



## Article

# Embedded Zero-Shot Image Classification Based on Bidirectional Feature Mapping

Huadong Sun <sup>1,2</sup> , Zhibin Zhen <sup>1,\*</sup>, Yinghui Liu <sup>1</sup>, Xu Zhang <sup>1,2</sup>, Xiaowei Han <sup>1,2</sup> and Pengyi Zhang <sup>1</sup> 

<sup>1</sup> School of Computer and Information Engineering, Harbin University of Commerce, Harbin 150028, China; kof97\_sun@163.com (H.S.); 18133807369@163.com (Y.L.); zx\_hit@163.com (X.Z.); hanxiaowei2017@163.com (X.H.); zpy@s.hrbcu.edu.cn (P.Z.)

<sup>2</sup> Heilongjiang Provincial Key Laboratory of Electronic Commerce and Information Processing, Harbin 150028, China

\* Correspondence: wozzb0608@outlook.com

**Abstract:** The zero-shot image classification technique aims to explore the semantic information shared between seen and unseen classes through visual features and auxiliary information and, based on this semantic information, to complete the knowledge migration from seen to unseen classes in order to complete the classification of unseen class images. Previous zero-shot work has either not extracted enough features to express the relationship between the sample classes or has only used a single feature mapping method, which cannot fully explore the information contained in the features and the connection between the visual–semantic features. To address the above problems, this paper proposes an embedded zero-shot image classification model based on bidirectional feature mapping (BFM). It mainly contains a feature space mapping module, which is dominated by a bidirectional feature mapping network and supplemented with a mapping network from visual to category label semantic feature space. Attention mechanisms based on attribute guidance and visual guidance are further introduced to weight the features to reduce the difference between visual and semantic features to alleviate the modal difference problem, and then the category calibration loss is utilized to assign a larger weight to the unseen class to alleviate the seen class bias problem. The BFM model proposed in this paper has been experimented on three public datasets CUB, SUN, and AWA2, and has achieved 71.9%, 62.8%, and 69.3% and 61.6%, 33.2%, and 66.6% accuracies under traditional and generalized zero-sample image classification settings, respectively. The experimental results verify the superiority of the BFM model in the field of zero-shot image classification.

**Keywords:** zero-shot image classification; knowledge transfer; auxiliary information; attention mechanism; bidirectional feature mapping



**Citation:** Sun, H.; Zhen, Z.; Liu, Y.; Zhang, X.; Han, X.; Zhang, P. Embedded Zero-Shot Image Classification Based on Bidirectional Feature Mapping. *Appl. Sci.* **2024**, *14*, 5230. <https://doi.org/10.3390/app14125230>

Academic Editor: Serafim Kalliadasis

Received: 7 May 2024

Revised: 11 June 2024

Accepted: 14 June 2024

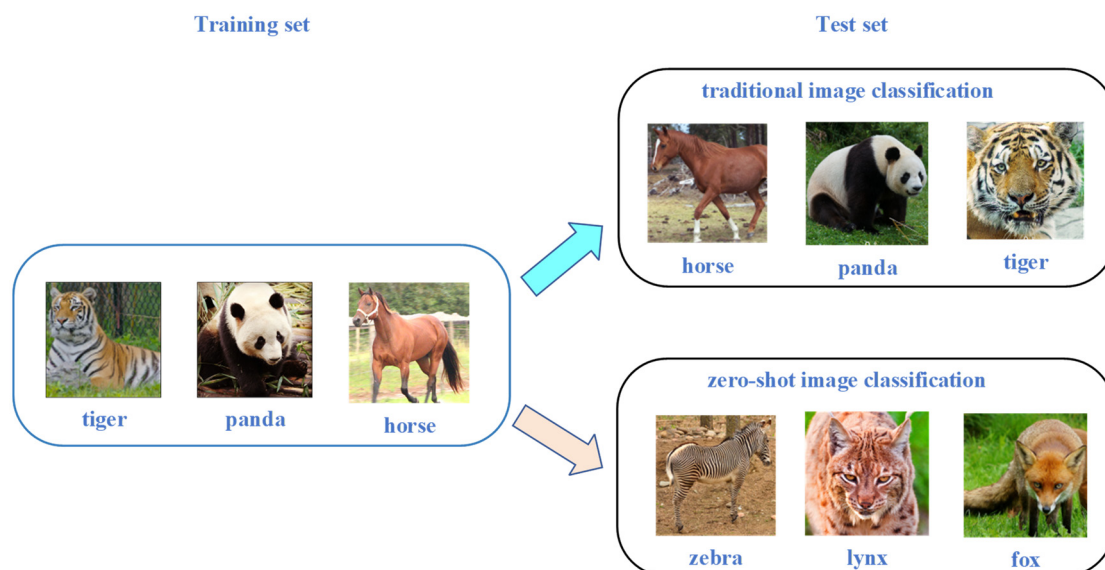
Published: 17 June 2024



**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

Image classification [1] is a basic and critical task in image processing, aiming at recognizing and classifying different images. However, this image classification technique requires a large amount of labeled data for training in order to accomplish the image classification task and can only classify trained categories. In order to classify untrained categories, scholars, inspired by the human cognitive processing of fresh things, proposed the idea of zero-shot learning. The learning logic of zero-shot learning is to mimic the logic of human cognition regarding new things, so that the model can reason on the basis of the knowledge learned on the seen class, so as to achieve the goal of classifying the samples of the unseen class. The difference between zero-shot image classification and traditional image classification is whether the training set contains the category samples from the test set, as shown in Figure 1.

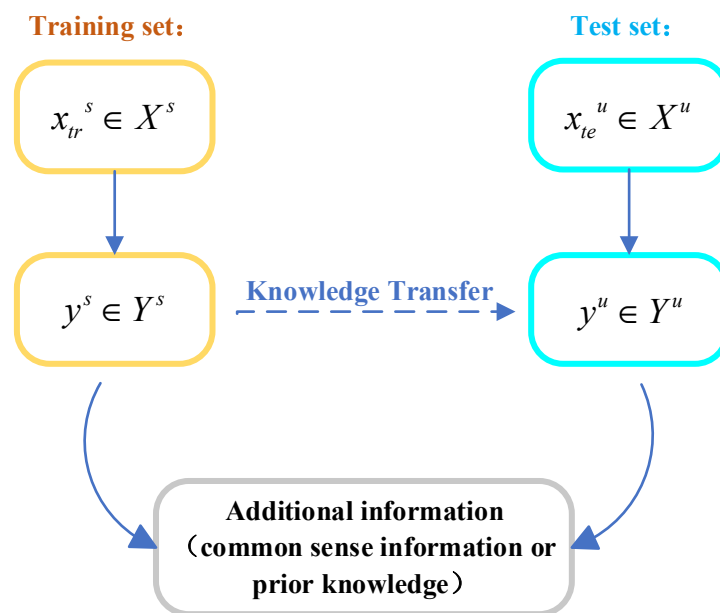


**Figure 1.** Difference between traditional image classification and zero-shot image classification.

Since its development, the idea of zero-shot learning has attracted much attention and has been introduced into various tasks, such as zero-shot image retrieval [2,3] and zero-shot video classification [4] tasks in the field of computer vision, zero-shot sketch retrieval [5] tasks in the cross-modal field, zero-shot semantic segmentation [6,7] and zero-shot text classification [8] tasks in the field of natural language processing, and zero-shot visual quizzing [9] tasks in the multimodal field. The research object of this paper is the zero-shot image classification task, i.e., the zero-shot learning idea is applied to the task of image classification.

Through an example of human cognition regarding new things the learning logic of zero-shot image classification can be understood: a child who has never seen a panda, according to the “panda looks like a bear, the color is like a zebra is black and white” paragraph description, can successfully find the panda in a zoo. This example combines the description of the panda with known things to form the appearance of the panda, which is the transfer of knowledge from the seen class to the unseen class in zero-shot learning. Through the above description, it can be learned that the information about the description of the panda is a bridge between the seen class and the unseen class. Therefore, in order to enable the trained model to reason on the basis of the knowledge learned on the seen class for the purpose of classifying the samples of the unseen class, it is necessary to add the descriptive information for each class in the zero-shot image classification, which is called auxiliary information here to communicate the seen class and the unseen class, as shown in Figure 2.

Zero-shot image classification accomplishes knowledge transfer from seen to unseen classes through partial semantic information shared between the seen and unseen classes. Auxiliary information is usually characterized by relevant common-sense information or prior knowledge and is used to enhance the model’s ability to learn shared partial semantic information. In early zero-shot studies, manually defined attributes (e.g., “wing color”, “beak shape”, etc.) were often used as auxiliary information. With the development of technology, word embeddings [10] of category labels, text embeddings [11] of linguistic descriptions of images, knowledge graphs [12], etc., can be used as auxiliary information. All of these types of auxiliary information can characterize the semantic information of the categories and the similarity between different categories; so they can be used as auxiliary information to help the model realize the knowledge migration from the seen to the unseen classes to complete the zero-shot image classification task. In zero-shot image classification, the quality of auxiliary information is closely related to the effectiveness of knowledge migration and the final model classification result.



