


Article

An Image Retrieval Method for Lunar Complex Craters Integrating Visual and Depth Features

Yingnan Zhang ^{1,2} , Zhizhong Kang ^{1,2,*}  and Zhen Cao ^{1,2}

¹ School of Land Science and Technology, China University of Geosciences, Beijing 100083, China; zyn@glut.edu.cn (Y.Z.); caozhen@email.cugb.edu.cn (Z.C.)

² Subcenter of International Cooperation and Research on Lunar and Planetary Exploration, Center of Space Exploration, Ministry of Education of the People's Republic of China, Beijing 100083, China

* Correspondence: zzkang@cugb.edu.cn

Abstract: In the geological research of the Moon and other celestial bodies, the identification and analysis of impact craters are crucial for understanding the geological history of these bodies. With the rapid increase in the volume of high-resolution imagery data returned from exploration missions, traditional image retrieval methods face dual challenges of efficiency and accuracy when processing lunar complex crater image data. Deep learning techniques offer a potential solution. This paper proposes an image retrieval model for lunar complex craters that integrates visual and depth features (LC²R-Net) to overcome these difficulties. For depth feature extraction, we employ the Swin Transformer as the core architecture for feature extraction and enhance the recognition capability for key crater features by integrating the Convolutional Block Attention Module with Effective Channel Attention (CBAMwithECA). Furthermore, a triplet loss function is introduced to generate highly discriminative image embeddings, further optimizing the embedding space for similarity retrieval. In terms of visual feature extraction, we utilize Local Binary Patterns (LBP) and Hu moments to extract the texture and shape features of crater images. By performing a weighted fusion of these features and utilizing Principal Component Analysis (PCA) for dimensionality reduction, we effectively combine visual and depth features and optimize retrieval efficiency. Finally, cosine similarity is used to calculate the similarity between query images and images in the database, returning the most similar images as retrieval results. Validation experiments conducted on the lunar complex impact crater dataset constructed in this article demonstrate that LC²R-Net achieves a retrieval precision of 83.75%, showcasing superior efficiency. These experimental results confirm the advantages of LC²R-Net in handling the task of lunar complex impact crater image retrieval.

Keywords: LC²R-Net; CBAM; ECA; impact crater; image retrieval; deep learning; triplet loss function



Citation: Zhang, Y.; Kang, Z.; Cao, Z. An Image Retrieval Method for Lunar Complex Craters Integrating Visual and Depth Features. *Electronics* **2024**, *13*, 1262. <https://doi.org/10.3390/electronics13071262>

Academic Editors: Peter Odry and Vladimir László Tadić

Received: 4 March 2024

Revised: 26 March 2024

Accepted: 28 March 2024

Published: 28 March 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

Impact craters on the lunar surface are significant witnesses to the history of the Solar System. Their size, shape, and distribution provide key insights into understanding the geological history of the Moon and other celestial bodies [1–4]. With the advancement of space exploration technology, we are now able to obtain high-resolution imagery of the lunar surface. Over the past few decades, lunar exploration projects such as NASA's Apollo program, the Lunar Reconnaissance Orbiter, and China's Chang'e program have accumulated a vast amount of lunar data, which have been used for in-depth studies [5–9]. These images contain rich information, such as the morphology, structure, and geological features of impact craters, as well as the distribution of rocks related to impact events. However, this also presents a challenge: how to effectively retrieve and analyze the vast amount of crater imagery data [10]. Content-based image retrieval (CBIR) technology may be an effective solution to this problem.

Content-based image retrieval systems have a wide range of applications in lunar and planetary science research. Traditional CBIR methods typically rely on the visual

content of images, such as texture, shape, and color features, to index and retrieve images. These methods depend on handcrafted features such as Speeded Up Robust Features (SURF) [11], Hu moments, and Gabor features [12]. Although these features are effective in certain scenarios, their application is limited in the complex lunar environment, where they struggle to capture detailed information within images effectively. Moreover, these methods often require meticulous feature engineering and parameter tuning, which is not only time-consuming but also limits their generalization ability and scalability. With the rise of deep learning technologies, particularly the successful application of Convolutional Neural Networks (CNNs) in remote sensing image recognition and classification tasks [13–15], researchers have begun to explore the use of deep features to enhance the performance of CBIR systems.

The successful application of deep learning methods in the field of remote sensing image retrieval has demonstrated their significant advantages over traditional manual feature methods [16–18]. Deep learning models are capable of automatically extracting abstract feature representations from images by learning from large-scale datasets. These deep features can better capture the semantic information and contextual relationships within images, thus enhancing the accuracy and robustness of CBIR systems. Given the complexity of remote sensing image description, some scholars have begun to explore strategies for feature fusion. Yan et al. have found through research that CNN features and SIFT features are highly complementary and can significantly improve the performance of image retrieval tasks [19]. Cheng et al. have proposed a distributed retrieval system architecture suitable for high-resolution satellite images by combining deep features with traditional manual features [20]. Their work indicates that the combination of deep features and traditional manual features can provide a more comprehensive image representation method. Although these methods have achieved commendable results, these models generally use the cross-entropy loss function during training, which has certain limitations in image retrieval tasks. In image retrieval tasks, we are more concerned with the similarity between images rather than just the accuracy of categories. This means that models need to be able to not only identify the categories of images but also capture the subtle differences within categories and the significant distinctions between categories. To this end, deep metric learning has become a key technology for addressing such problems.

Deep metric learning integrates the advantages of deep learning and metric learning. This approach automatically extracts image features through deep learning models and optimizes the distances between features through metric learning, making the distances in feature space closer for similar images while expanding the distances between dissimilar images. Deep metric learning has shown significant effectiveness in multiple applications in the field of remote sensing, including image retrieval [21–25], image classification [26,27], and object recognition [28,29], etc. However, the task of image retrieval for lunar impact craters demands more complex and meticulous feature extraction requirements, and these models do not always effectively capture all the key features of the craters. In this context, the Swin Transformer, as a novel deep learning architecture [30], has demonstrated its powerful performance in various visual tasks, which has inspired us to utilize this technology to address the challenges of lunar impact crater image retrieval.

The Swin Transformer is a neural network architecture based on the Transformer, widely applied as an efficient feature extractor in computer vision tasks [31–35]. Compared to traditional CNNs, the Swin Transformer not only effectively models global contextual information but also captures features at different scales through its hierarchical structure, which is crucial for understanding complex scenes and relationships within images. Inspired by this, we propose a lunar complex crater image retrieval model (LC²R-Net) that fuses visual and depth features. We employ the Swin Transformer as the core architecture for depth feature extraction and integrate LBP and Hu moments for visual feature extraction. Moreover, to evaluate the effectiveness of our method, we have constructed a lunar crater image retrieval dataset and conducted extensive experiments. Our main contributions are as follows:

1. The Swin Transformer is utilized as the feature extraction structure, and the CBAMwith-ECA module is integrated into the linear embedding and patch merging modules. Through the attention mechanism, the channel and spatial relevance of features are enhanced, allowing for a comprehensive capture of the details and structural information within images. This enhancement improves the model's capability to recognize and extract image features. It directs the model's focus toward the global context, elevating the perceptibility of key features while concurrently suppressing less important features and noise information.
2. By integrating visual features (texture features, shape features) with deep features, we balance the contribution of different features through a weighted approach, emphasizing important features during the fusion process. Furthermore, we apply PCA to condense the dimensionality of the integrated feature set. This process not only trims down the number of feature dimensions but also amplifies the retrieval process's swiftness and effectiveness.
3. Within the network's training framework, we integrate a triplet loss function coupled with a strategy for mining difficult negative examples. This approach is designed to prompt the network to cultivate features with greater discrimination. By utilizing triplet loss, we optimize the embedded space, ensuring that vectors of akin images are positioned in closer proximity, whereas those of non-akin images are segregated, thereby markedly boosting the precision of our retrieval system.

The structure of this document is laid out in the following manner: Section 2 introduces the work related to content-based image retrieval. Section 3 delineates the method we propose. Section 4 details the dataset used for the retrieval of complex impact crater images on the lunar surface. Section 5 is dedicated to an in-depth presentation of experimental outcomes and their subsequent analysis. Finally, Section 6 summarizes the findings and conclusions of this study.

The related codes are publicly available at <https://github.com/ZYNHYF/lunar-complex-crater-image-retrieval> (released on 27 March 2024).

2. Related Works

In this section, we provide a detailed overview of prior research work related to content-based image retrieval. We categorize these studies into three groups: methods based on traditional features, methods based on deep features, and methods based on metric learning.

2.1. Methods Based on Traditional Features

Early CBIR systems primarily relied on traditional image processing techniques to extract features, such as color histograms, texture features, and shape descriptors. The advantages of these methods lie in their computational simplicity and intuitive understanding, and they have been extensively studied by scholars. Tekeste et al. conducted a comparative study to explore the impact of different LBP variants on the results of remote sensing image retrieval [36]. Aptoula applied global morphological texture descriptors to remote sensing image retrieval and, despite the shorter length of the extracted feature vectors, achieved high retrieval scores [37]. Xie et al. proposed an image retrieval method that combines a dominant color descriptor with Hu moments, leveraging the advantages of color and shape detection [38]. Chen et al. introduced a feature descriptor based on the relationships between prominent craters on the lunar surface and a composite feature model composed of different features. Based on these characteristics, similarity measurement rules and a retrieval algorithm were proposed and detailed [39]. Hua et al. utilized a general saliency-based landmark detection algorithm to identify regions of interest on the lunar surface, then indexed and retrieved them using feature vectors extracted from the region-of-interest images, evaluating the performance of saliency-based landmark detection [12]. However, these methods also have apparent limitations; they perform well under specific conditions, particularly when the image content structure is simple and changes little. Nevertheless,

they often fail to effectively handle high-level semantic information and have limited robustness in complex scenes.

2.2. Methods Based on Deep Features

With the advancement of deep learning technology, methods based on Convolutional Neural Networks have become a research hotspot in the field of Content-Based Image Retrieval. These methods automatically extract deep representations of image content by learning multi-level abstract features. Compared to handcrafted features, deep features are better at capturing the complex patterns and high-level semantic information in images. Wang et al. designed a Multi-Attention Fusion Network with dilated convolution and label smoothing capabilities, using label smoothing to replace the cross-entropy loss function, which yielded competitive retrieval results [40]. Ye et al. proposed a query-adaptive feature fusion method based on a CNN regression model, which can accurately predict the DCG values of the ranked image list to assign weights to each feature, thereby enhancing retrieval precision [41]. Wang et al. introduced a novel Wide Context Attention Network (W-CAN), utilizing two attention modules to adaptively learn relevant local features in spatial and channel dimensions, thus obtaining discriminative features with extensive contextual information [42]. Chaudhuri et al. designed a GCN-based Context Attention Network, including node and edge attention. Beyond highlighting the fundamental features within each node, edge attention enables the network to learn the most critical neighborhood structures from the RAG within each target class image [43]. Furthermore, methods based on deep features can also leverage transfer learning to adapt these pre-trained models to specific domains or datasets [40], further improving retrieval performance.

