electronics

MDPI

Article

Dual-Contrastive Attribute Embedding for Generalized Zero-Shot Learning

Qin Li [1], Yujie Long [2], Zhiyi Zhang [1] and Kai Jiang [1,*]

[1] School of Computer Science and Software Engineering, Shenzhen University of Information Technology, Shenzhen 518172, China; liqin@sziit.edu.cn (Q.L.); zhangzhiyi@sziit.edu.cn (Z.Z.)
[2] School of Computer Science, Guangdong University of Foreign Studies South China Business College, Guangzhou 510545, China; 2520601158@e.gwng.edu.cn
* Correspondence: jiangkai@sziit.edu.cn

Abstract

Zero-shot learning (ZSL) aims to categorize target classes with the aid of semantic knowledge and samples from previously seen classes. In this process, the alignment of visual and attribute modality features is key to successful knowledge transfer. Several previous studies have investigated the extraction of attribute-related local features to reduce visual-semantic domain gaps and overcome issues with domain shifts. However, these techniques do not emphasize the commonality of features across different objects belonging to the same attribute, which is critical for identifying and distinguishing the attributes of unseen classes. In this study, we propose a novel ZSL method, termed dual-contrastive attribute embedding (DCAE), for generalized zero-shot learning. This approach simultaneously learns both class-level and attribute-level prototypes and representations. Specifically, an attribute embedding module is introduced to capture attribute-level features and an attribute semantic encoder is developed to generate attribute prototypes. Attribute-level and class-level contrastive loss terms are then used to optimize an attribute embedding space such that attribute features are compactly distributed around corresponding prototypes. This double contrastive learning mechanism facilitates the alignment of multimodal information from two dimensions. Extensive experiments with three benchmark datasets demonstrated the superiority of the proposed method compared to current state-of-the-art techniques.

Keywords: zero-shot learning; attribute-level features; contrastive learning

Academic Editor: Daniel Gutiérrez Reina

Received: 8 September 2025
Revised: 30 October 2025
Accepted: 30 October 2025
Published: 5 November 2025

Citation: Li, Q.; Long, Y.; Zhang, Z.; Jiang, K. Dual-Contrastive Attribute Embedding for Generalized Zero-Shot Learning. *Electronics* **2025**, *14*, 4341. https://doi.org/10.3390/electronics14214341

1. Introduction

The development of deep neural networks has facilitated the pursuit of complex object classification tasks, which typically require large quantities of labeled data [1,2]. Zero-shot learning (ZSL), modeled after the ability of humans to predict labels for unseen targets (given prior knowledge), aims to transfer knowledge learned from seen classes to target classes. This process utilizes semantic auxiliary information [3–6], which typically includes human-defined attributes [7,8], word vectors [9], and their combinations. Zero-shot learning methods can be divided into conventional ZSL (CZSL) and generalized ZSL (GZSL), depending on the target classes used in the test phase [3]. The primary goal of CZSL is to recognize unseen classes, while GZSL aims to recognize both seen and unseen classes.

ZSL has achieved significant progress in recent years, with several studies focusing on embedding and generative methods. Embedding-based models typically learn mapping functions, which are then used to link visual features and semantic information. Generative-based models reduce large biases toward seen classes by generating visual features for

unseen classes. Most embedding methods employ convolutional neural networks or transformers to extract global visual features from images [10]. However, such features often contain significant amounts of attribute-independent background information, which enlarges the gap between visual features and semantic details. As such, attention-based ZSL methods employ attribute descriptions to guide the learning of discriminative local visual features used to address the challenges discussed above [11–15]. The performance of ZSL methods can also be improved significantly by aligning attribute features and class semantics. However, previous studies have typically extracted attribute representations only for certain pictures, without taking into account the commonality between the attribute features of different object classes, which can lead to relatively discrete distributions of representations for the same attributes in an embedding space. Furthermore, this approach has the potential to prevent models from recognizing attributes in unseen classes, which is the cause of the domain shift problem [16]. In order to address these issues, the same attribute features for different objects must be more compactly distributed in the embedding space, with features for different attributes further apart. This approach can effectively improve attribute identification, as shown in Figure 1.