**Figure 2.** A learning paradigm for zero-shot image classification.

After the zero-shot image classification technique was proposed, a wide variety of methods have emerged. However, it has been found that most of the current methods can be categorized into two paradigms, one is the embedding-based zero-shot image classification method and the other is the generative-based zero-shot image classification method. Embedding-based methods map image features and semantic information into some space and subsequently classify them according to strategies such as nearest neighbor search. In generalized zero-shot image classification, embedded models usually suffer from the problem of bias towards seen classes. Moreover, previous embedded models usually adopt a single feature mapping approach, such as visual–semantic mapping or semantic–visual mapping, and this single mapping approach cannot fully explore the information embedded in the features and the connection between two features. Although generative-based zero-shot image classification methods can solve the seen class bias problem, they require a large amount of computational resources for training.

Based on the above observations, we propose a feature space mapping module dominated by a bidirectional feature mapping network (BFM), which consists of an attribute-to-visual mapping network and a visual-to-attribute mapping network, which captures a more comprehensive relationship between visual and semantic features to make full use of the information contained in both visual and semantic features. A visual-to-category labeled semantic feature space mapping network is used as a supplement to enable the model to learn richer semantic information. Attention mechanisms based on attribute guidance and visual guidance are then used to weight the features to reduce the discrepancy between visual and semantic features to mitigate the modal difference problem, and then the category calibration loss is used to assign a larger weight to unseen classes and a smaller weight to seen classes to mitigate the seen class bias problem. In addition to this, we introduced the self-attention mechanism to dynamically focus on the sample visual features and enhance the representation of visual features. The experimental results verify the superiority of BFM. Our contributions are as follows: (1) We propose a novel feature space mapping module, which can fully utilize the feature representation information to capture a more comprehensive relationship between visual and semantic features. (2) We introduce a category calibration loss that enables the model to assign a larger weight to the unseen class to alleviate the seen class bias problem that exists in embedded models. (3) We introduce a self-attention mechanism to dynamically focus on the visual features extracted using the pre-trained deep network to extract key visual features and enhance the representation of visual features.

## 2. Related Work

### 2.1. Embedded Zero-Shot Image Classification

Embedded zero-shot image classification methods require learning a feature space mapping function on the seen class in order to achieve knowledge migration from the seen class to the unseen class and to classify the unseen class. Embedding-based methods map image features and semantic information into a certain space and subsequently perform classification based on strategies such as nearest neighbor search. In general, embedded zero-shot image classification methods can be categorized into three types: visual–semantic mapping, semantic–visual mapping, and mapping in a common space. The DeVISE model proposed by Frome et al. [13] uses a pre-trained skip-gram model and a CNN to extract the semantic and visual features of an image as inputs and chooses hinge loss as the model's loss function to construct a deep zero-shot classification model. Yu et al. [14] used a generalized dictionary model to map the visual features and labels of an image into the common space and then used a self-training strategy to incorporate reliable test samples into the model learning process to make the model performance improve. Li et al. [15] proposed a novel ecological supervision-based approach to learn classifier weights by applying knowledge graphs and graph convolution to a comparative learning framework, which accurately exploits the hierarchical structure between target species and explores potential relationships between categories. Kong et al. [16] use visual and semantic information to assist the model in mining inter-class relationships and leveraging learned knowledge, and then use graph convolution to optimize the classifier. Wang et al. [17] design a spatial attention mechanism to extract key visual features and propose a semantic fusion approach to enrich semantic knowledge. Sun et al. [18] mitigate the modal heterogeneity and domain drift problems by decoupling visual semantic features to complement different modal information.

Since embedding-based methods are trained using only seen class samples in the training phase, it leads to the problem that the model will be biased towards seen class samples in the generalized zero-shot image classification task. For this problem, existing methods generally mitigate the problem by designing the loss function or changing the feature mapping approach of the model. Previous embedded models only use a single feature mapping approach to learn the mapping function, which results in the model not being able to fully utilize the information of feature representations and only learn a limited number of feature relationships [13]. Therefore, it is important to deeply explore the intrinsic feature connections between visual and semantic features for the embedded zero-shot image classification task.

### 2.2. Attention Mechanism

The attention mechanism in deep learning originates from the human visual attention mechanism. The human visual attention mechanism is that when people see a picture or an object, they will first quickly scan the whole thing, and then focus on some key areas to help them make a recognition judgment. The core of the visual attention mechanism is to focus on the most useful information for judgment from a large amount of information and then assign it a greater weight for judgment. Similarly, the core of the attention mechanism is to select the information that is more effective for the target task from a large amount of input information and assign it a larger weight to realize the goal of focusing on important information. Based on the properties of the attention mechanism, some scholars have introduced the attention mechanism into the zero-shot image classification task; for example, Xie et al.'s [19] proposed attention mechanism focuses on the image local region at the same time, but also through a kind of thresholding computer system to remove the redundant attention local region. This method assists the zero-shot knowledge migration task by mining the implicit information in the local regions. Huynh et al. [20] find the image local region focused on each attribute through the attention mechanism to mine the association between the attribute and the visual features of the image. Naeem et al. [21] propose a cross-modal attention mechanism, through which semantic embeddings with

visual discriminative properties can be obtained from large-scale documents, which helps to mine the linkage between the visual features and the semantic features. Wang et al. [17] design a spatial attention mechanism to extract key visual features and propose a semantic fusion method to enrich semantic knowledge.

### 3. Model and Proposed Method

#### 3.1. Motivation

Previous work has simply (1) used a pre-trained deep network to extract visual features of an image, ignoring the fitness of that visual feature to a zero-shot classification task [13], (2) employed only the singularity of manually labeled attribute information as auxiliary information, ignoring the semantic information embedded in the category labels [22], or (3) leveraged a single mapping approach that cannot adequately mine the information embedded in the features and the link between the visual features and the semantic links between visual features and semantic features. Based on these observations, we hypothesize that the performance of the embedded zero-shot image classification model is closely related to the expressiveness of visual and semantic features and the potential semantic links between the two features, which provides a solid foundation for effective knowledge transfer.

To enhance the representation of visual features, we introduce the self-attention mechanism to attentively weight the visual features extracted with the pre-trained deep network. The adaptability of the self-attention mechanism lies in its ability to process and analyze each element in the input data (e.g., each pixel in an image or each region in a feature map) and compute the interrelationships among them. This allows the model to capture information on a global scale, not limited to local features. In order to improve the singularity of auxiliary information, we add category labeled word vector information together with manually labeled attribute information as auxiliary information to complete the knowledge migration work. In order to fully explore the information embedded in the features and the connection between visual features and semantic features, we propose a feature space mapping network based on a bidirectional feature mapping approach. The strategy of this network is to fully explore the information embedded in the features from the bidirectional mapping process of visual–semantic and semantic–visual features and to extract the intrinsic connection between the features through attribute-guided and visually guided attention.

#### 3.2. Embedded Zero-Shot Image Classification Based on Bidirectional Feature Mapping

As shown in Figure 3, our proposed embedded zero-shot image classification model based on bidirectional feature mapping includes a feature extraction module and a feature mapping module. The feature extraction module includes visual feature extraction and word vector feature extraction. Visual features are extracted using a pre-trained ResNet101 network; word vectors are extracted using a pre-trained skip-gram language model. The feature mapping module is dominated by a bidirectional feature mapping network with a visual–semantic mapping network as a spoke. The bidirectional feature network learns the intrinsic knowledge of visual attributes, and the visual–semantic mapping network aims to learn richer semantic knowledge. The category calibration loss balances the weights of seen and unseen classes. The overall block diagram of the model is shown in Figure 4.