2.3. Methods Based on Metric Learning

Metric learning methods aim to learn an optimized distance metric such that similar images are closer in the feature space while dissimilar images are farther apart. These methods are often combined with deep learning, adjusting the feature space through the loss function during training. Zhang et al. constructed a Triplet Non-Local Neural Network (T-NLNN) model that combines deep metric learning with non-local operations, significantly improving the performance of high-resolution remote sensing image retrieval [21]. Cao et al. proposed a method based on a triplet deep metric learning network to enhance the retrieval performance of remote sensing images [22]. Zhong et al. introduced an L2-normed attention and multi-scale fusion network (L2AMF-Net) to achieve accurate and robust lunar image patch matching [44]. Additionally, some scholars focus on constructing new loss functions to enhance retrieval performance. Fan et al. proposed a ranking loss result, thereby building a global optimization model based on feature space and retrieval outcomes, which can be optimized in an end-to-end manner [45]. Zhao et al. designed a similarity-preserving loss-based deep metric learning strategy, utilizing the ratio of easy to hard samples within classes to dynamically weigh the selected hard samples in experiments, learning the structural characteristics of intra-class samples [46]. Fan et al. introduced a distribution consistency loss to address the problem of imbalanced sample distribution in remote sensing datasets, constructing an end-to-end fine-tuned network suitable for remote sensing image retrieval, achieving state-of-the-art performance [47]. Compared to methods based on deep features, metric learning-based approaches are particularly suitable for tasks requiring refined retrieval and sensitivity to similarity.

3. Proposed Method

The LC²R-Net model proposed in this paper achieves the task of lunar complex impact crater image retrieval by fusing low-level visual features with deep features of images. By integrating these two complementary types of features, a more comprehensive image representation is formed, which enhances the model's ability to recognize and differentiate complex impact craters, thereby improving the accuracy of retrieval. Figure 1 outlines the overall process by which LC²R-Net completes the retrieval task.

As illustrated in Figure 1, the core steps include: (1) extracting deep features of impact crater images using an improved Swin Transformer model, which, by integrating the CBAMwithECA attention module, mines potential information within feature maps across channel and spatial dimensions, achieving comprehensive calibration and meticulous optimization of features, thereby enhancing the model’s capability to represent image features; (2) utilizing LBP and Hu moments to extract texture and shape features of impact crater images as low-level visual features; (3) the extracted low-level visual and deep features are weighted and fused, followed by dimensionality reduction to create a more compact and efficient feature representation for retrieval tasks; (4) finally, the model employs the fused feature representation to perform the image retrieval task, matching query images with images in the database, and identifying the most similar images based on the queried features.

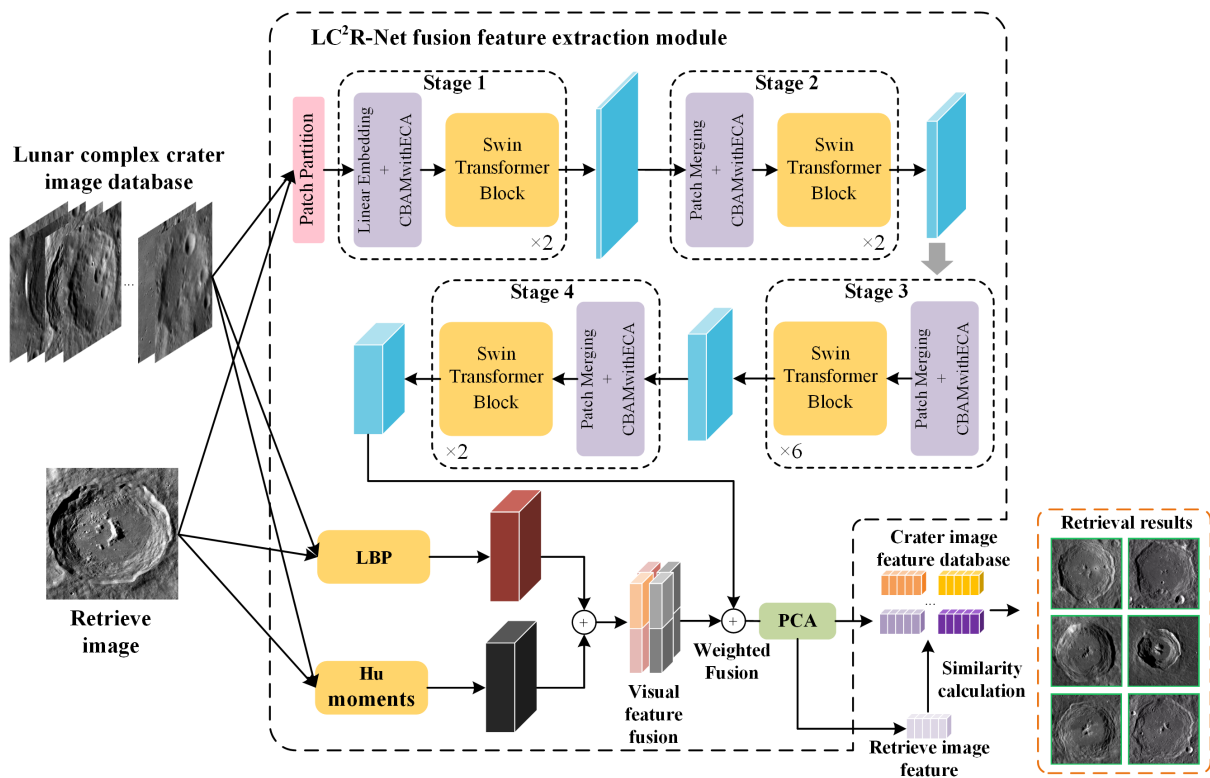


Figure 1. Image retrieval framework for lunar complex crater images based on the LC²R-Net.

3.1. Visual Feature Extraction

In order to effectively capture the distinctive visual attributes of intricate lunar crater images, we utilized two resilient techniques for visual feature extraction: LBP [48] and Hu Moments [49]. LBP serves as a potent texture descriptor, capturing local texture nuances within an image by contrasting the grayscale intensity of a pixel with its neighboring pixels. To elaborate, the LBP value for each pixel is computed by juxtaposing the grayscale intensity of the surrounding pixels with that of the central pixel. This operation can be mathematically represented as follows:

$$LBP(x_c, y_c) = \sum_{i=0}^{P-1} s(I(x_i, y_i) - I(x_c, y_c))2^i \tag{1}$$

where P is the number of pixels in the domain, and $s(z)$ is a sign function defined as:

$$s(z) = \begin{cases} 1, & z \geq 0 \\ 0, & z < 0 \end{cases} \tag{2}$$

In this way, for each pixel point $I(x_c, y_c)$ is transformed into a P -bit binary number, i.e., the LBP code, which reflects the texture structure of the region around that pixel. By computing the histogram of the LBP code for the entire image, we can obtain the feature vector characterizing the texture of the image.

Hu moments are similarly employed to serve as descriptors of shape features, encapsulating the geometric characteristics of an image through the synthesis of its central moments, which are inherently invariant to transformations such as translation, scaling, and rotation of the image. The initial step in this process involves the calculation of the image's raw moments and central moments. Raw moments are defined by the following equation:

$$m_{pq} = \sum_x \sum_y x^p y^q I(x, y) \quad (3)$$

where $I(x, y)$ is the pixel intensity of the image at coordinate (x, y) , and p and q are the orders of the moments. Then, the center of mass of the image is computed using the original moments (\bar{x}, \bar{y}) :

$$\bar{x} = \frac{m_{10}}{m_{00}}, \bar{y} = \frac{m_{01}}{m_{00}} \quad (4)$$

where m_{00} is the zero-order primitive moment, representing the total luminance of the image, and m_{10} and m_{01} are the first-order primitive moments, related to the position of the center of mass of the image in the x and y directions. Subsequently, the center moments with respect to the center of mass are calculated:

$$\mu_{pq} = \sum_x \sum_y (x - \bar{x})^p (y - \bar{y})^q I(x, y) \quad (5)$$

The central moment describes the shape of the image, and seven Hu moments are calculated from the central moment in the following form:

$$\begin{aligned} H_1 &= \mu_{20} + \mu_{02} \\ H_2 &= (\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2 \\ H_3 &= (\mu_{30} - 3\mu_{12})^2 + (3\mu_{21} - \mu_{03})^2 \\ H_4 &= (\mu_{30} + \mu_{12})^2 + (\mu_{21} + \mu_{03})^2 \\ H_5 &= (\mu_{30} - 3\mu_{12})(\mu_{30} + \mu_{12}) \left[(\mu_{30} + \mu_{12})^2 - 3(\mu_{21} + \mu_{03})^2 \right] \\ &+ (3\mu_{21} - \mu_{03})(\mu_{21} + \mu_{03}) \left[3(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2 \right] \\ H_6 &= (\mu_{20} - \mu_{02}) \left[(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2 \right] \\ &+ 4\mu_{11}(\mu_{30} + \mu_{12})(\mu_{21} + \mu_{03}) \\ H_7 &= (3\mu_{21} - \mu_{03})(\mu_{30} + \mu_{12}) \left[(\mu_{30} + \mu_{12})^2 - 3(\mu_{21} + \mu_{03})^2 \right] \\ &- (\mu_{30} - 3\mu_{12})(\mu_{21} + \mu_{03}) \left[3(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2 \right] \end{aligned} \quad (6)$$

where μ_{ij} is the central moment with respect to the center of mass, and H_i is the i th Hu moment.

3.2. Deep Feature Extraction

The deep features extracted through neural networks can effectively describe the semantic information of complex lunar impact crater images. The strategy for extracting deep features in this paper is to use the Swin Transformer as the backbone network, removing the classification head from the network to extract deep feature representations. In addition, this paper introduces the CBAMwithECA module at the patch merging layer and the linear embedding layer of the Swin Transformer. The CBAMwithECA module combines the spatial attention mechanism of Effective Channel Attention (ECA) and Convolutional Block

Attention Module (CBAM), mining information in both the channel and spatial dimensions of the feature map. This achieves comprehensive calibration and optimization of features, further enhancing the model's capability to express features.

3.2.1. Backbone: Swin Transformer

The core advantage of the Swin Transformer lies in its unique hierarchical structure, which encodes images via a partitioning strategy, thereby effectively capturing multi-scale features within the image. Specifically, the input image is segmented into patches by the Patch Partition module, followed by the construction of feature maps at varying scales through four stages. Beyond the initial stage, which begins with a Linear Embedding layer, the subsequent three stages each commence with a Patch Merging operation and then proceed with a series of stacked Swin Transformer Blocks to achieve a deep feature representation of the image.

1. Patch Partition

At the outset of the Swin Transformer's processing pipeline, the Patch Partition layer plays a pivotal role in decomposing the incoming image into a sequential array of patches. Given an image with dimensions $H \times W \times 3$, where H and W denote the height and width, and the numeral 3 indicates the RGB color channels, this layer segments the image into a grid of 4×4 patches. These patches are subsequently flattened along the channel dimension and undergo a linear projection into an elevated dimensional space, culminating in a feature map with dimensions of $\frac{H}{4} \times \frac{W}{4} \times 48$. This feature map is then subject to a linear transformation within the Linear Embedding layer, producing an output feature map dimensionally characterized as $\frac{H}{4} \times \frac{W}{4} \times C$. The ensuing feature map is channeled into the initial Swin Transformer Block, referred to as Stage 1, for additional refinement. This mechanism is conceptually analogous to the convolutional operation found in conventional convolutional neural networks, and the intricacies of this process are graphically depicted in Figure 2.

