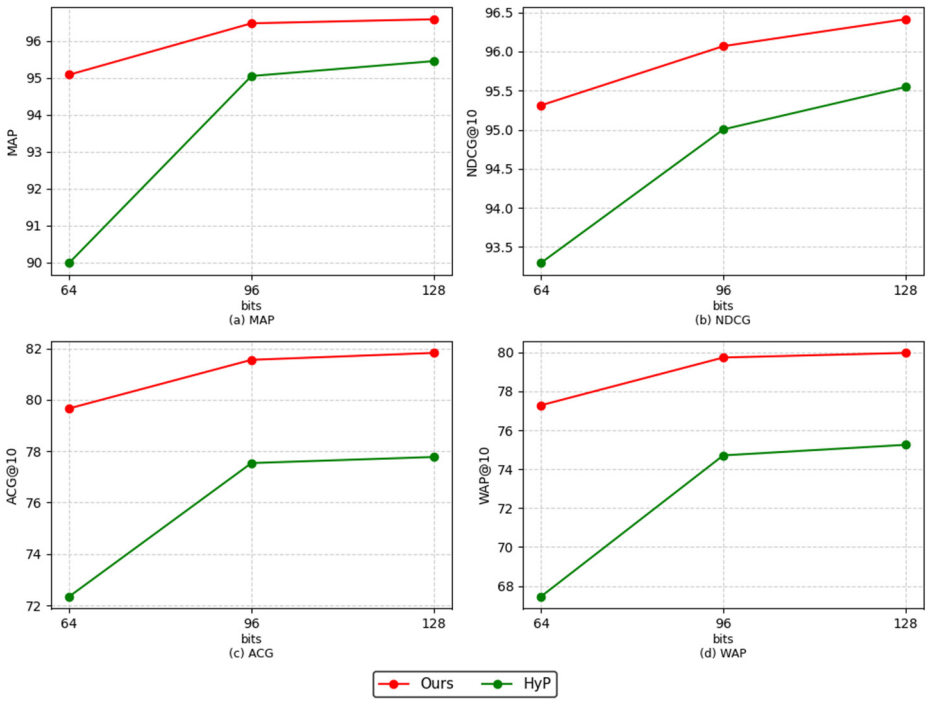


Figure 12. Performance curves of MAP, NDCG, ACG, and WAP for different hash code lengths on top 10 retrieved images from MLRS dataset.

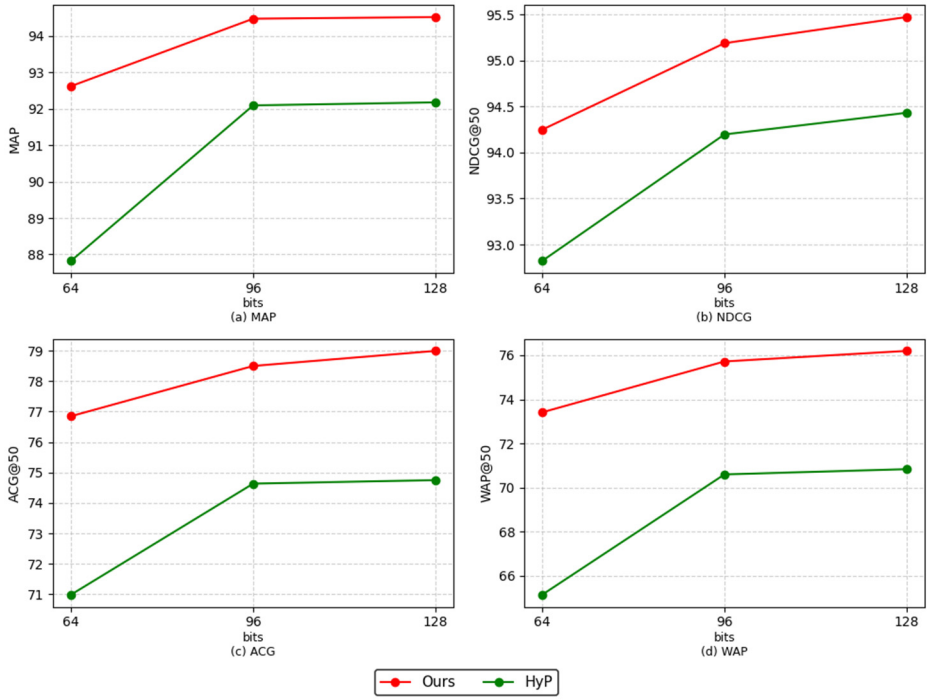


Figure 13. Performance curves of MAP, NDCG, ACG, and WAP for different hash code lengths on top 50 retrieved images from MLRS dataset.