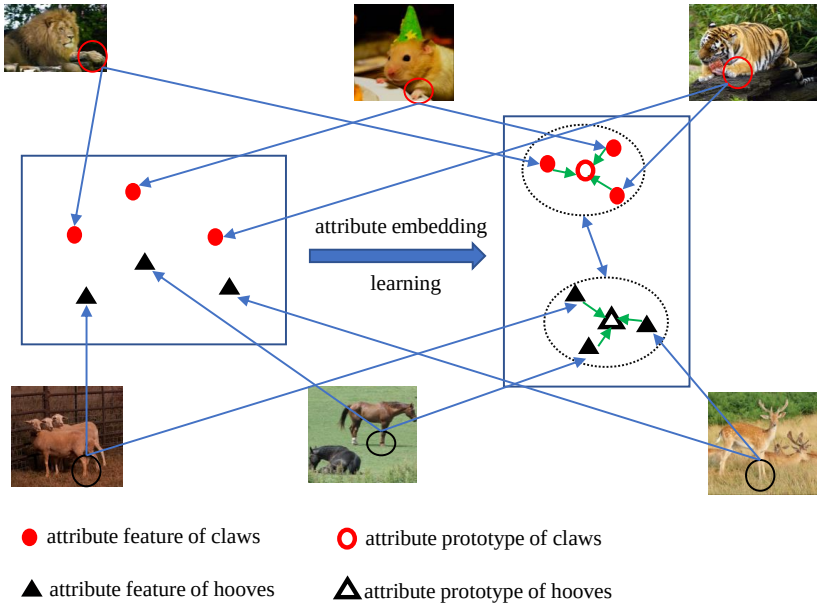

Figure 1. An illustration of attribute features in the embedding space. In existing methods, the commonality of similar attribute features cannot be learned from various objects with obvious visual differences, which leads to the domain shift problem. In the present study, this issue is addressed by explicitly learning attribute-level features and attribute prototypes used to improve the discriminative capacity of the embedding space.

In light of these observations, we propose a novel GZSL method, termed dual-contrastive attribute embedding (DCAE) for generalized zero-shot learning, which is designed to overcome the domain shift problem. In this process, an attribute semantics encoder is first developed to map one-hot attribute vectors into a set of prototypes. Feature maps can then be obtained simultaneously using an attribute filtering network. An included embedding module then outputs attribute-level visual features and triplet loss is used to align the features with corresponding attribute prototypes, allowing features from the same attribute to be distributed around corresponding prototypes. In addition, visual feature distributions can be made more compact, with distributions for distinct attributes further apart, by using a contrastive loss strategy based on adaptive hard sample selection. This approach constrains the attribute embedding space, which allows the model to more effectively identify attributes. In addition, a class-level contrastive loss term, constructed

by concatenating attribute features, is used to optimize the attribute vectors for each image in a batch. In addition to the learning of attribute embedding spaces, class-level features were also extracted from images and optimized using corresponding class prototypes. In summary, the primary contributions of the proposed DCAE method can be described as follows:

- A novel framework is introduced to extract attribute-level features and generate prototypes used to learn highly discriminative embedding spaces. In this way, the model can better recognize attributes and overcome the domain shift problem.
- An attribute-level contrastive loss term is proposed based on adaptive hard sample selection and used to improve the discriminative capacity of attribute representations. In addition, class-level contrastive loss is employed to optimize these representations and enhance GZSL performance.
- A double contrastive learning mechanism is developed to facilitate the alignment of multimodal information from two dimensions. Experiments with three challenging benchmark datasets (CUB [17], AWA2 [18], and SUN [19]) demonstrate that our method outperforms many state-of-the-art GZSL models.

The remainder of this paper is organized as follows. Section 2 introduces related works investigating zero-shot learning and contrastive learning. Section 3 introduces the proposed method, including basic notations, network architectures, and loss functions. The performance of DCAE is further evaluated using several benchmark datasets in Section 4 and conclusions are provided in Section 5.

2. Related Work

This section introduces related work on ZSL methods and contrastive learning frameworks.

2.1. Zero-Shot Learning

Existing ZSL methods can be divided into generative- and embedding-based models. Generative methods typically utilize generative adversarial networks (GANs) [5,20] or variational autoencoders (VAEs) [21,22] to produce samples for unseen classes from existing data and construct a complete dataset. These techniques involve training a classifier on both generated and seen samples, thereby converting a GZSL problem into a traditional supervised classification problem. The benefit of generative models is their ability to effectively reduce classification bias for seen classes. However, the visual features in unseen classes generated by generative models are similar to those of seen classes, which limits further performance improvements in GZSL. In contrast, embedding methods [23–26] can more directly transfer semantic knowledge from seen classes to unseen classes by associating visual features with semantic information. In other words, these techniques map low-level visual features to a semantic embedding space or map semantic information to a visual embedding space, by learning an embedding function. The learned function is then used to identify unseen classes by measuring the similarity between predicted and prototype representations of sample points in the embedding space. SCILM [27] involves a novel strategy that balances model training by randomly selecting an equal number of samples from each training category. In addition, attention mechanisms have recently been employed for extracting rich attribute features to narrow semantic gaps. Similarly, MSDN [28] gradually learns intrinsic attribute representations through attention and knowledge distillation mechanisms. DPPN [29] constructs two prototypes to record prototypical visual patterns for classes and attributes, which is beneficial for improving the transferability of attribute knowledge. Unlike these studies, we focus on mining commonalities between the same attributes exhibited by objects with obvious visual differences. Additionally, some

cross-disciplinary works in optimization and decision-making have been studied in other domains [30,31].