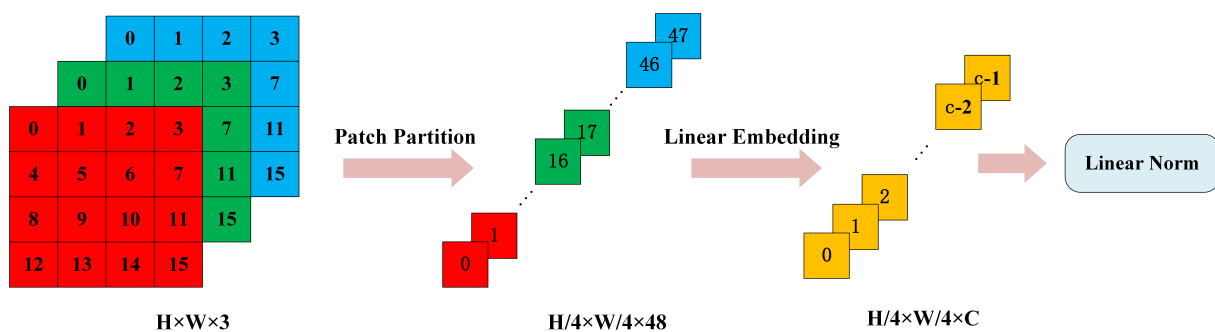


Figure 2. Schematic diagram of the Patch Partition operation.

2. Patch Merging

The Patch Merging technique in the Swin Transformer architecture functions analogously to the pooling layers found in classical convolutional neural networks, effectively generating a pyramidal hierarchy of representations through the downsampling of feature maps. Imagine an input feature map with dimensions $H \times W \times C$. The Patch Merging operation commences with the segmentation of the feature map into distinct 2×2 pixel blocks, treating each as a separate patch. Within these patches, corresponding pixels are extracted and amalgamated, yielding four distinct feature maps, each with a reduced size of $\frac{H}{2} \times \frac{W}{2} \times C$. These quartet of feature maps are then concatenated along the channel axis, resulting in a singular, enhanced feature map with dimensions $\frac{H}{2} \times \frac{W}{2} \times 4C$.

Following the concatenation, the resultant feature map is normalized by a LayerNorm layer, which precedes the final transformation. A fully connected layer then undertakes a linear transformation on the concatenated feature map, specifically targeting its channel

depth. This transformation modifies the channel depth from $4C$ to $2C$, effectively halving it. The procedural specifics of this Patch Merging operation are visually detailed in Figure 3.

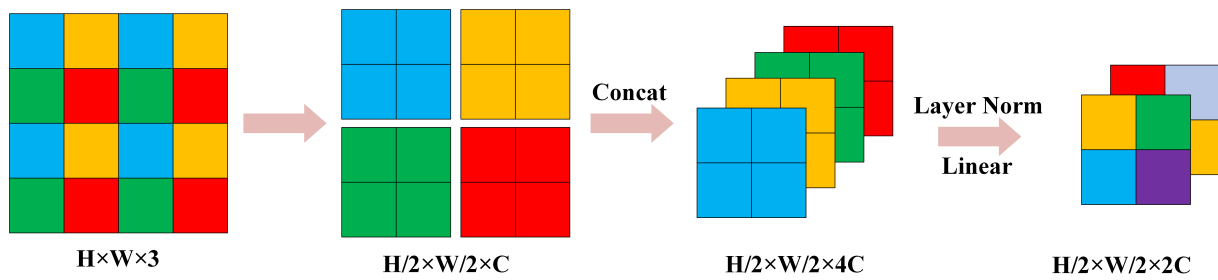


Figure 3. Schematic diagram of the Patch Merging operation.

3. Swin Transformer Block

The Swin Transformer Block represents the fundamental building block of the Swin Transformer architecture. As illustrated in Figure 4, this block is structured as a sequence of two Transformer Blocks. Each Transformer Block is crafted from a series of components: an initial layer normalization (LN), a multi-head self-attention mechanism (MSA), a subsequent layer normalization (LN), and a multilayer perceptron (MLP). To facilitate stable training and mitigate the vanishing gradient issue in deep networks, both the MSA and MLP are equipped with skip connections.

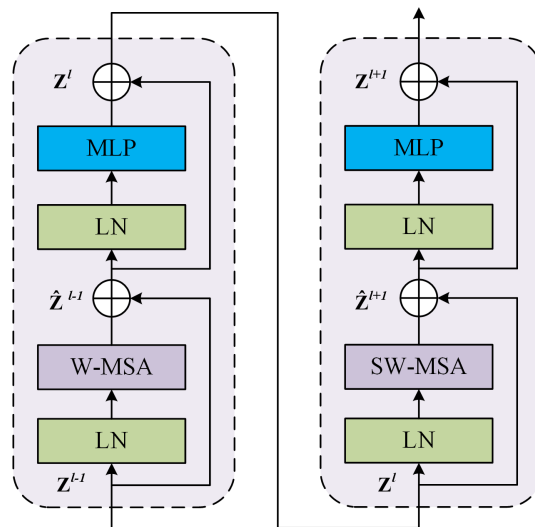


Figure 4. Swin Transformer Block.

The distinguishing feature between the two Transformer Blocks within the Swin Transformer Block is the type of self-attention mechanism employed. The first block integrates a window-based multi-head self-attention ($W-MSA$), which confines the self-attention process within predetermined window boundaries to lower computational demands and hone in on local feature extraction. Conversely, the second block incorporates shifted window multi-head self-attention ($SW-MSA$). By offsetting the window alignment, $SW-MSA$ broadens the receptive field of the model, enabling feature interactions across neighboring windows, which in turn amplifies the model’s global contextual comprehension. This operation is encapsulated in Equation (7):

$$\begin{aligned}
 \hat{Z}^l &= W\text{-MSA} \left[LN \left(Z^{l-1} \right) \right] + Z^{l-1} \\
 Z^l &= MLP \left[LN \left(\hat{Z}^l \right) \right] + \hat{Z}^l \\
 \hat{Z}^{l+1} &= SW\text{-MSA} \left[E \left(Z^l \right) \right] + Z^l \\
 Z^{l+1} &= MLP \left[LN \left(\hat{Z}^{l+1} \right) \right] + \hat{Z}^{l+1}
 \end{aligned}
 \tag{7}$$

where Z^{l-1} and Z^{l+1} represent the input and output of the Swin Transformer Block, respectively, while $W\text{-MSA}$, $SW\text{-MSA}$, and MLP denote the window-based multi-head self-attention, the shifted window multi-head self-attention, and the multilayer perceptron modules, respectively.

3.2.2. CBAMwithECA Attention Module

Due to the high homogeneity and rich detail of lunar complex crater imagery, relying solely on the self-attention mechanism is insufficient to fully capture the prominent features of impact craters. Therefore, we have introduced the CBAMwithECA module [50], which combines the channel attention of ECA [51] and the spatial attention of CBAM [52] to further enhance the representational capability of features. As shown in Figure 5, the core of ECA-Net is the adaptive computation of the size k of the one-dimensional convolutional kernel, which depends on the number of input channels C and the hyperparameters γ and b . The formula is calculated as follows:

$$k = \left\lfloor \frac{\log_2(C) + b}{\gamma} \right\rfloor_{odd}
 \tag{8}$$

where $\lfloor t \rfloor_{odd}$ represents the odd number closest to t , ensuring that the 1D convolutional kernel has symmetric padding. In ECA-Net, adaptive average pooling and a 1D convolutional layer are used to learn the channel attention weights:

$$M_{channel} = \sigma(\text{Conv1D}(\text{AvgPool}(x)))
 \tag{9}$$

where the input feature map is denoted by x and is a four-dimensional tensor within the real number space $\mathbb{R}^{B \times C \times H \times W}$, where B , C , H , and W represent the batch size, number of channels, height, and width, respectively. The channel attention mechanism is encapsulated by $M_{channel}$, which is a tensor of dimensions $\mathbb{R}^{B \times C \times 1 \times 1}$, capturing the importance of each channel. The Sigmoid function, symbolized by σ , is utilized to activate and normalize the elements of $M_{channel}$. Subsequently, the feature map x is modulated by $M_{channel}$ to produce the channel-wise enhanced feature map x_{ca} , which is formulated as follows:

$$x_{ca} = M_{channel} \odot x
 \tag{10}$$

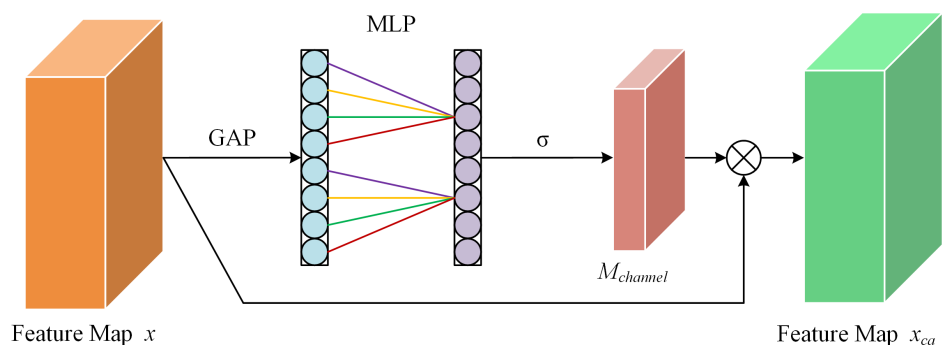


Figure 5. ECA channel attention module.

Concerning the mechanism for spatial attention, illustrated in Figure 6, the process begins by subjecting the feature map x_{ca} to both average pooling and max pooling operations across the channel axis, resulting in a pair of distinct feature descriptors. Subsequently, these descriptors are merged and proceed through a convolutional layer with a kernel size of 7×7 , culminating in the formation of the spatial attention map.

$$M_{spatial} = \sigma(\text{Conv2D}(\text{Concat}(\text{Avgpool}(x_{ca}), \text{Maxpool}(x_{ca})))) \tag{11}$$

The spatial attention map is then applied to the feature map x_{ca} , resulting in a weighted feature map.

$$x_{sa} = M_{spatial} \odot x_{ca} \tag{12}$$

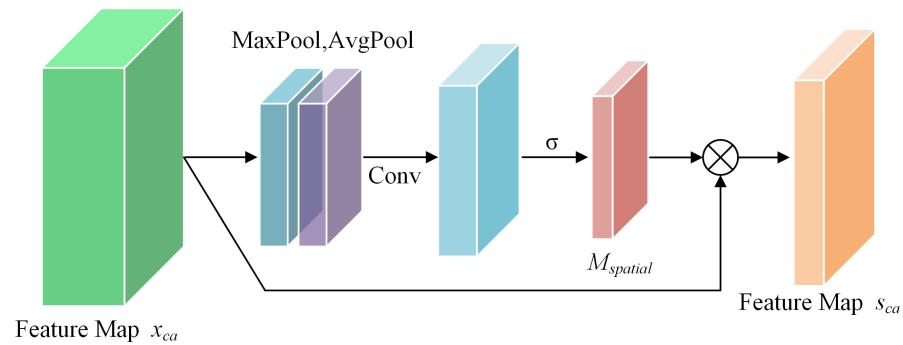


Figure 6. CBAM spatial attention module.