The experimental results show that both methods achieve the best performance when hash code length is set to 128 bits, indicating that longer hash codes are more effective at retaining the original high-dimensional feature information. In the mAP results for retrieving the top 10 most similar images, the proposed method improves by 1.134% compared to the HyP method. For retrieving the top 50 most similar images, the proposed method improves by 2.342%. This demonstrates that the proposed method maintains excellent retrieval accuracy even when dealing with large-scale datasets. In terms of ACG, the proposed method outperforms the HyP method by 4.049% for the top 10 images and by 4.241% for the top 50 images. This consistently shows that the retrieval results in the proposed method maintain a high similarity to the query images in terms of feature type and visual content. These results further validate the effectiveness of the proposed method in multi-label remote sensing image retrieval tasks.

Summarizing the results of comparison experiments in this section, our method demonstrates the effectiveness of the SGDSH approach when compared with other state-of-the-art methods across three different datasets. This can be attributed to the following key factors: The Multi-scale Feature Fusion Module effectively learns and fuses various types of features from remote sensing images at different scales, enhancing the model's feature extraction capability and ensuring a comprehensive representation of different feature types. Our proposed pairwise similarity loss strategy effectively reduces the distance between images with more common labels. This strategy helps ensure that images with similar labels are ranked higher in retrieval results, thereby improving the retrieval performance. The sorting strategy combined with category labels enhances visual similarity between query image and retrieved images. This results in satisfying retrieval outcomes, even when the dataset is large and images have inconsistent spatial resolutions. These experimental results fully confirm the effectiveness and practicality of the SGDSH method in the multi-label remote sensing image retrieval task.

4.6. Visual Analysis

To visualize the performance of different methods in the task of multi-label remote sensing image retrieval, we present a comparison of query retrieval results of the proposed method with those of HyP, DMHR, IDHN, and FAH on the UCM-multi and AID-multi datasets. It is important to note that all methods use a uniform 128-bit hash code, and the query image does not serve as a retrieved image, meaning it is excluded from the retrieval process. To further validate the effectiveness of the similarity metric strategy proposed in this paper, we also include the results of our method without Jaccard weighting, referred to as "our method". The retrieval results are displayed in Figures 14 and 15.