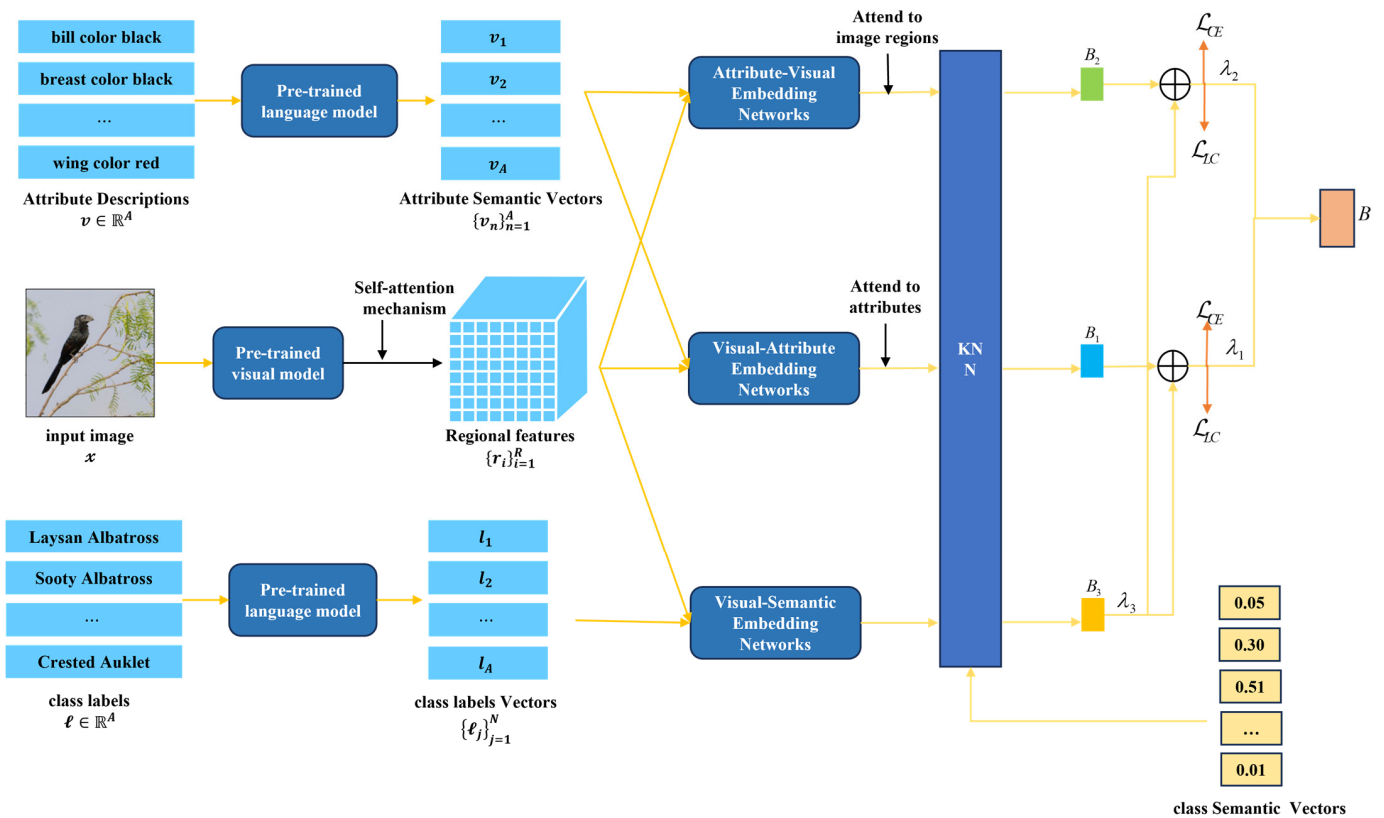


Figure 3. Framework diagram of embedded zero-shot image classification model based on bidirectional feature mapping.

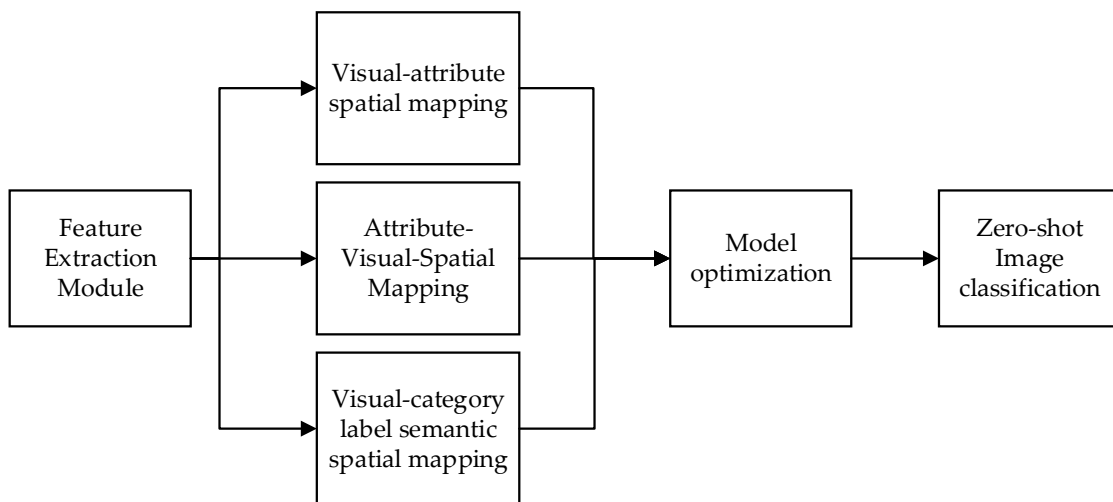


Figure 4. Overall block diagram of the model.

Suppose the training set  $D^s = \{(x_i^s, y_i^s) | x_i^s \in X^s, y_i^s \in Y^s\}$  where  $x_i^s \in X^s$  denotes the seen class samples  $i$ , and  $y_i^s \in Y^s$  denotes their corresponding category labels. The test set  $D^u = \{(x_i^u, y_i^u) | x_i^u \in X^u, y_i^u \in Y^u\}$  where  $x_i^u \in X^u$  denotes the unseen class samples  $i$ , and  $y_i^u \in Y^u$  denotes their corresponding category labels.  $z^c = [z_1^c, \dots, z_A^c]^T$  denotes the category attributes vector that characterizes the relationship between category labels and attributes. Word vector representation for each attribute  $A = \{a_1, \dots, a_K\}$  and each category label  $L = \{l_1, \dots, l_N\}$  are learned using the skip-gram model.



### 3.2.1. Feature Extraction Module

Extracting expressive visual features is important in embedded zero-shot image classification. As the first module of our BFM model, we propose a predominantly attentional mechanism, being designed to be able to compute and highlight the relative importance of each image region, which in turn directs the network to focus more on those features that are particularly important for distinguishing between different categories. Based on these needs, the self-attention mechanism [23] is chosen in this section to be introduced into the model of this paper. The self-attention mechanism is adapted in that it is able to process and analyze each element in the input data (e.g., each pixel in the image or each region in the feature map) and compute the interrelationships between them. This allows the model to capture information on a global scale, not limited to local features. The dot product form of the self-attention mechanism is shown in Equation (1).

$$\text{attention}(Q, K, V) = \text{softmax}(Q * K^T) * V \quad (1)$$

Assuming that the input is  $X = [x_1, \dots, x_N] \in \mathbb{R}^{D_x \times N}$ , the query matrix  $Q = [q_1, \dots, q_N] \in \mathbb{R}^{D_q \times N}$ , the key matrix  $K = [k_1, \dots, k_N] \in \mathbb{R}^{D_k \times N}$ , and the value matrix  $V = [v_1, \dots, v_N] \in \mathbb{R}^{D_v \times N}$ , and they are obtained after three linear mappings:

$$\begin{aligned} Q &= W_q X \in \mathbb{R}^{D_q \times N} \\ K &= W_k X \in \mathbb{R}^{D_k \times N} \\ V &= W_v X \in \mathbb{R}^{D_v \times N} \end{aligned} \quad (2)$$

where  $W_q$ ,  $W_k$ , and  $W_v$  are the learnable parameter matrices, respectively.

In this paper, the self-attention module is placed after the visual features have been extracted using the deep neural network and before the input of the feature space mapping module because, at this stage, it helps the model to better identify and focus on the features that are critical to the classification task, which provides the model with a more flexible and efficient way to understand and characterize the input data.

### 3.2.2. Feature Mapping Module

We propose a feature mapping module that is dominated by a bidirectional feature mapping network and supplemented with a visual–semantic mapping network. This module can fully utilize the feature intrinsic information and can learn more semantic information and the intrinsic connection between visual and semantic features. In the following section, each important module in the bidirectional feature mapping network will be described in detail separately, mainly including the spatial mapping module, the attention module, and the model optimization module.

The purpose of the attention module is mainly to weight the mapped and aligned features so that the model focuses on the information that is more useful for the target task and improves the performance and efficiency of the model. The visual-guided and attribute-guided attention mechanisms are used on the visual attribute features and attribute visual features after mapping alignment, respectively, to further enhance the visual details and attribute semantic information to the aligned features while compensating for possible missing information and to highlight the attribute semantics-related visual regions and attribute-specific semantic information in the image. Schematic diagrams of the visual-guided and attribute-guided attention mechanisms are shown in Figures 5 and 6, respectively.

The corresponding mathematical formulas for the visual-guided and attribute-guided attention-based mechanisms are shown in Equations (3) and (4), respectively:

$$\tau_r^a = \frac{\exp(f^r W_3 v_a)}{\sum_{r=1}^R \exp(f^r W_3 v_a)} \quad (3)$$

$$\beta_a^r = \frac{\exp(v_a W_4 f^r)}{\sum_{a=1}^A \exp(v_a W_4 f^r)} \tag{4}$$

where  $v_a$  denotes the  $a$ th attribute semantic feature vector, and  $f^r$  denotes the visual feature of the  $r$ th local region of the image;  $W_3$  is a learnable matrix for measuring the similarity between the attribute semantic feature and each local visual feature of the sample; and  $W_4$  is a learnable matrix for calculating the similarity between the local visual feature and each attribute semantic feature.