Finally, the feature map weighted by the attention mechanism is added to the original input feature map to realize a residual connection, resulting in the enhanced feature map $x_{enhanced}$:

$$x_{enhanced} = x + x_{sa} \tag{13}$$

The process of inserting the CBAMwithECA module into the linear embedding module is illustrated in Figure 7.

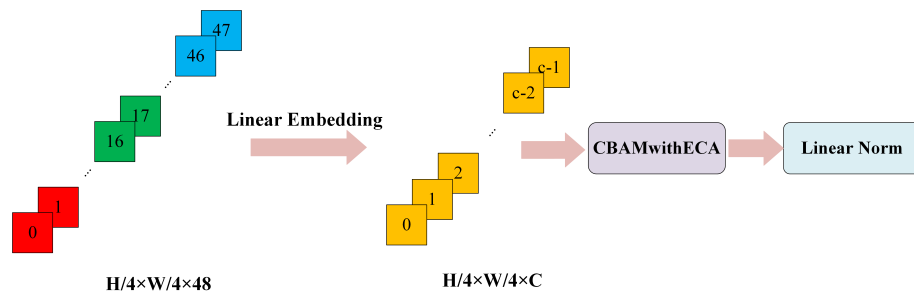


Figure 7. Insertion of the CBAMwithECA module into the linear embedding module.

The process of inserting the CBAMwithECA module into the patch merging module is shown in Figure 8.

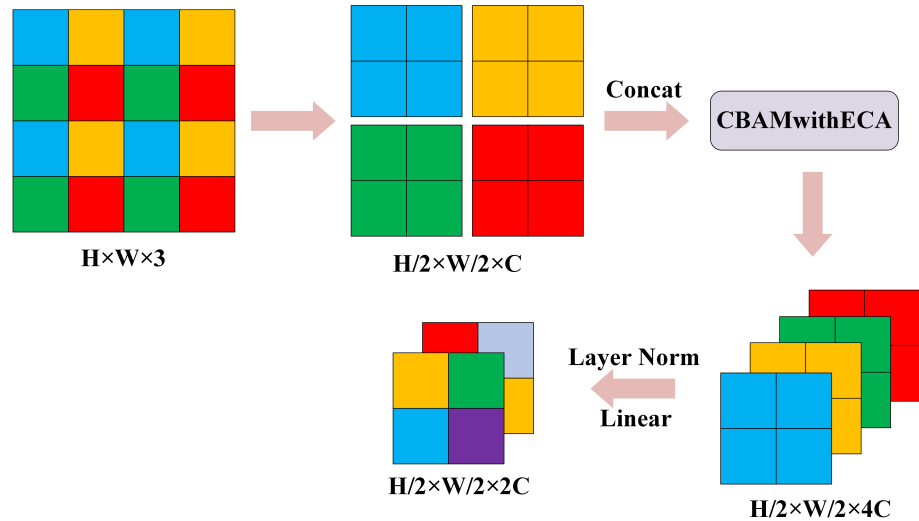


Figure 8. Insertion of the CBAMwithECA module into the patch merging module.

3.2.3. Loss Function

In our research, the training phase utilized a triplet loss function [53] to guide the optimization. This loss function operates on triplets, which include an anchor image, a corresponding positive image, and a contrasting negative image. The primary objective is to amplify the model’s ability to discriminate between varying classes. This is achieved by diminishing the distance metric between the anchor and the positive instance while concurrently enlarging the gap between the anchor and the negative instance. The functional form of the triplet loss is delineated below:

$$L_{triplet} = \sum_{i=1}^N \max(0, \|f(a_i) - f(p_i)\|^2 - \|f(a_i) - f(n_i)\|^2 + \alpha) \quad (14)$$

where $f(\cdot)$ denotes the output of the feature mapping function derived from the Gempool layer. The terms a_i , p_i , and n_i correspond to the anchor, positive, and negative images within the i -th triplet, respectively. Here, N signifies the aggregate count of triplets, while α represents a predetermined margin parameter that delineates the threshold between the proximities of positive and negative pairs. This loss function is instrumental in clustering akin features of images and concurrently dispersing the features of dissimilar images, an aspect that is crucial in areas where the terrain, such as the lunar surface, exhibits high degrees of similarity.

During training, a hard negative sample mining strategy [53] was employed to enhance the effectiveness of the triplet training. For each anchor image, we select negative samples with lower structural similarity by calculating the Structural Similarity Index $SSIM$ with all negative samples in the dataset. $SSIM$ is used to quantify the visual similarity between two images, and its formula is as follows:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (15)$$

where x and y are utilized to denote two distinct image windows. The terms μ_x and μ_y refer to their respective mean intensity values. Variance for each window is indicated by σ_x^2 and σ_y^2 , while σ_{xy} represents the covariance between the two windows. Constants c_1 and c_2 are incorporated within the formulation to prevent the occurrence of division by zero, ensuring numerical stability.

During the hard negative sample mining process, for each anchor image a , we select the negative sample n with the smallest $SSIM$ value from all negative sample images, satisfying the following condition:

$$n = \arg \min_{n'} d(f(a), f(n')) \quad (16)$$

The hard negative mining strategy ensures that the anchor image and the selected negative sample image have significant structural differences, providing more challenging samples for training and enhancing the model's discriminative ability.

3.3. Feature Fusion and Retrieval

Feature fusion is used to concatenate visual features and depth features to get a fused feature vector. Let the visual feature be F_{VC} and the depth feature be F_{DC} . Set the feature weight of F_{VC} to λ and the feature weight of F_{DC} to $1 - \lambda$. Change the importance of the feature by adjusting the size of λ . The fused feature is shown in Equation (17).

$$F = (\lambda F_{VC}, (1 - \lambda) F_{DC}) \quad (17)$$

Since the dimensionality of the fused feature vectors is too high, the fused features are downsampled using PCA. The principle is to maximize the variance of the downsampled features, and if the downsampled features are uncorrelated, then it can be expressed as an optimization problem, as shown in Equation (18).

$$\max_W tr(W^T S_i W), s, t, W^T W = I \quad (18)$$

where S_i represents the covariance matrix of the sample features, $tr(W^T S_i W)$ is the variance of the sample features after dimensionality reduction, $w^T w^T = I$ denotes the constraint conditions, and I is the identity matrix.

After the dimensionality reduction in the fused features, cosine similarity is used to calculate the similarity between different impact crater images, as shown in Equation (19).

$$\cos\theta = \frac{\sum_{i=1}^n (A_i \times B_i)}{\sqrt{\sum_{i=1}^n A_i^2 \times \sum_{i=1}^n B_i^2}} \quad (19)$$

where A_i denotes the composite feature vector derived from the query image, whereas B_i signifies the composite feature vector corresponding to the lunar impact crater images within the image repository. The ultimate retrieval outcomes are the k highest-ranked images determined by their respective cosine similarity measures.

4. Lunar Complex Crater Dataset

The lunar surface is home to a multitude of impact craters that cover much of its terrain. To date, a vast number of lunar craters have been identified in images and Digital Elevation Model (DEM) data through expert visual inspection as well as automated detection methods, leading to the establishment of numerous crater databases. This paper selects 3234 craters ranging from 20 to 30 km in diameter from the lunar impact crater database (2015 revision) maintained by the Lunar and Planetary Institute as the research subjects to construct the Lunar Complex Impact Crater Dataset; the data can be obtained from https://www.lpi.usra.edu/lunar/surface/Lunar_Impact_Crater_Database_v08Sep2015.xls (accessed on 1 January 2020). Utilizing 100-m resolution imagery and DEM data provided by the Lunar Reconnaissance Orbiter (LRO), an analysis based on the morphological texture features and profile characteristics of the craters is conducted (when a crater contains two or more types of local structures, the most prominent feature is chosen as the basis for classification). These craters are categorized into six types, including simple craters, floor-fractured craters, central peak craters, multi-impacted floor craters, lunar oceanic

remnant impact craters, and impact residual craters. Example images for each category are shown in Figure 9.

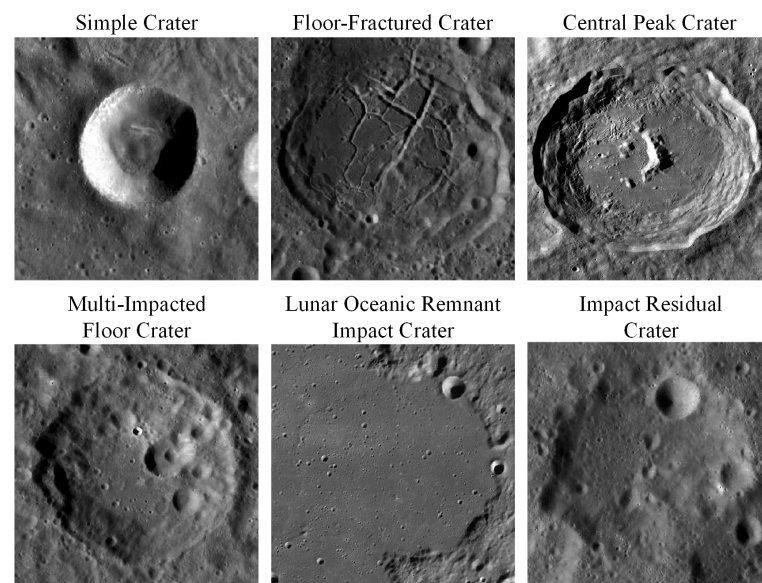


Figure 9. Images of six different types of impact crater samples: simple craters, floor-fractured craters, central peak craters, multi-impacted floor craters, lunar oceanic remnant impact craters, and impact residual craters.

Due to the specificity of impact crater types, the number of different categories of impact craters in the constructed dataset is severely imbalanced. To prevent overfitting during network training, we employed a series of data augmentation techniques to expand the original dataset. These techniques include random rotation, random horizontal flipping, color jittering, random affine transformations, and random Gaussian blur, all aimed at simulating the various conditions that impact craters may encounter during actual imaging processes. Ultimately, we obtained 5597 images, of which 80% were randomly selected to constitute the training data, with the remainder used for model validation.

5. Experiments and Analysis

This section presents a comprehensive evaluation of the performance of the proposed method through a series of extensive experiments and provides a clear and accurate description of the experimental results.

5.1. Implementation Details

5.1.1. Experimental Setup

All experiments in this study were conducted on a deep learning server equipped with an Intel(R) Xeon(R) Platinum 8255C CPU and an RTX 3090 (24GB) GPU. The software environment consisted of Pytorch 1.10.0 and Python 3.8, with the operating system being Ubuntu 20.04. During the model training phase, weights trained on the ImageNet dataset were used as the initial parameters. The model was optimized using the Adam optimizer, and a cosine annealing algorithm was employed to dynamically adjust the learning rate. Parameters were updated every 4 batches, with each batch containing 16 samples. The detailed parameters are shown in Table 1. The experiments returned the top 20 images in the retrieval results to evaluate the model's retrieval accuracy.

Table 1. Experimental parameter configuration.

| Parameter Name | Parameter Configuration |
|-----------------------|-------------------------|
| Initial learning rate | 5×10^{-6} |
| Weight decay | 1×10^{-5} |
| Margin α | 2 |
| Training epochs | 25 |

5.1.2. Evaluation Metrics

During the experimental phase of this research, we employed three principal metrics to assess the efficacy of the lunar complex crater image retrieval system: mean average precision mAP , average normalized modified retrieval rank $ANMRR$, and the time taken for retrieval.