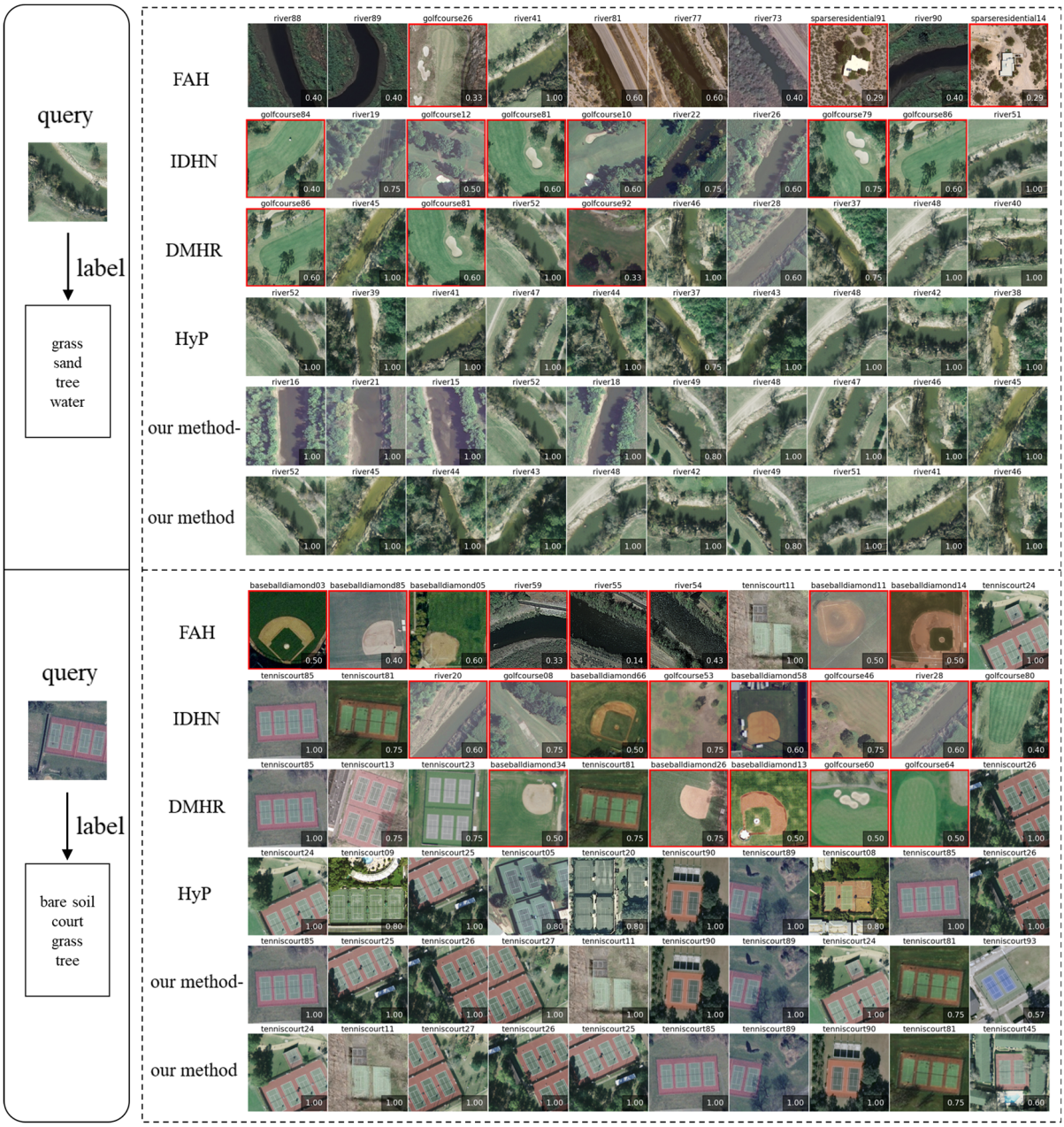


Figure 14. Example of retrieval results using different methods on the UCM-multi dataset. Each row displays the top 10 retrieval results. The leftmost image is the query image, with the names of the retrieved images shown at the top. Incorrectly retrieved images are highlighted with a red box. The label similarity between the ground truth labels of the retrieved images and the query image is shown in the bottom-right corner.