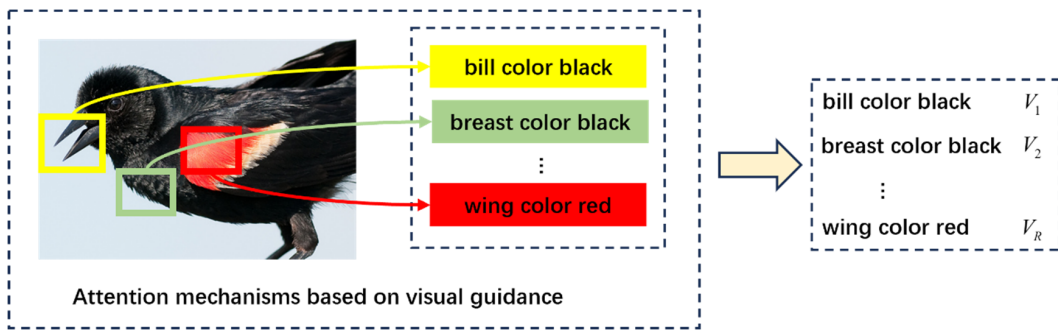


Figure 5. Schematic diagram of the attention mechanism based on visual guidance.

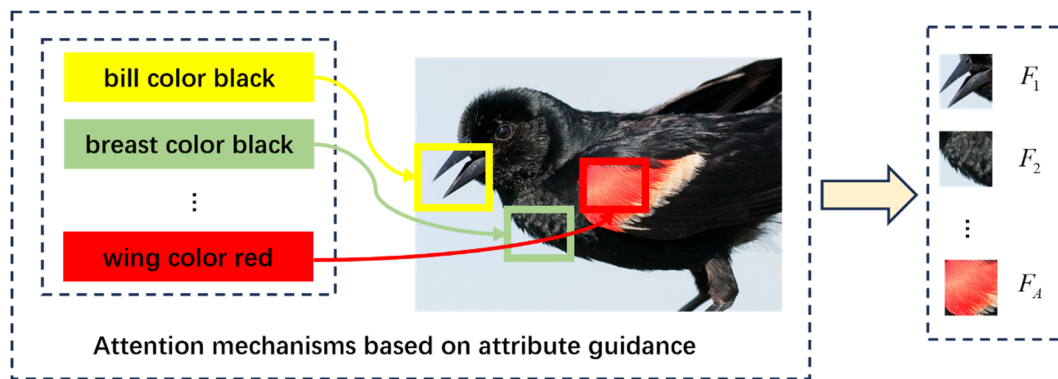


Figure 6. Schematic of attention mechanism based on attribute guidance.

In this way, two sets of attentional weights  $\{\tau_r^a\}_{a=1}^A$  and  $\{\beta_a^r\}_{r=1}^R$  can be obtained. Then, based on these two sets of attentional weights, attentional features can be obtained, i.e., the obtained attentional weights  $\{\tau_r^a\}_{a=1}^A$  and  $\{\beta_a^r\}_{r=1}^R$  are subjected to a weighting operation with visual attribute features and attribute visual features, respectively, which are computed as shown in Equations (5) and (6):

$$V_r = \sum_{a=1}^A \tau_r^a s(f, v) \tag{5}$$

$$F_a = \sum_{r=1}^R \beta_a^r s(v, f) \tag{6}$$

where  $V_r$  and  $F_a$  denote the final obtained visual attribute features and attribute visual features.

The purpose of the spatial mapping module is mainly to learn the feature spatial mapping function, so that the model can realize the knowledge migration from seen to unseen classes more effectively. The main part of the module consists of two branches, i.e., a bidirectional feature mapping network, which includes mapping of attribute semantic features to visual space and mapping of visual features to attribute semantic space. The auxiliary part includes the mapping of visual features to the semantic space of category labels, which is used to assist the learning of the main part so that the model learns more



semantic knowledge. The structure of the feature space mapping module is shown in Figure 7. In the feature space mapping module,  $\mathcal{V}$  denotes the visual feature space,  $\mathcal{A}$  denotes the attribute semantic space,  $\mathcal{L}$  denotes the category label semantic space, and  $\mathcal{Z}$  denotes the category semantic vector space. The circle, diamond, triangle, and rectangle represent the visual features, attribute features, category label semantic features, and category semantic vectors of the sample, respectively.

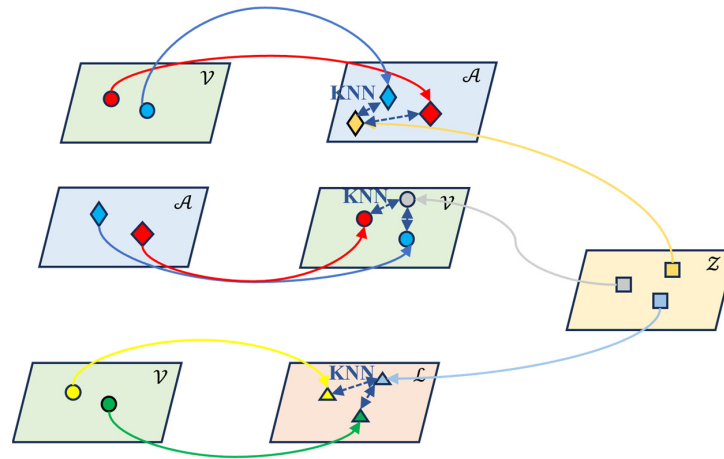


Figure 7. Feature space mapping module.

The main computational process of the main part of the spatial mapping module is as follows: first, the input features, i.e., sample visual features and attribute semantic features, are prepared; second, the visual features and attribute semantic features are mapped to the semantic space and the visual space, respectively, and the semantic features and the visual features in the mapping space are aligned by learning a key matrix, with the formulas as shown in Equations (7) and (8):

$$s(f, v) \triangleq fW_1v \tag{7}$$

$$s(v, f) \triangleq vW_2f \tag{8}$$

where Equation (7) represents the mapping of semantic features to visual space,  $v$  represents attribute semantic features,  $f$  represents sample visual features, and  $W_1$  is a learnable mapping matrix to compute the correlation between visual features and attribute semantic features. Similarly, Equation (8) represents the mapping of visual features to attribute semantic space, and  $W_2$  is a learnable mapping matrix to compute the correlation between attribute semantic features and visual features.

The main computational process of the auxiliary part of the spatial mapping module is as follows: mapping the visual features into the category label semantic space and then aligning the visual features and the category label semantic features in the mapping space by learning a feature mapping matrix as shown in Equation (9):

$$s(f, \ell) \triangleq fW_5\ell \tag{9}$$

where  $\ell$  denotes the category labeling semantic features.

After completing the feature mapping step described above, the two alignment features obtained from the main part of the mapping module are subjected to the visual-guided and attribute-guided attention-based mechanisms, respectively, to obtain the more discriminative alignment features  $V_r$  and  $F_a$ , whose computational formulas are shown in Equations (5) and (6), respectively.

Then, the features aligned in the mapping space are subjected to a similarity measure (KNN algorithm) with the category semantic vector  $z^c$  to complete the classification operation, which is computed as shown in Equations (10)–(12):

$$B_1 = V_r \times z^c \tag{10}$$

$$B_2 = F_a \times z^c \tag{11}$$

$$B_3 = s(f, \ell) \times z^c \tag{12}$$

where the category semantic vector  $z^c$  represents the relationship matrix of categories and semantic features.

The classification result  $B_3$  of the obtained auxiliary part is weighted and fused to the two branches of the main body at the decision level, respectively, so that the model learns more semantic information. The calculation process is shown in Equations (13) and (14):

$$B'_1 = B_1 + \lambda_3 B_3 \tag{13}$$

$$B'_2 = B_2 + \lambda_3 B_3 \tag{14}$$

where  $\lambda_3$  controls the weight of the category label semantic information fusion.

Subsequently, according to the different data environments, the two classification results  $B'_1$  and  $B'_2$  are assigned corresponding weights to obtain the final classification results, as shown in Equation (15).

$$B = \lambda_1 B'_1 + \lambda_2 B'_2 \tag{15}$$

where  $\lambda_1$  and  $\lambda_2$  are the weights of the visual-to-semantic and semantic-to-visual branches, respectively, and  $\lambda_1 = 1 - \lambda_2$ .

In the network structure of the feature space mapping module shown in Figure 3,  $B'_1$  and  $B'_2$ , which incorporate the category label information, are constrained by using the cross-entropy loss on them with the aim of minimizing the error between the predicted and true values. The cross-entropy loss function [4]  $\mathcal{L}_{CE}$  is shown in Equation (16):

$$\mathcal{L}_{CE} = -\frac{1}{n} \sum_{i=1}^n \log \frac{\exp((B'_j)_i \times Y_c)}{\sum_{\hat{c} \in C^s} \exp((B'_j)_i \times Y_{\hat{c}})} \tag{16}$$

where  $C^s$  denotes the seen class,  $Y_c$  denotes the label corresponding to the sample in the form of a onehot vector,  $(B'_j)_i$  denotes the value of the  $i$ th element in the  $B'_1$  or  $B'_2$  vector, which denotes the category score of the  $i$ th image,  $j = 1, 2$ , and  $n$  denotes the number of samples included in a training batch during the training process.