1. Mean Average Precision (mAP)

When performing image retrieval for lunar complex craters, for a given query image and an image database with a total of N images, the Average Precision (AP) is defined as follows:

$$AP = \frac{1}{n} \sum_{k=1}^N P(k) \cdot rel(k) \quad (20)$$

where n is indicative of the aggregate count of images in the repository which are categorized under the identical impact crater classification as the query image. The index k refers to the ordinal position within the ranked retrieval outcomes. The function $P(k)$ quantifies the precision attained at the juncture of the k -th result in the retrieval sequence. The function $rel(k)$ operates as a binary indicator, assigning a value of 1 when the k -th result in the retrieval sequence is of the same impact crater category as the query image, and 0 in all other instances. The mAP , is derived by computing the mean of precision values across all query instances, which is elucidated in Equation (21).

$$mAP = \frac{1}{Q} \sum_{q=1}^Q AP(q) \quad (21)$$

where Q stands for the cumulative quantity of all the queries processed, while $AP(q)$ signifies the Average Precision AP computed for each distinct query. The mAP value, which falls within the interval $[0, 1]$, serves as a performance indicator for the retrieval system; a value approaching 1 denotes the superior performance of the system.

2. Average Normalized Modified Retrieval Rank ($ANMRR$)

In the dataset of images, every image is allocated a ranking $Rank(i)$, with i denoting the image's sequence in the outcome set. Given a query's reference image S_K , the count of analogous images within the dataset is denoted as $G(S_K)$. Within the uppermost K images of the search outcomes, should the $Rank(i)$ of an image surpass K , the $Rank(i)$ is recalibrated as per the subsequent expression:

$$Rank(i) = \begin{cases} Rank(i) & Rank(i) \leq K \\ 1.25 \times K & Rank(i) > K \end{cases} \quad (22)$$

For each query S_K , its average rank $AvgRank(S_K)$ is calculated as follows:

$$AvgRank(S_K) = \frac{1}{G(S_K)} \sum_{i=1}^{G(S_K)} Rank(i) \quad (23)$$

The normalized and corrected retrieval rank is defined as $NMRR(S_K)$:

$$NMRR(S_K) = \frac{AvgRank(S_K) - 0.5 \times (K + 1)}{1.25 \times K - 0.5 \times (K + 1)} \quad (24)$$

In assessing the efficacy of image retrieval approaches within a collection of images, suppose that M queries have been executed. To compute the aggregate mean normalized modified retrieval rank, denoted as $ANMRR$, the following procedure is adopted:

$$ANMRR = \frac{1}{M} \sum_{j=1}^M NMRR(S_{K_j}) \quad (25)$$

The value of $ANMRR$ is within the range $[0, 1]$. It should be noted that the lower the value of $ANMRR$, the higher the retrieval precision.

3. Retrieval Time

The retrieval duration stands as a crucial metric for gauging the performance of an image retrieval system. It spans from the moment the query image is submitted to the point when a full set of search outcomes is obtained. The efficiency of the system is inversely proportional to the retrieval time; the less time it takes to complete the search, the more efficient the system is considered to be.

5.2. Comparison of LC^2R -Net with Other Methods

To verify the effectiveness of the LC^2R -Net model and its advantages over traditional methods in the task of complex lunar crater image retrieval, we selected several widely used convolutional neural network models and Transformer models for comparative analysis. These included VGG16 [54], ResNet101 [55], DenseNet121 [56], EfficientnetV2-S [57], and Vision Transformer (ViT) [58]. The dataset, optimization algorithms, loss functions, and hyperparameters during training were consistent with those used for LC^2R -Net. In LC^2R -Net, λ was set to 0.2, and features were reduced to 128 dimensions using the PCA method. The augmented dataset was used for training, while the original, unmodified dataset was used for testing. The retrieval precision of each model was compared by calculating the mAP for each category. Table 2 presents a detailed comparison of the performance between LC^2R -Net and the aforementioned models. The results indicate that LC^2R -Net achieves better retrieval precision, with the mAP of 83.75%. Compared to VGG16, ResNet101, DenseNet121, EfficientnetV2-S, and Vision Transformer, the mAP of LC^2R -Net is higher by 32.31%, 39.85%, 30.65%, 26.58%, and 21.52%, respectively. These results further demonstrate the significant advantage of LC^2R -Net in integrating low-level visual features and deep features for lunar image retrieval, achieving more precise retrieval results compared to methods relying on the deep features of traditional CNN models.

Table 2. Mean average precision by category on the lunar complex crater dataset for different methods.

| Category | Methods | | | | | |
|-------------------------------------|---------|-----------|-------------|------------------|--------|--------------|
| | VGG16 | ResNet101 | DenseNet121 | EfficientnetV2-S | ViT | LC^2R -Net |
| Simple Crater | 55.33% | 54.99% | 58.73% | 63.77% | 61.91% | 80.82% |
| Floor-Fractured Crater | 31.32% | 24.05% | 53.51% | 42.08% | 55.33% | 99.77% |
| Central Peak Crater | 50.80% | 43.83% | 47.06% | 58.33% | 52.87% | 70.52% |
| Multi-Impacted Floor Crater | 43.50% | 38.99% | 41.88% | 45.44% | 46.08% | 64.32% |
| Lunar Oceanic Remnant Impact Crater | 80.62% | 59.98% | 73.61% | 81.40% | 89.78% | 98.22% |
| Impact Residual Crater | 47.09% | 41.56% | 43.81% | 52.03% | 52.23% | 68.44% |
| Average | 51.44% | 43.90% | 53.10% | 57.17% | 62.23% | 83.75% |

In Table 2, the retrieval accuracy for Multi-Impacted Floor Craters and Impact Residual Craters is significantly lower compared to other categories. The reason is that the features

of the crater images in these categories bear a high visual similarity to those of other categories, making it difficult to distinguish between them even with the use of fused features. Nonetheless, in the face of such challenges of feature similarity, the LC²R-Net model still demonstrates superior performance compared to traditional convolutional neural network models that rely solely on deep features. This indicates the effectiveness of LC²R-Net in integrating multi-level features, particularly in dealing with image categories with high feature similarity, significantly enhancing the accuracy of retrieval. It is noteworthy that among the mentioned convolutional neural networks, EfficientNetV2-S significantly outperforms VGG16, ResNet-101, and DenseNet121. The reason lies in EfficientNetV2-S's effective balancing of model depth, width, and resolution through scaling methods and the introduction of several novel architectures, thereby preserving more image detail information, which is crucial for retrieval tasks. Furthermore, the Vision Transformer surpasses traditional convolutional neural network models in performance, indicating that models based on self-attention mechanisms can more effectively capture global dependencies, thereby enhancing the model's generalization capability.

To more visually demonstrate the effectiveness of LC²R-Net, Figures 10–12 present some retrieval examples. Taking the top 10 returned images as an example, the retrieval results of LC²R-Net are shown in Figure 10, and the comparative retrieval results of LC²R-Net and other methods are shown in Figures 11 and 12.

5.3. Ablation Study

To evaluate the performance of the LC²R-Net model in the task of image retrieval for complex lunar craters, this section conducts ablation experiments on the feature fusion and attention mechanisms within the LC²R-Net network. The experiments are carried out on the complex lunar crater dataset constructed for this paper, utilizing *mAP* and *ANMRR* as metrics to assess retrieval performance. Table 3 presents the ablation study for the attention mechanism.

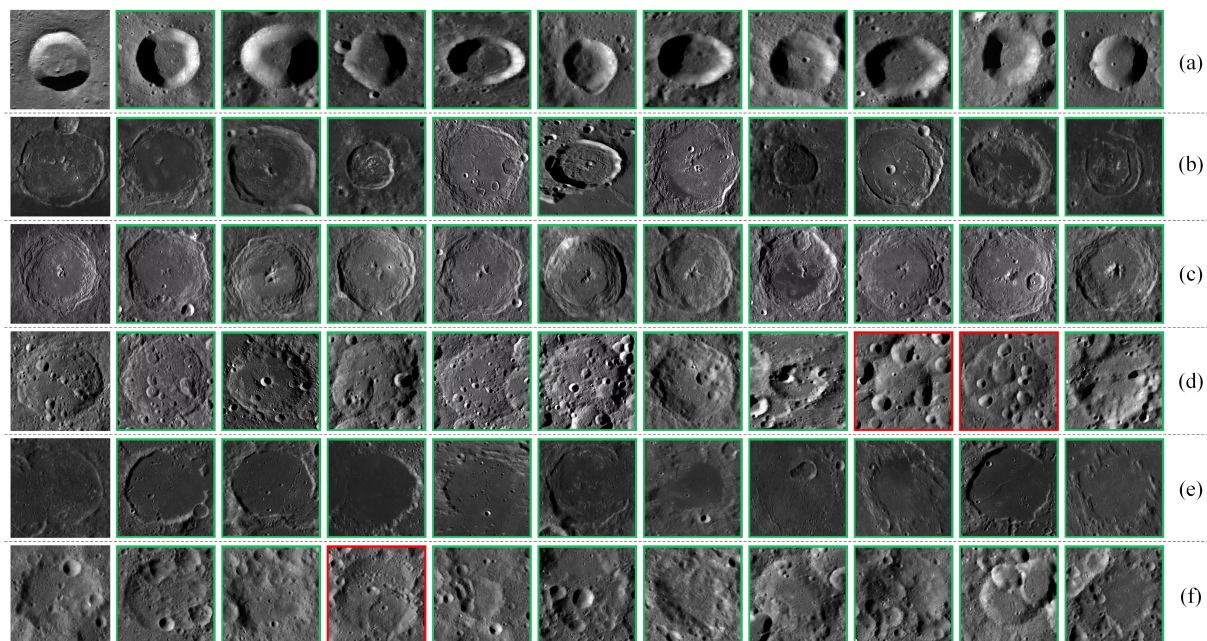


Figure 10. Retrieval results of LC²R-Net for various crater categories (the first image in each row is the query image, green borders indicate correct retrieval results, and red borders indicate incorrect retrieval results): (a) Simple crater. (b) Floor-fractured crater. (c) Central peak crater. (d) Multi-impacted floor crater. (e) Lunar oceanic remnant impact crater. (f) Impact residual crater.