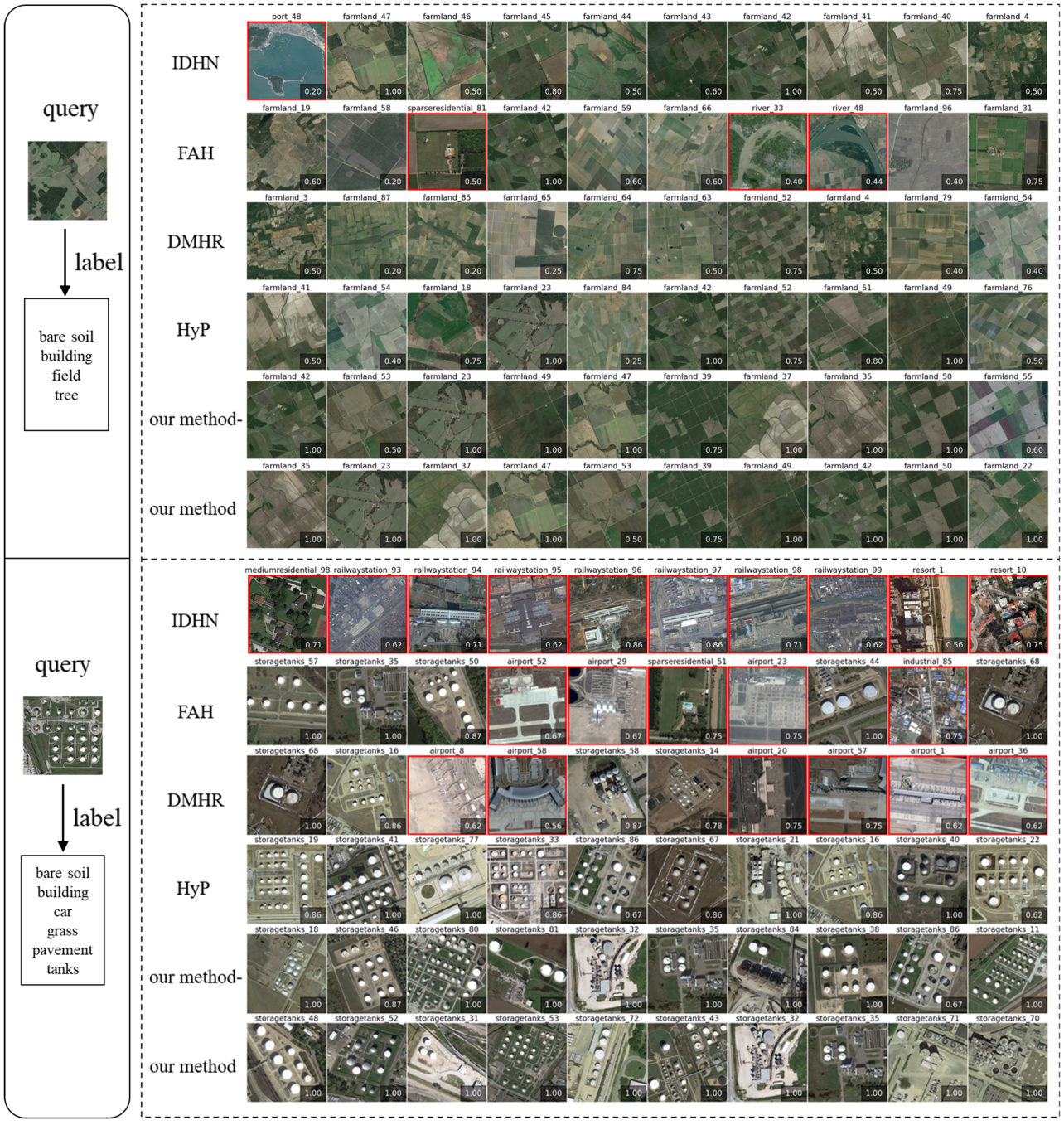


Figure 15. Example of different methods of retrieval of AID-multi dataset. The leftmost image is the query image, with the names of the retrieved images shown at the top. Incorrectly retrieved images are highlighted with a red box. The label similarity between the ground truth labels of the retrieved images and the query image is shown in the bottom-right corner.

We specifically selected natural landscape images and artificial facility images as two types of retrieval queries. The retrieval results demonstrate that our proposed method ensures all returned results are correct for both queries, with retrieved images showing a higher similarity to the actual labels of query images. This means the retrieved images share more labels with the query images. Notably, by weighting the Hamming distance, our method preferentially returns images that share more labels with the query, leading to retrieval results that are closer to the ideal outcome. This indicates that the proposed method can effectively recognize different feature classes within remote sensing images,

extract features from each class, and deliver strong ranking performance, thus providing more accurate and reliable retrieval results for practical applications.