In embedded zero-shot image classification, the training set and the test set are not intersected at all, so when the model is tested, the prediction results will be biased towards the seen class, which leads to the model overfitting to the seen class, in order to alleviate the problem of the model's bias towards the seen class; in this section, the category calibration loss  $\mathcal{L}_{LC}$  is introduced to reduce the model's bias towards the seen class, and the formulas are shown in Equation (17):

$$\mathcal{L}_{LC} = -\frac{1}{n} \sum_{i=1}^n \left( \sum_{c'=1}^{C^u} \log \frac{\exp((B'_j)_i \times Y_{c'} + \psi(c' \in C^u))}{\sum_{c'' \in C^u} \exp((B'_j)_i \times Y_{c''} + \psi(c'' \in C^u))} \right) \tag{17}$$

where  $C^u$  denotes the unseen class and  $n$  denotes the number of samples included in a training batch during the training process.  $\psi(c)$  is an indicator function,  $\psi(c) = 1$  when  $c \in C^u$ , and  $\psi(c) = -1$  otherwise. The indicator function is added to assign a large weight to the unseen class and a small weight to the seen class as a way to suppress the model's bias towards the seen class.

Ultimately, the overall loss function defining the feature space mapping module is shown in Equation (18):

$$\mathcal{L}_{cll} = \mathcal{L}_{CE} + \lambda_{LC}\mathcal{L}_{LC} \quad (18)$$

where  $\lambda_{LC}$  is the weight of the control category calibration loss.

## 4. Results and Analysis

### 4.1. Datasets

The experiments conducted in this paper use three benchmark datasets, CUB (Caltech-UCSD-Birds) [24], SUN (SUN Attribute) [25], and AWA2 (Animals with Attributes2) [26]. The CUB dataset is a sample of 11,788 images with a total of 200 categories of bird species in a fine-grained dataset. The SUN dataset is a fine-grained dataset of scene species with 14,340 image samples and a total of 717 categories. The AWA2 dataset is a coarse-grained dataset of animals with 37,322 image samples and a total of 50 categories. In addition, in order to be fair when the proposed method is compared with other methods, this paper uses the standard dataset division defined by Xian et al. [26], including the division ratio of the training, validation, and test sets, as well as the division ratio of the seen and unseen classes, and the statistics of the dataset information are shown in Table 1.

**Table 1.** Basic information statistics of the datasets.

Datasets	Attributes	Seen Classes	Unseen Classes	Training Samples	Test Samples
Caltech-UCSD-Birds (CUB)	312	150	50	7057	4731
SUN Attribute (SUN)	102	645	72	10,320	4020
Animals with Attributes2 (AWA2)	85	40	10	23,527	13,795

### 4.2. Evaluation Protocols

In the CZSL setting, the unseen class is evaluated using top-1 precision, denoted as *acc*. Under the GZSL setting, the seen class (denoted as *S*) and the unseen class (denoted as *U*) are evaluated separately using top-1 precision. In addition, the overall performance of the GZSL model was measured using the harmonic mean, denoted as *H* ( $H = 2 \times \frac{S \times U}{S + U}$ ).

### 4.3. Implementation Details

We use the ResNet-101 network pre-trained on ImageNet-1K as the visual feature extraction network and do not perform any fine-tuning operations. We optimized the model using RMSProp optimizer with hyperparameters set to momentum = 0.9 and weight decay = 0.0001. Based on the experimental setting and experience, we set the batch size and learning rate to 50 and 0.0001, respectively.

### 4.4. Comparison

We first compare our BFM model with some classical and state-of-the-art methods under the CZSL setting. The experimental results on different datasets under the CZSL setting are provided in Table 2. The BFM model proposed in this paper achieves optimal and sub-optimal accuracies of 71.9% and 62.8% on the CUB and SUN datasets, respectively, which indicates that the model extracts the intrinsic representational information that is effective in discriminating fine-grained image samples. Meanwhile, although the model does not achieve the optimal classification results on the AWA2 dataset, the top-1 accuracy is only 69.3. This is because AWA2 is a coarse-grained dataset, which covers a wider range of knowledge, and our model is not good enough to learn them all in the limited information, so the model does not achieve the optimal results on the AWA2 dataset. However, compared with the best results, the gap is also within 3%, which can show that the model is competitive to the existing models.

**Table 2.** Comparison of image classification models under traditional zero-shot learning setting. Note that in the table, bold fonts indicate the best performance in this metric, and underlined fonts indicate the second best performance in this metric. "--" indicates that the method corresponds to the results not given in the literature.

Methods	Models	CUB	SUN	AWA2
		acc (%)	acc (%)	acc (%)
Generative Methods	f-CLSWGAN	57.3	60.8	68.2
	f-VAEGAN-D2	61.0	<b>64.7</b>	71.1
	Composer	69.4	62.6	<u>71.5</u>
	cycle-CLSWGAN	58.4	60.0	66.3
	LisGAN	58.8	61.7	--
Embedding-based Methods	TCN	59.5	61.5	71.2
	DAZLE	66.0	59.4	67.9
	LFGAA	67.6	61.5	68.1
	SGMA	<u>71.0</u>	--	68.8
	DSAN	57.4	62.4	<b>72.3</b>
	BFM (ours)	<b>71.9</b>	<u>62.8</u>	69.3

The experimental results on different datasets under the GZSL setting are provided in Table 3. The BFM model proposed in this paper has the best classification results on the CUB dataset, with a reconciled mean of 61.6%. The results on the AWA2 dataset are also closer to those of the SOTA method, reaching 66.6%. However, BFM is less effective on the SUN dataset because the SUN dataset has many categories and fewer samples per category, and because it is a scenario class dataset, it is more difficult for the embedded-based model to extract the effective information. The results in the table also show that generative based models are more advantageous on the SUN dataset. In addition, for the evaluation index, i.e., the classification accuracy of seen class samples, it can be seen that many models have achieved very good results on this index, but many models are less effective on the index, i.e., the model's classification accuracy of unseen class samples is worse, which indicates that some models do not compensate much for the seen class favoritism problem, resulting in the model's overfitting to the seen class. While the BFM model proposed in this paper is not as effective as other models in some indicators, the BFM model is more effective in addressing the problem of seen class favoritism, which effectively alleviates the seen class favoritism problem.

**Table 3.** Comparison of image classification models in a generalized zero-shot learning setting. Note that in the table, bold fonts indicate the best performance in this metric, and underlined fonts indicate the second best performance in this metric. "--" indicates that the method corresponds to the results not given in the literature.

Methods	Models	CUB			SUN			AWA2		
		S (%)	U (%)	H (%)	S (%)	U (%)	H (%)	S (%)	U (%)	H (%)
Generative Methods	f-CLSWGAN	57.7	43.7	49.7	36.6	42.6	39.4	61.4	57.9	59.6
	f-VAEGAN-D2	60.1	48.4	53.6	38.0	45.1	<b>41.3</b>	70.6	57.6	63.5
	Composer	63.8	56.4	<u>59.9</u>	22.0	<b>55.1</b>	31.4	77.3	<b>62.1</b>	<b>68.8</b>
	cycle-CLSWGAN	61.0	45.7	52.3	33.6	49.4	40.0	64.0	56.9	60.2
	LisGAN	57.9	46.5	51.6	37.8	42.9	<u>40.2</u>	--	--	--
Embedding-based Methods	TCN	52.0	52.6	52.3	37.3	31.2	34.0	65.8	61.2	63.4
	DAZLE	59.6	<u>56.7</u>	58.1	24.3	<u>52.3</u>	33.2	75.7	60.3	67.1
	LFGAA	<b>80.9</b>	36.2	50.0	<b>40.0</b>	18.5	25.3	<b>93.4</b>	27.0	41.9
	SGMA	<u>71.3</u>	36.7	48.5	--	--	--	<u>87.1</u>	37.6	52.5
	DSAN	56.6	46.9	51.3	41.1	33.2	36.7	78.8	58.6	<u>67.2</u>
	BFM (ours)	61.3	<b>61.9</b>	<b>61.6</b>	25.3	48.4	33.2	72.8	<u>61.3</u>	66.6