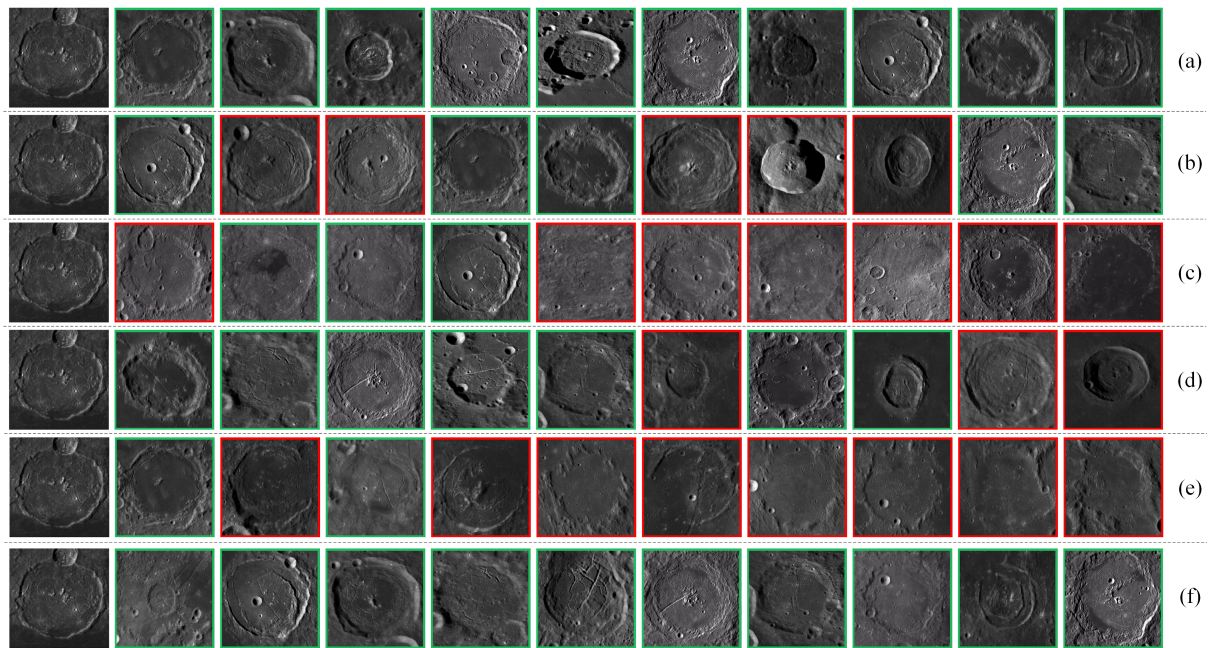


Figure 11. Examples of retrieving central peak craters using different methods (the first image in each row is the query image, green borders indicate correct retrieval results, and red borders indicate incorrect retrieval results): (a) LC²R-Net. (b) VGG-16. (c) ResNet-101. (d) DenseNet-121. (e) EfficientNetV2-S. (f) ViT.

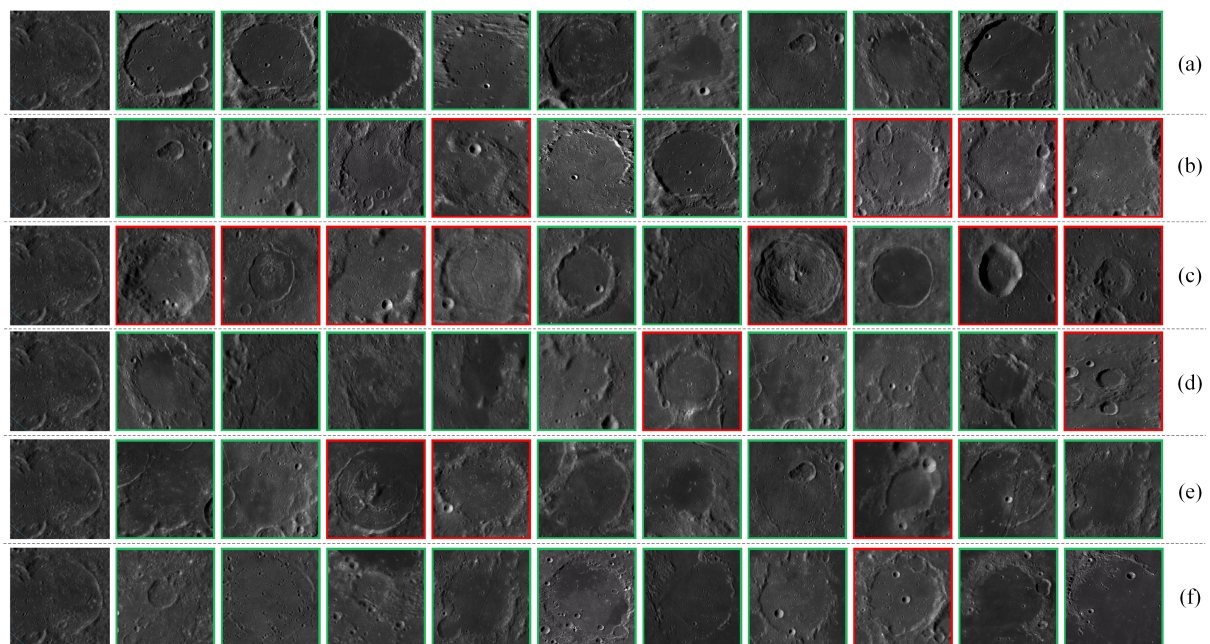


Figure 12. Examples of retrieving Lunar Oceanic Remnant Impact Craters using different methods (the first image in each row is the query image, green borders indicate correct retrieval results, and red borders indicate incorrect retrieval results): (a) LC²R-Net. (b) VGG-16. (c) ResNet-101. (d) DenseNet-121. (e) EfficientNetV2-S. (f) ViT.

Table 3. Ablation study on the attention mechanism.

| Methods | mAP/% | ANMRR |
|-----------------------|-------|--------|
| Swin-T | 83.01 | 0.0755 |
| Swin-T + CBAMwithECA | 83.65 | 0.0725 |
| LC ² R-Net | 83.75 | 0.0721 |

As shown in Table 3, the features extracted using the Swin-T network achieve *mAP* and *ANMRR* of 83.08% and 0.0755, respectively, on the dataset. By integrating the CBAMwith-ECA attention module, the model's *mAP* is improved by 0.64%, and the *ANMRR* is reduced by 0.003. These results confirm that the introduction of attention mechanisms can more effectively highlight key features in images, enhance the discrimination ability for images of different categories, and thereby improve the accuracy of lunar complex crater image retrieval tasks.

Ablation studies were conducted on the feature fusion module with the value of λ set to 0.2. The results are shown in Table 4, where LBP represents texture features, and Hu denotes shape features.

Table 4. Feature fusion ablation study.

| Methods | mAP/% | ANMRR |
|-----------------------|-------|--------|
| LBP | 39.85 | 0.3717 |
| Hu | 29.81 | 0.4064 |
| LBP + Hu | 41.37 | 0.3616 |
| LC ² R-Net | 83.75 | 0.0721 |

The data in Table 4 reveal the limitations of relying solely on visual features for retrieving complex images such as lunar impact craters, resulting in lower image retrieval accuracy. Furthermore, although combining texture (LBP) and shape (Hu) features (LBP + Hu) can improve retrieval performance to some extent, the retrieval accuracy on the complex lunar crater dataset only increased by 1.52% and 11.56%, respectively, when using these features in isolation. However, when deep features were fused, the *mAP* increased by 43.9% and 53.94%, and the *ANMRR* decreased by 0.2996 and 0.3343, respectively. It is noteworthy that the contribution of texture features to retrieval performance was greater than that of shape features, which may be due to the high visual similarity of lunar crater images. These results fully demonstrate the effectiveness of fusing deep and visual features in improving image retrieval accuracy.

5.4. Parametric Analyses

In the LC²R-Net model, the fusion of visual and deep features involves a key parameter λ , which is used to adjust the weight between different features. The specific calculation method is detailed in Section 3.3. This section designs a series of experiments to illustrate the impact of the value of λ on the performance of LC²R-Net by adjusting its value (ranging from 0 to 1, with an interval of 0.1). The features are reduced to 128 dimensions using the PCA method, and the results are shown in Table 5.

From Table 5, it is evident that when the value of λ is set to 0.2, the *mAP* of the LC²R-Net in the lunar complex crater dataset reaches 83.75%, with the *ANMRR* of 0.0721. The retrieval accuracy of the fused features is higher than that of using depth features alone when the value of λ ranges from 0 to 0.3. However, when the value of λ exceeds 0.4, the retrieval accuracy using fused features or visual features alone is lower than that of using depth features alone. This indicates that the depth features extracted by the Swin Transformer are more effective than traditional visual features in performing image retrieval tasks for lunar complex craters.

Table 5. Impact of different λ values on the retrieval performance of LC²R-Net.

| Method | λ | mAP/% | ANMRR |
|-----------------------|-----------|-------|--------|
| LC ² R-Net | 0 | 83.65 | 0.0725 |
| | 0.1 | 83.67 | 0.0716 |
| | 0.2 | 83.75 | 0.0721 |
| | 0.3 | 83.71 | 0.0728 |
| | 0.4 | 83.29 | 0.0756 |
| | 0.5 | 83.19 | 0.0769 |
| | 0.6 | 82.91 | 0.0752 |
| | 0.7 | 82.46 | 0.0798 |
| | 0.8 | 81.73 | 0.0811 |
| | 0.9 | 78.79 | 0.0934 |
| | 1.0 | 37.99 | 0.3943 |

5.5. Comparison of Retrieval Time

In addition to accuracy, retrieval efficiency is also extremely important in practical applications. To evaluate the performance of different models, we conducted tests on the retrieval time for each model using the lunar complex crater dataset. Each model was subjected to 20 retrieval trials, and the average retrieval time was calculated. The retrieval time consumed by each model is shown in Table 6.

The data in Table 6 indicate that as the dimensionality of deep features is reduced, there is a downward trend in model retrieval time. The incorporation of the CBAMwithECA module results in a slight increase in the retrieval time for the Swin-T model. Among all the models compared, the LC²R-Net model, which employs PCA for dimensionality reduction, achieves the shortest retrieval time of only 0.1041 seconds, performing the best among all models. This result demonstrates that the LC²R-Net model successfully reduces the dimensionality and complexity of features while maintaining retrieval efficiency. Additionally, the retrieval time for traditional visual features is also short, which is due to the fact that deep features are denser; even with lower dimensions, they incur greater computational and storage costs compared to sparse visual features. These results highlight the efficiency advantages of the LC²R-Net model in the task of lunar crater image retrieval.

Table 6. Comparison of retrieval time by different methods on the lunar complex crater dataset.

| Methods | Feature Vector Length | Retrieval Times/s |
|-----------------------|-----------------------|-------------------|
| VGG-16 | 4096 | 0.2134 |
| ResNet101 | 2048 | 0.2046 |
| DenseNet121 | 1024 | 0.1922 |
| EfficientNetV2-S | 1280 | 0.1942 |
| ViT | 768 | 0.1878 |
| Swin-T | 768 | 0.1884 |
| Swin-T + CBAMwithECA | 768 | 0.1907 |
| LBP + Hu | 2367 | 0.1630 |
| LC ² R-Net | 128 | 0.1041 |

5.6. Impact of PCA Dimensionality Reduction on Retrieval Accuracy

The LC²R-Net model proposed in this paper initially integrates the low-level visual features with the deep features of lunar crater images to generate a feature vector with 3135 dimensions. Subsequently, to enhance the efficiency of retrieval, PCA is employed for feature dimensionality reduction, enabling more efficient retrieval. Therefore, experiments were conducted with different feature dimensions (16, 32, 64, 128, 256, and the original 3135 dimensions) to observe the impact on retrieval accuracy and retrieval time, with the value of λ set to 0.2. The results are shown in Figure 13.