5. Discussion

To enhance the feature extraction capability of remote sensing images, this study proposes a multi-scale feature fusion model, Semantically Guided Deep Supervised Hashing (SGDSH), based on Swin Transformer. The model effectively leverages detailed information in shallow features and rich semantic information in high-level features, while suppressing irrelevant information. This end-to-end deep hashing framework enables multi-scale feature fusion, optimizing the retrieval process. To improve the quality of hash code extraction, a classification layer is added after the hash layer, using real feature labels to guide training of the hash code. The predicted label information is then incorporated into the image retrieval task. The experimental results demonstrate that the inclusion of label information during computation and similarity ranking significantly improves the accuracy of multi-label remote sensing image retrieval. This improvement indicates that the proposed method effectively captures both the categories and distributions of ground objects within remote sensing images. It prioritizes the retrieval of images with higher semantic similarity to the query in terms of land cover types. This leads to retrieval results that better reflect the actual composition of the query image, making the method more suitable for practical applications in remote sensing image analysis. For example, in disaster assessment, retrieving semantically similar historical images from remote sensing databases aids in damage evaluation and response planning. In agricultural monitoring, identifying similar land cover types across different periods helps track crop growth patterns and detect anomalies. By improving the efficiency and accuracy of multi-label remote sensing image retrieval, this method provides a valuable tool for large-scale geospatial analysis and decision-making.

While the proposed method exhibits strong retrieval performance, there are still some limitations that need to be addressed in future work. Firstly, this paper introduces a classification layer to incorporate predicted labels into the similarity measure calculation. However, improving the accuracy of the predicted labels remains an important issue to address. Secondly, in judging the similarity of image pairs, this paper uses a label similarity threshold of 0.4 as the criterion for determining whether two images are similar. The impact of using higher similarity thresholds on model training and retrieval performance has not been explored. Additionally, the retrieval speed has not been experimentally analyzed. In our model, we adopt an improved Swin Transformer framework as the backbone. While comparative experiments demonstrate that our method achieves higher retrieval accuracy than other state-of-the-art approaches, the increased model complexity of the Transformer-based model introduces a higher computational overhead, leading to longer inference times compared to traditional CNN-based models. Moreover, the similarity metric proposed in this paper enhances retrieval accuracy; it also requires the storage of image label information when constructing the feature library. This significantly increases storage costs, especially when building large-scale remote sensing image feature libraries. Furthermore, the two-stage calculation process for the similarity metric—first computing hash code and label information for the query image, then combining them with hash codes and category labels of all the images in the database—inevitably adds a time overhead. In future research, we aim to develop a lightweight Transformer-based architecture that maintains the feature extraction capability of remote sensing images while reducing model complexity. Additionally, we plan to integrate hash codes and predicted label information into a unified compressed feature representation, aiming to reduce storage requirements and improve retrieval efficiency while preserving feature retrieval accuracy. Through the

above research directions, we aim to develop a more lightweight and efficient multi-label remote sensing image retrieval method.

6. Conclusions

In this paper, we propose an end-to-end multi-label remote sensing image retrieval method called Semantically Guided Deep Supervised Hashing (SGDSH), based on an improved Swin Transformer model. By incorporating an innovative multi-scale feature fusion module, we enhance the model's ability to extract deep features from remote sensing images. Additionally, our method leverages the semantic information in the data, with a carefully designed weighted pairwise similarity loss function and semantic labels, enabling the model to generate hash features with superior differentiation. To further improve the ranking performance of hash features, we introduce image category labels to compute similarity and apply a weighted Hamming distance between image features. This approach significantly enhances retrieval accuracy and ensures that the retrieved images share more common labels with the query image.

The experimental results demonstrate that the SGDSH method achieves outstanding retrieval performance across all three datasets. In the retrieval task of returning the top 10 most similar images, ACG and wAP metrics for SGDSH reach 94.148%, 93.388%, and 81.824% and 93.924%, 93.178%, and 79.975%, respectively. For the top 50 images, these metrics are 86.203%, 89.579%, and 78.992% and 85.438%, 89.042%, and 76.194%. Comparative experiments with other methods further highlight the advantages of SGDSH, showing significant improvements in ACG and WAP over the best-performing comparison method. Specifically, for the top 10 images returned, SGDSH outperforms the other methods by 2.673%, 7.853%, and 4.049% and 2.827%, 8.163%, and 4.722%, respectively, and, for the top 50 images, it improves by 2.181%, 7.62%, and 4.241% and 2.339%, 8.07%, and 5.358%. Additionally, visualization analysis confirms that SGDSH not only ensures high retrieval accuracy but also enhances label similarity between the query and retrieved images. These results validate the effectiveness of SGDSH in large-scale remote sensing image retrieval.