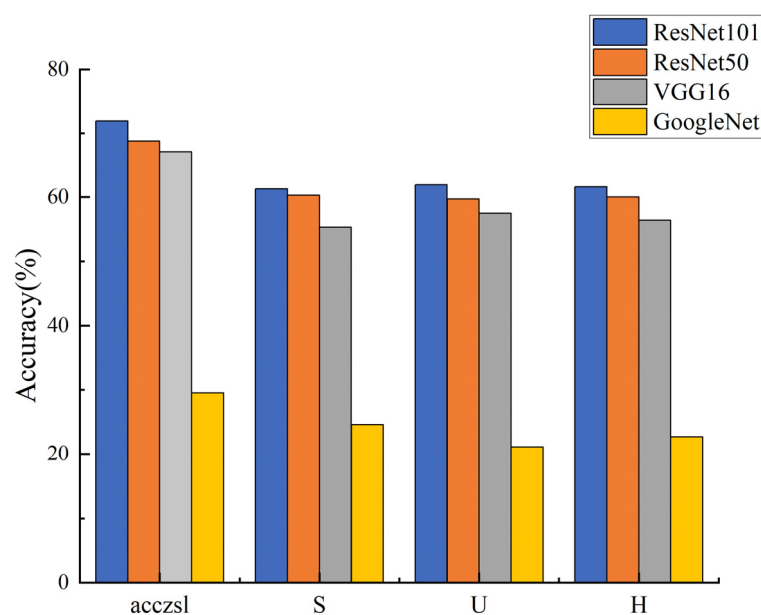
Compared with the two generative models that perform better, our model has lower complexity. The number of training parameters of the BFM model proposed in this paper is about  $4.56 \times 10^6$ . The number of training parameters of model f-VAEGAN-D2 is about  $1.48 \times 10^7$ , and the number of training parameters of model com is about  $5.81 \times 10^6$ . The number of training parameters of the BFM model is 69.2% and 21.5% lower than that of the other models, respectively. Under the same training environment, the training time of our BFM model is about half that of model com.

#### 4.5. Ablation Studies

To further understand our BFM model, we performed ablation experiments on it to evaluate the effectiveness of the model's modules.

##### 4.5.1. Visual Feature Extraction Network

Figure 8 shows the experimental comparison of several classical models among the three commonly used deep feature extraction networks, through which the results reveal that the ResNet-101 [27] network is the most suitable as a visual feature extraction network for the zero-shot image classification task. This is because the VGG [28] model is deeper and uses more parameters for the fully connected layer, which is computationally intensive; it is prone to the problem of gradient vanishing, which leads to difficulties in training, whereas ResNet101 employs a global average pooling layer at the end, which converts the feature maps into fixed-length vectors, which helps to extract more representative features and reduces the number of model parameters. Although GoogleNet [29] uses the Inception module, and it effectively reduces the number of parameters, it may face problems such as gradient vanishing while training deep networks.



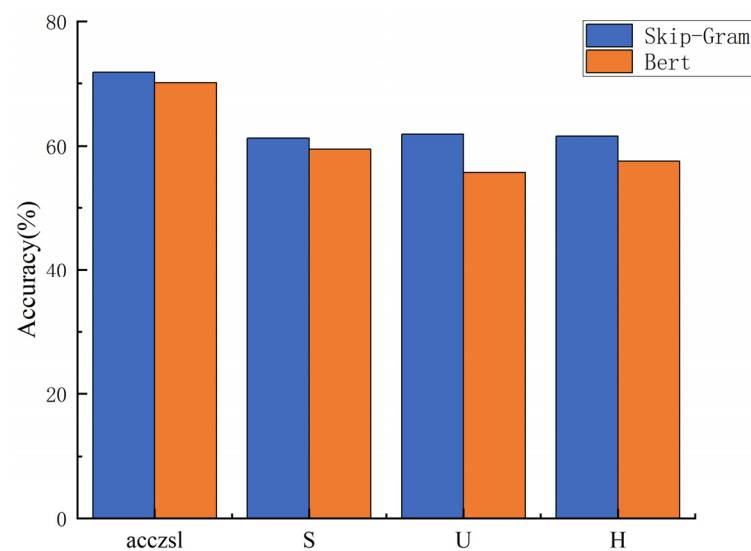
**Figure 8.** Results of ablation experiments with different visual extraction networks.

In contrast, ResNet101 uses a residual connection structure that overcomes the problem of vanishing gradients and makes the network easier to train. This design helps to propagate the gradient better, which improves the model performance and convergence speed. And, compared to ResNet50, the ResNet101 network has a deeper network structure with a residual connectivity module, which can better capture the high-level semantic information of the image in the zero-shot learning task, which helps to improve the classification performance. And the effective addition of the residual connection module helps to solve the problems of gradient vanishing and gradient explosion and facilitates the flow of

information. This design makes the model easier to train and also improves the performance of the model.

#### 4.5.2. Semantic Feature Extraction Network

Figure 9 shows the experimental comparison of the two semantic feature extraction models; through the comparison results in the figure, it is found that the skip-gram [30] language model is the most suitable for the semantic feature extraction work in the zero-shot image classification task. This is because, in the zero-shot image classification task, most of the words or phrases are short words or phrases, and there is no long text, and the skip-gram language model performs well in dealing with short text and sparse data; and the Bert [31] model is larger, and the dimensionality of the semantic features obtained is also relatively high, which may reduce the training efficiency of the model. As a result, in this paper, the skip-gram language model is chosen as the semantic feature extraction network for the zero-shot image classification task.



**Figure 9.** Results of ablation experiments with different language models.

#### 4.5.3. Model Component Ablation Experiments

In this section, model component ablation experiments are conducted as a way to analyze the effectiveness of different modules in the BFM model for traditional and generalized zero-shot image classification. The experiments explore the effect of removing the attribute-to-visual branch A-V, visual-to-attribute branch V-A, category tag word vector  $W2v_L$ , visual self-attention feature  $V_{att}$ , and category calibration loss  $\mathcal{L}_{LC}$  on the BFM model on the CUB dataset. Note that here the attribute-to-visual branch A-V represents the complete branch plus the visually guided attention-based mechanism; similarly, the visual-to-attribute branch V-A represents the complete branch plus the attribute-guided attention-based mechanism.

In this section, the components are added sequentially to the BFM, and the changes in the results after the addition reflect the validity of the components in the model. Table 4 shows the results of the BFM model component ablation experiments. “✗” means that this module is not present in the model. “✓” means that this module is present in the model.

Among them, the first row is the baseline model; based on the baseline model, the A-V branch with the attention mechanism, the V-A branch, the category-labeled word vector  $W2v_L$ , the self-attention mechanism of visual features  $V_{att}$ , and the  $\mathcal{L}_{LC}$  loss function are added step by step, and, finally, the experimental results of the method model proposed in this paper obtains a substantial improvement compared with the baseline model. For example, in the conventional zero-shot image classification task, the top-1 accuracy ( $acc_{zsl}$ ) of the unseen class is improved by 45.4% compared to the baseline model; in the generalized



zero-shot image classification task, the top-1 accuracy (S) of the seen class is improved by 23.7%, that of the unseen class is improved by 59.6%, and that of the harmonic mean (H) is improved by 57.2%. The third row of the table gains a 32.4%/43% improvement in acc/H on the basis of the second row, which is due to the fact that, at this point, the model possesses the property of mining feature information in both directions, which enriches the feature representations and promotes the knowledge migration from the seen to the unseen class. Finally, the problem of bias towards seen classes during testing is mitigated by adding a class calibration loss function, which is effective as can be seen from the results in the table, where the difference between the seen class top-1 accuracy (S) and the unseen class top-1 accuracy (U) is only 0.6%.

Table 4. Results of BFM model component ablation experiments.

A-V	V-A	W2v <sub>L</sub>	V <sub>att</sub>	L <sub>LC</sub>	ZSL		GZSL	
					acc (%)	S (%)	U (%)	H (%)
✗	✗	✗	✗	✗	26.4	37.6	2.3	4.4
✓	✗	✗	✗	✗	35.7	32.8	4.3	7.7
✓	✓	✗	✗	✗	68.1	49.6	51.8	50.7
✓	✓	✓	✗	✗	69.6	69.6	18.7	29.5
✓	✓	✓	✓	✗	71.8	70.2	18.3	29.1
✓	✓	✓	✓	✓	71.9	61.3	61.9	61.6

4.6. Hyperparametric Analysis

4.6.1. Category Calibration Loss Weight Analysis

The weight settings of the category-calibrated loss function are analyzed here, and the model performance with different weight coefficients is shown in Figure 10. Based on the results in Figure 10, we set λ<sub>LC</sub> for the CUB/AWA2 dataset to 0.1. This is because the seen and unseen classes in the CUB and AWA2 datasets are animals, and there is more semantic knowledge shared between the classes, i.e., the seen classes are more similar to the unseen classes, which leads to a more serious bias of the model towards the seen classes, so the loss weights need to be set larger. While the SUN dataset is a scene class dataset, the image subject is more complex, and since the number of its seen classes is much larger than the number of unseen classes, the model usually overfits the unseen classes on this dataset. Therefore, we set the λ<sub>LC</sub> of SUN to 0.001.