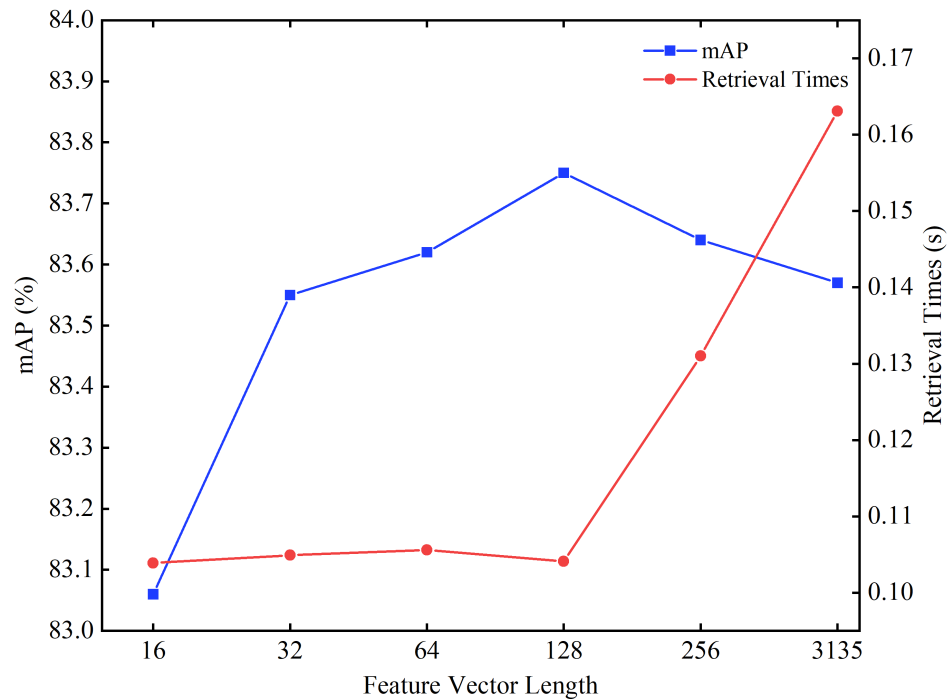


Figure 13. The impact of feature dimensions on the retrieval of lunar complex crater images.

In Figure 13, the retrieval accuracy peaks at a feature vector dimensionality of 128, with the *mAP* reaching 83.75%. Within the range of increasing feature dimensions from 16 to 128, the retrieval accuracy consistently improves. This phenomenon indicates that within this range of dimensions, as the richness of feature information increases, the system is able to more accurately distinguish and retrieve lunar crater images. However, when the feature vector dimensionality exceeds 128, the retrieval accuracy begins to decline. This decrease is due to the excessive expansion of the feature space, which introduces redundant information or increases noise, thereby negatively impacting the model's discriminative ability. When the feature dimensionality is below 128, the retrieval time remains relatively stable, suggesting that at this level of dimensionality, the system's computational efficiency is less affected by the number of features. In contrast, retrieval time significantly increases when the dimensionality exceeds 128, reflecting the computational burden brought about by higher dimensions. These results demonstrate that the PCA dimensionality reduction technique plays a significant role in enhancing the accuracy and efficiency of lunar crater image retrieval.

5.7. The Impact of Data Augmentation on Retrieval Accuracy

In this study, we address the challenge of imbalanced class distribution within our dataset of lunar impact craters, a factor that may lead to overfitting of certain classes by the neural network during the training process. To mitigate this issue, we have employed data augmentation algorithms to expand our dataset and enhance the model's generalization capabilities. To assess the specific impact of data augmentation on the performance of lunar impact crater image retrieval, we conducted model training on both the original dataset and the augmented dataset. Throughout the training process, to ensure comparability of results, we maintained consistency in our algorithmic optimization strategies, loss function, and hyperparameter settings. For LC²R-Net, the λ was set to 0.2, and feature dimensionality was reduced to 128 dimensions using the PCA method. The experimental results are presented in Figure 14.

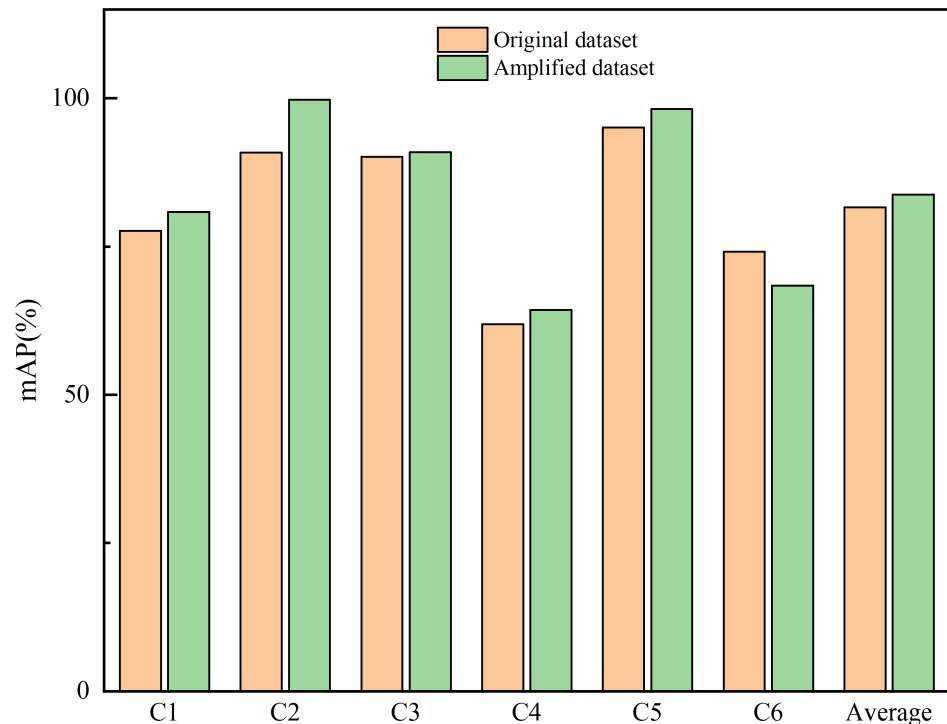


Figure 14. The impact of image augmentation algorithms on the retrieval of lunar complex craters. C1: Simple crater. C2: Floor-fractured crater. C3: Central peak crater. C4: Multi-impacted floor crater. C5: Lunar oceanic remnant impact crater. C6: Impact residual crater.

In Figure 14, it is observed that when the model is trained on the augmented dataset, its performance on the retrieval task is significantly superior to that of the model trained directly on the original dataset. Specifically, the mAP saw a notable increase, improving from 81.63% on the original dataset to 83.75%. This enhancement is reflected not only at a global average level but also across the majority of individual classes, indicating the universality of data augmentation in boosting model performance. However, it is important to note that for the specific category of impact residual crater, the mAP of the model trained on the augmented dataset was actually lower than that of the model trained on the original dataset. This phenomenon suggests that data augmentation does not invariably lead to positive effects. The performance decline in this particular category is attributed to the failure to consider its unique characteristics during augmentation, which hindered the model's ability to effectively discern the differences between impact residual craters and other categories. Therefore, when implementing data augmentation, it is crucial to adopt targeted strategies for different categories to ensure that data augmentation effectively enhances the model's learning and recognition of the distinctions between categories rather than merely increasing the quantity of data.

5.8. Further Discussion

In this study, in order to enhance the model's capability to capture and represent the features of lunar complex crater images, we utilized the CBAMwithECA attention mechanism module during deep feature extraction. To discuss the impact of different attention modules on feature extraction and image retrieval tasks, we conducted comparative experiments by introducing the SE attention mechanism module and the CBAM attention mechanism module at the same position, respectively. The experimental results are shown in Table 7.

Table 7. The impact of different attention mechanisms on lunar complex crater image retrieval tasks.

| Methods | mAP/% | ANMRR |
|-----------------------|-------|--------|
| Swin-T | 83.01 | 0.0755 |
| Swin-T + SE | 82.16 | 0.0795 |
| Swin-T + CBAM | 78.23 | 0.1038 |
| Swin-T + CBAMwithECA | 83.65 | 0.0725 |
| LC ² R-Net | 83.75 | 0.0721 |

As shown in Table 7, the introduction of the SE module and the CBAM module into the Swin-T model did not enhance the model's performance. On the contrary, the addition of these attention mechanisms had a negative impact on the performance of the original Swin-T model. However, upon integrating the CBAMwithECA attention module, the model's performance saw a significant improvement, with the *mAP* increasing by 0.64% and the *ANMRR* decreased by 0.003. It is noteworthy that, in comparison to the attention modules, the SE module outperformed the CBAM module because the SE module provided a more effective feature weighting strategy in channel recalibration. These results indicate that the CBAMwithECA attention module outperforms both the SE and CBAM attention modules in the task of lunar crater image retrieval.

The experimental results adequately substantiate the efficacy of the method we proposed. By integrating the CBAMwithECA module into both the patch embedding and merging modules, LC²R-Net is enabled to capture image details with greater finesse, markedly boosting the model's capability in feature recognition and extraction when dealing with complex crater imagery. Furthermore, we employed a weighted strategy to merge visual and depth features, which not only facilitated an effective complementarity between the two but also accentuated their individual significance. Concurrently, the introduction of a triplet loss function and a hard negative sample mining strategy further encouraged the network to learn more distinctive feature representations, thereby realizing a significant improvement in precision for image retrieval tasks. These results demonstrate that our approach can substantially enhance the model's ability to learn and extract features, significantly improving the accuracy of image retrieval for complex lunar crater imagery.

6. Conclusions

In this paper, we propose the LC²R-Net model, which achieves lunar complex crater image retrieval by fusing the underlying visual features with deep features of images. During the model training phase, we employed a triplet loss function and a hard negative sample mining strategy to generate more distinctive features. In the deep feature extraction stage, we integrated the CBAMwithECA module into the Swin Transformer, successfully capturing the rich details and significant information in lunar crater images, thus enabling better differentiation between different types of lunar complex crater images. In the visual feature extraction stage, we extracted texture and shape features, which effectively complement the deep features. During the feature fusion stage, we introduced feature fusion weights to highlight the importance of different features in retrieval and performed PCA dimensionality reduction after feature fusion, significantly improving the model's retrieval efficiency. We conducted extensive experiments on the lunar complex crater dataset generated in this paper, and the results show that compared to traditional deep learning methods, LC²R-Net achieved the highest retrieval accuracy of 83.75% when the feature fusion weight was set to 0.2 and PCA dimensionality was reduced to 128 while maintaining a fast retrieval speed. Through ablation experiments, we detailed the key role of the CBAMwithECA module and the feature fusion strategy in improving retrieval performance. We explored the impact of different dimensionality reductions on retrieval performance and found that the setting of 128 dimensions offered the best retrieval performance. In addition, we compared the effects of different attention mechanisms on retrieval results, and the experiments proved that the CBAMwithECA attention module performed the best in this study. The LC²R-Net model not only advances the technology of lunar crater

image retrieval but also provides a new perspective for the application of deep learning in the analysis of complex geological images.

In future work, we will consider adopting deep hashing techniques to replace PCA dimensionality reduction to further optimize the precision and efficiency of image retrieval. Secondly, we will explore the feasibility of applying our method to video stream processing. Although current research focuses on single image frames, our proposed network architecture and algorithms can be extended through time series analysis to handle consecutive frames within video streams. This will involve additional training of the network to adapt to dynamic changes. Lastly, we plan to combine object detection methods with image retrieval techniques to explore the detection of different types of impact craters within single image frames to address more realistic application scenarios.

Author Contributions: Y.Z. analyzed the data and wrote the Python source code. Z.K. and Z.C. helped with project and study design, paper writing, and data analysis. 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 used to support the findings of this study are available from the corresponding author upon request.