Despite promising results, the proposed SGDSH method has some limitations, including higher inference time due to the Transformer-based backbone, increased storage costs from label information in the feature library, and additional retrieval time caused by the two-stage similarity computation, which may hinder large-scale retrieval. To address these issues, future research will focus on developing a lightweight Transformer-based architecture that reduces model complexity while retaining feature extraction capability, as well as integrating hash codes and label information into a unified compressed representation. These efforts aim to improve the scalability and efficiency of multi-label remote sensing image retrieval, providing valuable insights and opening up new research avenues in this field.

Author Contributions: B.L. designed the experiments and wrote the manuscript; W.L. supervised this study and reviewed the draft paper; S.L. revised the manuscript and gave some appropriate suggestions. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to institutional confidentiality.

Acknowledgments: We thank the China Remote Sensing Satellite Ground Station for supporting this experiment.

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

References

1. Zhou, W.; Newsam, S.; Li, C.; Shao, Z. PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 197–209. [CrossRef]
2. Sumbul, G.; Kang, J.; Demir, B. Deep learning for image search and retrieval in large remote sensing archives. In *Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science, and Geosciences*; Wiley Online Library: Hoboken, NJ, USA, 2021; pp. 150–160.
3. 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]
4. JS, S.K. A review on content-based image retrieval techniques. In Proceedings of the 2023 International Conference on Circuit Power and Computing Technologies (ICCPCT), Kollam, India, 10–11 August 2023; pp. 1251–1256.
5. 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]
6. Tong, X.-Y.; Xia, G.-S.; Lu, Q.; Shen, H.; Li, S.; You, S.; Zhang, L. Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sens. Environ.* **2020**, *237*, 111322. [CrossRef]
7. Han, J.; Li, P.; Tao, Y.; Ren, P. Encrypting hashing against localization. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–14. [CrossRef]
8. Yuan, Z.; Zhang, W.; Tian, C.; Rong, X.; Zhang, Z.; Wang, H.; Fu, K.; Sun, X. Remote sensing cross-modal text-image retrieval based on global and local information. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–16. [CrossRef]
9. Malik, K.; Robertson, C. Landscape similarity analysis using texture encoded deep-learning features on unclassified remote sensing imagery. *Remote Sens.* **2021**, *13*, 492. [CrossRef]
10. Hatakeyama, W.; Kawakita, S.; Izawa, R.; Kimura, M. GRASP EARTH: Intuitive Software for Discovering Changes on the Planet. *arXiv* **2022**, arXiv:2203.00955.
11. Pi, Y.; Duffield, N.; Behzadan, A.H.; Lomax, T. Visual recognition for urban traffic data retrieval and analysis in major events using convolutional neural networks. *Comput. Urban Sci.* **2022**, *2*, 2–19. [CrossRef] [PubMed]
12. Li, Y.; Ma, J.; Zhang, Y. Image retrieval from remote sensing big data: A survey. *Inf. Fusion* **2021**, *67*, 94–115. [CrossRef]
13. Wang, Z.; Wu, N.; Yang, X.; Yan, B.; Liu, P. Deep learning triplet ordinal relation preserving binary code for remote sensing image retrieval task. *Remote Sens.* **2021**, *13*, 4786. [CrossRef]
14. 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]
15. Han, L.; Paoletti, M.E.; Moreno-Álvarez, S.; Haut, J.M.; Pastor-Vargas, R.; Plaza, A. Hashing for Retrieving Long-Tailed Distributed Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–14. [CrossRef]
16. 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]
17. 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]
18. Huang, H.; Cheng, Q.; Shao, Z.; Huang, X.; Shao, L. DMCH: A Deep Metric and Category-Level Semantic Hashing Network for Retrieval in Remote Sensing. *Remote Sens.* **2023**, *16*, 90. [CrossRef]
19. Shao, Z.; Yang, K.; Zhou, W. Performance Evaluation of Single-Label and Multi-Label Remote Sensing Image Retrieval Using a Dense Labeling Dataset. *Remote Sens.* **2018**, *10*, 964. [CrossRef]
20. Shao, Z.; Zhou, W.; Deng, X.; Zhang, M.; Cheng, Q. Multilabel remote sensing image retrieval based on fully convolutional network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 318–328. [CrossRef]
21. Chaudhuri, B.; Demir, B.; Chaudhuri, S.; Bruzzone, L. Multilabel remote sensing image retrieval using a semisupervised graph-theoretic method. *IEEE Trans. Geosci. Remote Sens.* **2017**, *56*, 1144–1158. [CrossRef]
22. Cheng, Q.; Huang, H.; Ye, L.; Fu, P.; Gan, D.; Zhou, Y. A semantic-preserving deep hashing model for multi-label remote sensing image retrieval. *Remote Sens.* **2021**, *13*, 4965. [CrossRef]
23. Jin, W.; Cai, Z.; Pan, Y.; Fu, R. ICIHRN: An Interpretable Multi-Label Hash Retrieval Method for Satellite Cloud Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 8662–8672. [CrossRef]
24. 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]
25. Imbriaco, R.; Sebastian, C.; Bondarev, E.; de With, P.H. Toward multilabel image retrieval for remote sensing. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–14. [CrossRef]
26. Lu, W.; Nguyen, M. Enhancing Remote Sensing Image Retrieval: A Hierarchical Approach Integrating Visual and Semantic Similarities. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 1–5. [CrossRef]
27. Wang, J.; Shen, H.T.; Song, J.; Ji, J. Hashing for similarity search: A survey. *arXiv* **2014**, arXiv:1408.2927.