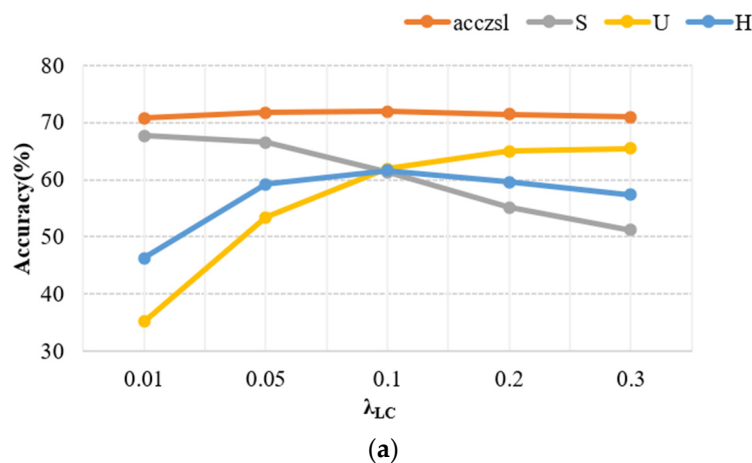
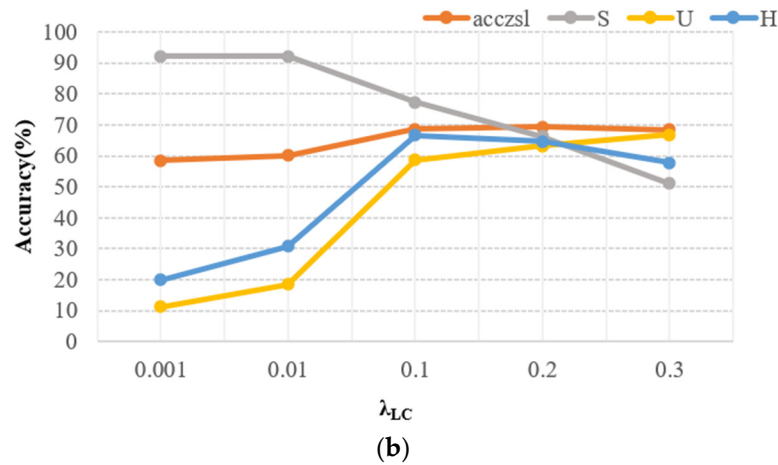


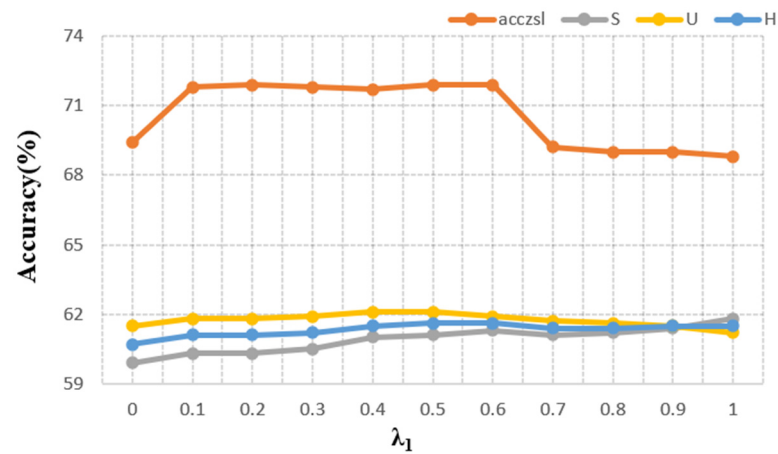
Figure 10. Cont.



**Figure 10.** Analysis of weighting factors for category calibration losses. (a) CUB dataset. (b) AWA2 dataset.

#### 4.6.2. Combined Coefficient Analysis

Parameters  $\lambda_1$  and  $\lambda_2$  are the weighting coefficients of the visual-to-attribute branch V-A and the attribute-to-visual branch A-V, respectively. Because  $\lambda_1 + \lambda_2 = 1$ , only  $\lambda_1$  is used to represent the horizontal axis in the figure, i.e.,  $\lambda_1$  denotes the proportion of branch V-A, and  $\lambda_2 = 1 - \lambda_1$  denotes the proportion of branch A-V. From Figure 11, it can be seen that the model is not sensitive to the changes in the weighting coefficients of the two branches. This is because the BFM model adopts a bidirectional parallel feature mapping method, so that the knowledge learned from the two branches is complementary, and the model learns more comprehensive and rich knowledge. The BFM achieves the best results when the parameter  $\lambda_1$  is set to 0.6 and 0.3 on CUB and SUN, respectively.

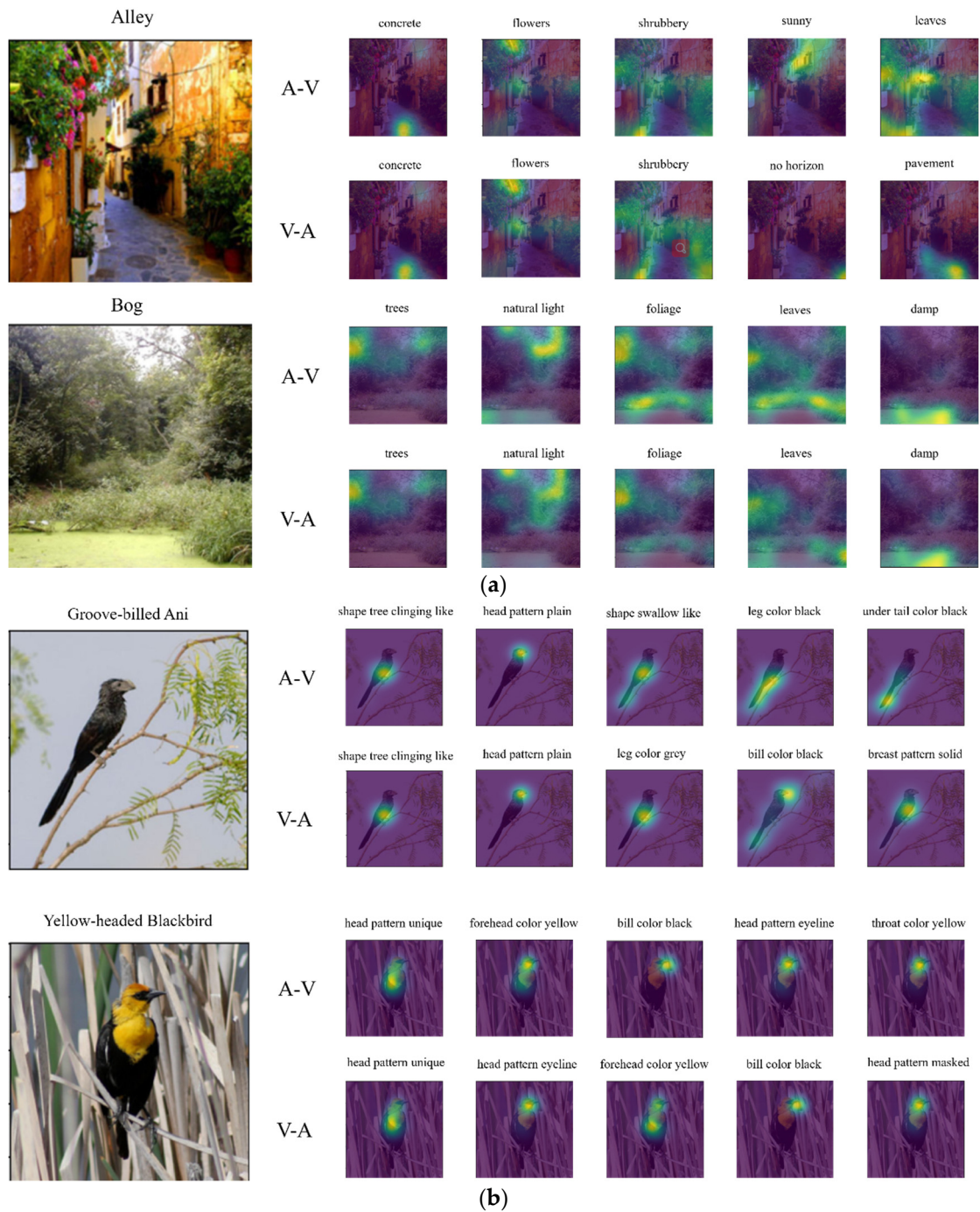


**Figure 11.** Analysis of weighting coefficients of V-A branches and A-V branches.

#### 4.7. Attention Map Visualization Analysis

To visualize the effectiveness of our BFM in extracting the intrinsic connection between visual features and semantic features, we visualize the attention graphs learned using the bidirectional mapping networks, e.g., BFM (A-V) and BFM (V-A). As shown in Figure 12, the A-V and V-A networks effectively learn visually guided and attribute-guided discriminative attribute-based features and visual features, respectively. In addition, the two networks learn the most important features separately and learn complementary feature knowledge from each other, which is conducive to the model obtaining richer semantic knowledge from the mutual learning of the two networks and better realizing the knowledge transfer from the seen class to the unseen class. For example, in the sample alley, A can learn the

feature “leaves” but not the feature “no horizon”; V can learn the feature “no horizon” but not the feature “leaves”. Therefore, our complete BFM can learn richer semantic knowledge, and the model achieves good performance in both seen and unseen classes.