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

References

1. Hartmann, W.K. Lunar cratering chronology. *Icarus* **1970**, *13*, 299–301. [[CrossRef](#)]
2. Ryder, G. Mass flux in the ancient Earth-Moon system and benign implications for the origin of life on Earth. *J. Geophys. Res. Planets* **2002**, *107*, 6–11. [[CrossRef](#)]
3. Chapman, C.R.; Cohen, B.A.; Grinspoon, D.H. What are the real constraints on the existence and magnitude of the late heavy bombardment? *Icarus* **2007**, *189*, 233–245. [[CrossRef](#)]
4. Bottke, W.F.; Norman, M.D. The late heavy bombardment. *Annu. Rev. Earth Planet. Sci.* **2017**, *45*, 619–647. [[CrossRef](#)]
5. Chen, M.; Lin, H.; Wen, Y.; He, L.; Hu, M. Sino-VirtualMoon: A 3D web platform using Chang’e-1 data for collaborative research. *Planet. Space Sci.* **2012**, *65*, 130–136. [[CrossRef](#)]
6. Di, K.; Li, W.; Yue, Z.; Sun, Y.; Liu, Y. A machine learning approach to crater detection from topographic data. *Adv. Space Res.* **2014**, *54*, 2419–2429. [[CrossRef](#)]
7. Sawabe, Y.; Matsunaga, T.; Rokugawa, S. Automated detection and classification of lunar craters using multiple approaches. *Adv. Space Res.* **2006**, *37*, 21–27. [[CrossRef](#)]
8. Vijayan, S.; Vani, K.; Sanjeevi, S. Crater detection, classification and contextual information extraction in lunar images using a novel algorithm. *Icarus* **2013**, *226*, 798–815. [[CrossRef](#)]
9. Yang, C.; Zhao, H.; Bruzzone, L.; Benediktsson, J.A.; Liang, Y.; Liu, B.; Zeng, X.; Guan, R.; Li, C.; Ouyang, Z. Lunar impact crater identification and age estimation with Chang’E data by deep and transfer learning. *Nat. Commun.* **2020**, *11*, 6358. [[CrossRef](#)]
10. Meyer, C.; Deans, M. Content based retrieval of images for planetary exploration. In Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems, San Diego, CA, USA, 29 October–2 November 2007; pp. 1377–1382.
11. Chen, H.Z.; Jing, N.; Wang, J.; Chen, Y.G.; Chen, L. A novel saliency detection method for lunar remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2013**, *11*, 24–28. [[CrossRef](#)]
12. Hua, K.A.; Shaykhian, G.A.; Beil, R.J.; Akpınar, K.; Martin, K.A. Saliency-based CBIR system for exploring lunar surface imagery. In Proceedings of the 2014 ASEE Annual Conference & Exposition, Indianapolis, Indiana, USA, 15–18 June 2014; pp. 24–1065.
13. Tombe, R.; Viriri, S. Adaptive deep co-occurrence feature learning based on classifier-fusion for remote sensing scene classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *14*, 155–164. [[CrossRef](#)]
14. Zhang, Z.; Jiang, T.; Liu, C.; Zhang, L. An effective classification method for hyperspectral image with very high resolution based on encoder–decoder architecture. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *14*, 1509–1519. [[CrossRef](#)]
15. Zhang, Y.; Zheng, X.; Yuan, Y.; Lu, X. Attribute-cooperated convolutional neural network for remote sensing image classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 8358–8371. [[CrossRef](#)]
16. 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](#)]
17. Napoletano, P. Visual descriptors for content-based retrieval of remote-sensing images. *Int. J. Remote Sens.* **2018**, *39*, 1343–1376. [[CrossRef](#)]
18. Ye, F.; Xiao, H.; Zhao, X.; Dong, M.; Luo, W.; Min, W. Remote sensing image retrieval using convolutional neural network features and weighted distance. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 1535–1539. [[CrossRef](#)]

19. Yan, K.; Wang, Y.; Liang, D.; Huang, T.; Tian, Y. Cnn vs. sift for image retrieval: Alternative or complementary? In Proceedings of the 24th ACM international conference on Multimedia, Amsterdam, The Netherlands, 15–19 October 2016; pp. 407–411.
20. Cheng, Q.; Shao, K.; Li, C.; Li, S.; Li, J.; Shao, Z. A distributed system architecture for high-resolution remote sensing image retrieval by combining deep and traditional features. In Proceedings of the Image and Signal Processing for Remote Sensing XXIV, Berlin, Germany, 10–13 September 2018; Volume 10789, pp. 413–432.
21. Zhang, M.; Cheng, Q.; Luo, F.; Ye, L. A triplet nonlocal neural network with dual-anchor triplet loss for high-resolution remote sensing image retrieval. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 2711–2723. [[CrossRef](#)]
22. Cao, R.; Zhang, Q.; Zhu, J.; Li, Q.; Li, Q.; Liu, B.; Qiu, G. Enhancing remote sensing image retrieval using a triplet deep metric learning network. *Int. J. Remote Sens.* **2020**, *41*, 740–751. [[CrossRef](#)]
23. Liu, Y.; Ding, L.; Chen, C.; Liu, Y. Similarity-based unsupervised deep transfer learning for remote sensing image retrieval. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 7872–7889. [[CrossRef](#)]
24. Zhang, Y.; Zheng, X.; Lu, X. Remote Sensing Image Retrieval by Deep Attention Hashing With Distance-Adaptive Ranking. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 4301–4311. [[CrossRef](#)]
25. Ding, C.; Wang, M.; Zhou, Z.; Huang, T.; Wang, X.; Li, J. Siamese transformer network-based similarity metric learning for cross-source remote sensing image retrieval. *Neural Comput. Appl.* **2023**, *35*, 8125–8142. [[CrossRef](#)]
26. Cheng, G.; Li, Z.; Han, J.; Yao, X.; Guo, L. Exploring hierarchical convolutional features for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 6712–6722. [[CrossRef](#)]
27. Chaudhuri, U.; Dey, S.; Datcu, M.; Banerjee, B.; Bhattacharya, A. Interband retrieval and classification using the multilabeled sentinel-2 bigearthnet archive. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 9884–9898. [[CrossRef](#)]
28. Li, Y.; Zhang, Y.; Huang, X.; Yuille, A.L. Deep networks under scene-level supervision for multi-class geospatial object detection from remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2018**, *146*, 182–196. [[CrossRef](#)]
29. Cheng, G.; Li, Q.; Wang, G.; Xie, X.; Min, L.; Han, J. SFRNet: Fine-Grained Oriented Object Recognition via Separate Feature Refinement. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5610510. [[CrossRef](#)]
30. 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, BC, Canada, 11–17 October 2021; pp. 10012–10022.
31. Lin, A.; Chen, B.; Xu, J.; Zhang, Z.; Lu, G.; Zhang, D. Ds-transunet: Dual swin transformer u-net for medical image segmentation. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 4005615. [[CrossRef](#)]
32. Ma, J.; Tang, L.; Fan, F.; Huang, J.; Mei, X.; Ma, Y. SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA J. Autom. Sin.* **2022**, *9*, 1200–1217. [[CrossRef](#)]
33. He, X.; Zhou, Y.; Zhao, J.; Zhang, D.; Yao, R.; Xue, Y. Swin transformer embedding UNet for remote sensing image semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4408715. [[CrossRef](#)]
34. Gao, L.; Liu, H.; Yang, M.; Chen, L.; Wan, Y.; Xiao, Z.; Qian, Y. STransFuse: Fusing swin transformer and convolutional neural network for remote sensing image semantic segmentation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 10990–11003. [[CrossRef](#)]
35. Liu, Z.; Tan, Y.; He, Q.; Xiao, Y. SwinNet: Swin transformer drives edge-aware RGB-D and RGB-T salient object detection. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 4486–4497. [[CrossRef](#)]
36. Tekeste, I.; Demir, B. Advanced local binary patterns for remote sensing image retrieval. In Proceedings of the IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 6855–6858.
37. Aptoula, E. Remote sensing image retrieval with global morphological texture descriptors. *IEEE Trans. Geosci. Remote Sens.* **2013**, *52*, 3023–3034. [[CrossRef](#)]
38. Xie, G.; Guo, B.; Huang, Z.; Zheng, Y.; Yan, Y. Combination of dominant color descriptor and Hu moments in consistent zone for content based image retrieval. *IEEE Access* **2020**, *8*, 146284–146299. [[CrossRef](#)]
39. Chen, H.Z.; Jing, N.; Wang, J.; Chen, Y.G.; Chen, L. Content Based Retrieval for Lunar Exploration Image Databases. In Proceedings of the Database Systems for Advanced Applications: 18th International Conference, DASFAA 2013, Wuhan, China, 22–25 April 2013; Proceedings, Part II 18; Springer: Berlin/Heidelberg, Germany, 2013; pp. 259–266.
40. Wang, S.; Hou, D.; Xing, H. A novel multi-attention fusion network with dilated convolution and label smoothing for remote sensing image retrieval. *Int. J. Remote Sens.* **2022**, *43*, 1306–1322. [[CrossRef](#)]
41. Ye, F.; Chen, S.; Meng, X.; Xin, J. Query-adaptive feature fusion base on convolutional neural networks for remote sensing image retrieval. In Proceedings of the 2021 International Conference on Digital Society and Intelligent Systems (DSInS), Chengdu, China, 3–4 December 2021; pp. 148–151.
42. 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](#)]
43. Chaudhuri, U.; Banerjee, B.; Bhattacharya, A.; Datcu, M. Attention-driven graph convolution network for remote sensing image retrieval. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 8019705. [[CrossRef](#)]
44. Zhong, W.; Jiang, J.; Ma, Y. L2AMF-Net: An L2-Normed Attention and Multi-Scale Fusion Network for Lunar Image Patch Matching. *Remote Sens.* **2022**, *14*, 5156. [[CrossRef](#)]
45. Fan, L.; Zhao, H.; Zhao, H. Global optimization: Combining local loss with result ranking loss in remote sensing image retrieval. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 7011–7026. [[CrossRef](#)]

46. Zhao, H.; Yuan, L.; Zhao, H. Similarity retention loss (SRL) based on deep metric learning for remote sensing image retrieval. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 61. [[CrossRef](#)]
47. Fan, L.; Zhao, H.; Zhao, H. Distribution consistency loss for large-scale remote sensing image retrieval. *Remote Sens.* **2020**, *12*, 175. [[CrossRef](#)]
48. Ojala, T.; Pietikäinen, M.; Harwood, D. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognit.* **1996**, *29*, 51–59. [[CrossRef](#)]
49. Hu, M.K. Visual pattern recognition by moment invariants. *IRE Trans. Inf. Theory* **1962**, *8*, 179–187.
50. Zhu, L.; Geng, X.; Li, Z.; Liu, C. Improving YOLOv5 with attention mechanism for detecting boulders from planetary images. *Remote Sens.* **2021**, *13*, 3776. [[CrossRef](#)]
51. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11534–11542.
52. 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, 8–14 September 2018; pp. 3–19.
53. Balntas, V.; Riba, E.; Ponsa, D.; Mikolajczyk, K. Learning local feature descriptors with triplets and shallow convolutional neural networks. *Bmvc* **2016**, *1*, 3.
54. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
55. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
56. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
57. Tan, M.; Le, Q. Efficientnetv2: Smaller models and faster training. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 10096–10106.
58. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.

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.