28. Rodrigues, J.; Cristo, M.; Colonna, J.G. Deep hashing for multi-label image retrieval: A survey. *Artif. Intell. Rev.* **2020**, *53*, 5261–5307. [CrossRef]
29. Li, P.; Ren, P. Partial randomness hashing for large-scale remote sensing image retrieval. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 464–468. [CrossRef]
30. 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]
31. Li, Y.; Zhang, Y.; Huang, X.; Zhu, H.; Ma, J. Large-scale remote sensing image retrieval by deep hashing neural networks. *IEEE Trans. Geosci. Remote Sens.* **2017**, *56*, 950–965. [CrossRef]
32. Liu, C.; Ma, J.; Tang, X.; Zhang, X.; Jiao, L. Adversarial hash-code learning for remote sensing image retrieval. In Proceedings of the IGARSS 2019—2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 4324–4327.
33. 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*, 1–14. [CrossRef]
34. Roy, S.; Sangineto, E.; Demir, B.; Sebe, N. Metric-learning-based deep hashing network for content-based retrieval of remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 226–230. [CrossRef]
35. Zhou, H.; Qin, Q.; Hou, J.; Dai, J.; Huang, L.; Zhang, W. Deep global semantic structure-preserving hashing via corrective triplet loss for remote sensing image retrieval. *Expert Syst. Appl.* **2024**, *238*, 122105. [CrossRef]
36. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
37. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8 September 2018; pp. 3–19.
38. Kang, J.; Fernandez-Beltran, R.; Hong, D.; Chanussot, J.; Plaza, A. Graph relation network: Modeling relations between scenes for multilabel remote-sensing image classification and retrieval. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 4355–4369. [CrossRef]
39. Hua, Y.; Mou, L.; Zhu, X.X. Relation network for multilabel aerial image classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 4558–4572. [CrossRef]
40. 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.
41. Xia, G.-S.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Zhang, L.; Lu, X. AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3965–3981. [CrossRef]
42. Xu, C.; Chai, Z.; Xu, Z.; Yuan, C.; Fan, Y.; Wang, J. Hyp^2 loss: Beyond hypersphere metric space for multi-label image retrieval. In Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2022; pp. 3173–3184.
43. Zhang, Z.; Zou, Q.; Lin, Y.; Chen, L.; Wang, S. Improved deep hashing with soft pairwise similarity for multi-label image retrieval. *IEEE Trans. Multimed.* **2019**, *22*, 540–553. [CrossRef]

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