**Figure 12.** Visualization results of visually guided attention-based mechanisms in the semantic-to-visual branch and attribute-guided attention-based mechanisms in the visual-to-semantic branch on the (a) SUN dataset and (b) CUB dataset.

### 5. Conclusions

In this paper, we propose an embedded zero-shot image classification model based on bidirectional feature mapping. The main part of the model is the feature mapping module, which is dominated by a bidirectional feature mapping network, which mainly

learns the intrinsic information of visual and attribute features and the intrinsic connection between them. It is supplemented with a visual–semantic mapping network to provide richer semantic knowledge for the model. To enhance the expressiveness of visual features, we introduce a self-attention mechanism to dynamically focus on the features and extract key visual features. To alleviate the seen class bias problem, we introduce the category calibration loss to balance the weights of seen and unseen classes. As a result, the model is able to capture richer and accurate intrinsic semantic representations for effective knowledge transfer. Experiments on three commonly used datasets demonstrate the superiority of BFM, and we hope that our work will contribute to the field of zero-shot image classification.

**Author Contributions:** Conceptualization, Z.Z. and H.S.; methodology, Z.Z. and H.S.; software, Z.Z.; validation, P.Z. and Y.L.; formal analysis, H.S. and Y.L.; investigation, X.Z.; resources, X.H.; data curation, Z.Z.; writing—original draft preparation, Z.Z. and H.S.; writing—review and editing, Z.Z. and Y.L.; visualization, P.Z.; supervision, H.S.; project administration, X.H.; funding acquisition, X.Z. and H.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by Harbin City Science and Technology Plan Projects grant number 2022ZCZJCG006, the Basic Research Support Program for Excellent Young Teachers in Provincial Undergraduate Universities in Heilongjiang Province grant number YQJH2023240, and the Science and Technology Collaborative Innovation Project in Heilongjiang Province grant number LJGXCG2022-085.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data presented in this study are available on request from the corresponding author. The datasets used in this paper are all public datasets. The dataset can be found at: CUB: [http://www.vision.caltech.edu/datasets/cub\\_200\\_2011/](http://www.vision.caltech.edu/datasets/cub_200_2011/) (accessed on 7 May 2024); AWA2: <https://cvml.ista.ac.at/AWA2/> (accessed on 7 May 2024); SUN: <https://cs.brown.edu/~gmpatter/sunattributes.html> (accessed on 7 May 2024).

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

## References

1. Wang, P.; Fan, E.; Wang, P. Comparative analysis of image classification algorithms based on traditional machine learning and deep learning. *Pattern Recognit. Lett.* **2021**, *141*, 61–67. [[CrossRef](#)]
2. Wang, X.; Peng, D.; Hu, P.; Gong, Y.; Chen, Y. Cross-domain alignment for zero-shot sketch-based image retrieval. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 7024–7035. [[CrossRef](#)]
3. Liu, H.; Qin, Z. Deep quantization network with visual-semantic alignment for zero-shot image retrieval. *Electron. Res. Arch.* **2023**, *31*, 4232–4247. [[CrossRef](#)]
4. Hong, M.; Zhang, X.; Li, G.; Huang, Q. Fine-grained feature generation for generalized zero-shot video classification. *IEEE Trans. Image Process.* **2023**, *32*, 1599–1612. [[CrossRef](#)] [[PubMed](#)]
5. Tursun, O.; Denman, S.; Sridharan, S.; Goan, E.; Fookes, C. An efficient framework for zero-shot sketch-based image retrieval. *Pattern Recognit.* **2022**, *126*, 108528. [[CrossRef](#)]
6. Liu, X.; Bai, S.; An, S.; Wang, S.; Liu, W.; Zhao, X.; Ma, Y. A meaningful learning method for zero-shot semantic segmentation. *Sci. China Inf. Sci.* **2023**, *66*, 210103. [[CrossRef](#)]
7. Wang, Y.; Tian, Y. Exploiting multi-scale contextual prompt learning for zero-shot semantic segmentation. *Displays* **2024**, *81*, 102616. [[CrossRef](#)]
8. Chen, Q.; Wang, W.; Huang, K.; Coenen, F. Zero-shot text classification via knowledge graph embedding for social media data. *IEEE Internet Things J.* **2022**, *9*, 9205–9213. [[CrossRef](#)]
9. Liu, C.; Wang, C.; Peng, Y.; Li, Z. ZVQAF: Zero-shot visual question answering with feedback from large language models. *Neurocomputing* **2024**, *580*, 127505. [[CrossRef](#)]
10. Qiao, R.; Liu, L.; Shen, C.; Hengel, A.V.D. Visually aligned word embeddings for improving zero-shot learning. *arXiv* **2017**, arXiv:1707.05427.
11. Reed, S.; Akata, Z.; Lee, H.; Schiele, B. Learning deep representations of fine-grained visual descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 49–58.
12. Yu, B.; Xie, C.; Tang, P.; Li, B. Semantic-visual shared knowledge graph for zero-shot learning. *PeerJ Comput. Sci.* **2023**, *9*, e1260. [[CrossRef](#)]



13. Frome, A.; Corrado, G.S.; Shlens, J.; Bengio, S.; Dean, J.; Ranzato, M.A.; Mikolov, T. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems 26, Proceedings of the 27th Annual Conference on Neural Information, Lake Tahoe, NV, USA, 5–10 December 2013*; Curran Associates, Inc.: Red Hook, NY, USA, 2014.
14. Yu, Y.; Ji, Z.; Li, X.; Guo, J.; Zhang, Z.; Ling, H.; Wu, F. Transductive zero-shot learning with a self-training dictionary approach. *IEEE Trans. Cybern.* **2018**, *48*, 2908–2919. [[CrossRef](#)] [[PubMed](#)]
15. Li, L.; Liu, L.; Du, X.; Wang, X.; Zhang, Z.; Zhang, J.; Zhang, P.; Liu, J. CGUN-2A: Deep graph convolutional network via contrastive learning for large-scale zero-shot image classification. *Sensors* **2022**, *22*, 9980. [[CrossRef](#)] [[PubMed](#)]
16. Kong, D.; Li, X.; Wang, S.; Li, J.; Yin, B. Learning visual-and-semantic knowledge embedding for zero-shot image classification. *Appl. Intell.* **2023**, *53*, 2250–2264. [[CrossRef](#)]
17. Wang, Y.; Feng, L.; Song, X.; Xu, D.; Zhai, Y. Zero-shot image classification method based on attention mechanism and semantic information fusion. *Sensors* **2023**, *23*, 2311. [[CrossRef](#)] [[PubMed](#)]
18. Sun, X.; Tian, Y.; Li, H. Zero-shot image classification via visual–semantic feature decoupling. *Multimed. Syst.* **2024**, *30*, 82. [[CrossRef](#)]
19. Xie, G.S.; Liu, L.; Jin, X.; Zhu, F.; Zhang, Z.; Qin, J.; Yao, Y.; Shao, L. Attentive region embedding network for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019*; pp. 9384–9393.
20. Huynh, D.; Elhamifar, E. Fine-grained generalized zero-shot learning via dense attribute-based attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020*; pp. 4483–4493.
21. Naeem, M.F.; Xian, Y.; Gool, L.V.; Tombari, F. I2dformer: Learning image to document attention for zero-shot image classification. In *Advances in Neural Information Processing Systems 35, Proceedings of the 36th Annual Conference on Neural Information Processing Systems, New Orleans, LA, USA, 28 November–9 December 2022*; Curran Associates, Inc.: Red Hook, NY, USA, 2022; pp. 12283–12294.
22. Akata, Z.; Reed, S.; Walter, D.; Lee, H.; Schiele, B. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015*; pp. 2927–2936.
23. Alamri, F.; Dutta, A. Multi-head self-attention via vision transformer for zero-shot learning. *arXiv* **2021**, arXiv:2108.00045.
24. Wah, C.; Branson, S.; Welinder, P.; Peron, P.; Belongi, S. *The Caltech-Ucsd Birds-200-2011 Dataset*; California Institute of Technology: Pasadena, CA, USA, 2011.
25. Patterson, G.; Xu, C.; Su, H.; Hays, J. The sun attribute database: Beyond categories for deeper scene understanding. *Int. J. Comput. Vis.* **2014**, *108*, 59–81. [[CrossRef](#)]
26. Xian, Y.; Lampert, C.H.; Schiele, B.; Akata, Z. Zero-shot learning—A comprehensive evaluation of the good, the bad and the ugly. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 2251–2265. [[CrossRef](#)]
27. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017*; pp. 1492–1500.
28. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
29. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015*; pp. 1–9.
30. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26, Proceedings of the 27th Annual Conference on Neural Information, Lake Tahoe, NV, USA, 5–10 December 2013*; Curran Associates, Inc.: Red Hook, NY, USA, 2014.
31. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.

**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.