2.2. Contrastive Learning

As a characteristic representation learning method, contrastive learning has achieved considerable performance in both unsupervised [32,33] and supervised tasks [34,35]. The primary aim of contrastive learning is to maximize the similarity of positive pairs while minimizing the similarity of negative pairs. As such, the construction of positive and negative samples is critical for learning robust and discriminative representations. SimCLR [32] considers the same image from two different views as a positive pair, while the remaining images are considered to be negative pairs. MoCo [36] treats contrastive learning as a dictionary look-up task and constructs a large and consistent dictionary for performing contrastive unsupervised learning. SCL [37] introduces a self-supervised mechanism based on contrastive learning for few-shot learning, aiming to minimize the distance between each training sample and its augmented variants, thereby enhancing sample discrimination. In supervised contrastive learning [34], images from the same class constitute positive pairs and images from different classes constitute negative pairs. CE-GZSL [38] introduces both instance-level and class-level contrastive learning into the zero-shot learning framework, in an attempt to learn highly discriminative embedding spaces. However, positive and negative pairs in these techniques are not representative. In contrast, the attribute-level comparative learning step in this study includes positive and negative sample selection, aiming to improve the discriminative capacity of the attribute embedding space.

3. Methodology

In this section, the notation and problem definitions for zero-shot learning are first developed sequentially. An illustration of the proposed framework and corresponding formulation are then provided in detail. More specifically, our model learns highly discriminative visual features through the joint optimization of class representation learning and attribute embedding learning.

3.1. Notation and Problem Settings

Zero-shot learning (ZSL) attempts to recognize unseen classes (Y^U) by transferring knowledge learned from seen classes (Y^S) with the assistance of class semantics. In this case, seen and unseen classes are disjoint (i.e., $Y^S \cap Y^U = \emptyset$) and the corresponding image spaces can be represented as $X = X^S \cup X^U$. The training set then contains attribute vectors and labeled images from seen classes (i.e., $T^s = \{x_i^s, y_i^s, \varphi(y_i^s)\}_{i=1}^{N^s}$). Here, x_i^s is an image in X^S, y_i^s is a corresponding class label, and $\varphi(y_i^s) \in R^K$ is a class semantic vector. The test set for unseen classes can thus be defined as $T^u = \{x_i^u, y_i^u, \varphi(y_i^u)\}_{i=1}^{N^u}$. Class semantic information is defined as $CS = \{\varphi(i)\}_{i=1}^{M}$, which is provided to facilitate knowledge transfer from seen to unseen classes. Here, M is the total number of categories in both sets of classes. The goal of CZSL is then to predict the labels for objects in unseen classes (i.e., $X^S \rightarrow Y^U$). However, the goal of GZSL is to predict objects from both classes (i.e., $X^S \rightarrow Y^U \cup Y^S$).

3.2. Overview

Figure 2 demonstrates the proposed framework, which includes five components: an image encoder network, a class semantics encoder, an attribute filter network, an attribute embedding module, and an attribute semantics encoder.

Prototype Generation Network. The proposed prototype generation module consists of two encoders: a **Class Semantics Encoder** and an **Attribute Semantics Encoder**, as illustrated in Figure 3. The class semantics encoder maps class semantics $\mathcal{CS} = \{\phi(i)\}_{i=1}^{M}$ to class prototypes $\mathcal{CP} = \{cp_i\}_{i=1}^{M}$, while the attribute semantics encoder maps attribute

one-hot vectors $\mathcal{AS} = \{as_j\}_{j=1}^{K}$ to attribute prototypes $\mathcal{AP} = \{ap_j\}_{j=1}^{K}$. By embedding semantic information into the visual representation space, increasingly separable and discriminative prototype features can be obtained.

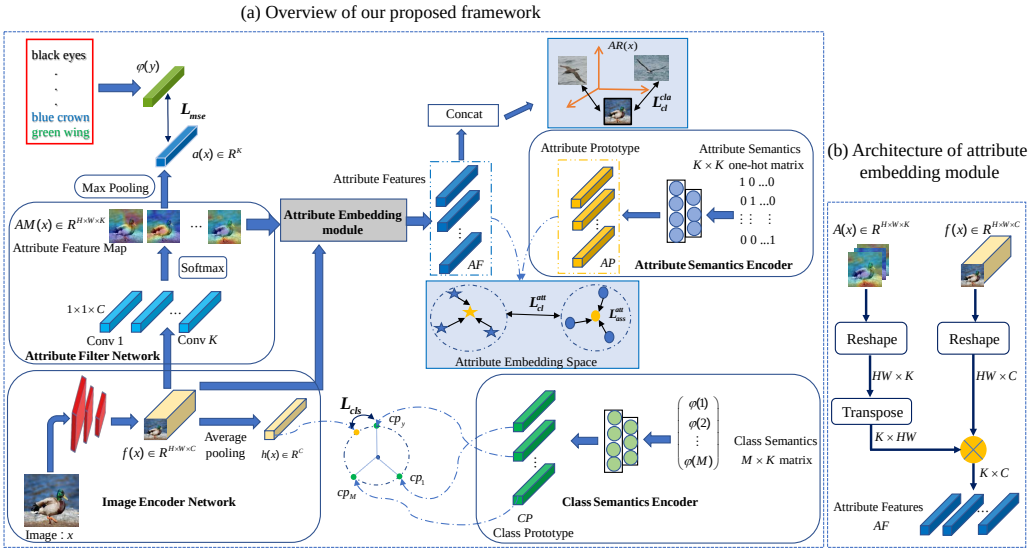

Figure 2. (**a**) An illustration of the proposed framework, which includes five components: an image encoder network, a class semantics encoder, an attribute filter network, an attribute embedding module, and an attribute semantics encoder. The two encoders on the right map attribute and category semantics into attribute and category prototypes, respectively. The image encoder and attribute embedding modules then extract class representations and attribute features from the images, respectively. (**b**) The attribute embedding module takes visual representation and attribute feature maps as input and then outputs attribute-level visual features after operations such as transposing, reshaping, and matrix multiplication.

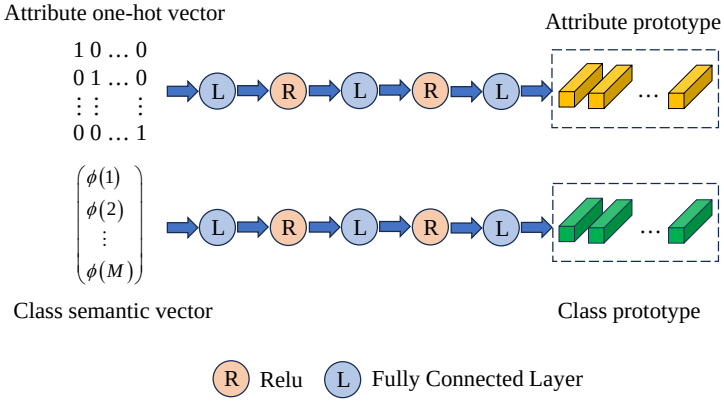

Figure 3. Prototype Generation Network.

Class Semantics Encoder: Directly aligning class features with semantic vectors in the semantic embedding space may lead to the hubness problem, as different classes often share overlapping attributes. Consequently, feature embeddings of different classes may cluster around similar semantic vectors, reducing discriminability. To mitigate this issue, the DCAE model projects class semantic vectors into a *high-dimensional visual embedding space* to obtain class prototypes $\mathcal{CP} = \{cp_i\}_{i=1}^{M}$, which are more suitable for classification tasks. The class semantics encoder is implemented using three fully connected layers with two ReLU activation functions. The input dimensionality corresponds to the number of attributes, while the output dimensionality is set to 2048, matching the feature dimension of the ResNet-101 backbone.

Attribute Semantics Encoder: To enhance the transferability of class features and alleviate the domain shift problem, we further design an attribute semantics encoder to discriminate and localize visual attributes more effectively. Similar to the class semantics encoder, it projects one-hot attribute vectors into an attribute embedding space to obtain attribute prototypes $\mathcal{AP} = \{ap_j\}_{j=1}^{K}$. This network adopts the same structure, consisting of three fully connected layers and two ReLU activations.

Initialization and Interpretability: All prototype parameters are initialized using *Xavier initialization* to ensure stable convergence and balanced feature scaling. Since both class and attribute prototypes are generated through the projection of semantic descriptions, they preserve clear semantic correspondence—each dimension can be interpreted in relation to specific attributes or categories. This design enhances the semantic interpretability of the learned prototypes while maintaining strong discriminative capability in both class and attribute embedding spaces.

Image Encoder Network. An image encoder network was leveraged to extract visual representations $f(x) \in R^{H \times W \times C}$ of the image x. Global average pooling was then applied over H and W to learn global visual features $h(x) \in R^{C}$.

Attribute Filtering Network. The attribute filtering network, which consists of K $1 \times 1 \times C$ convolutional kernels and a softmax operation, maps visual representations $f(x)$ into attribute feature maps $A(x) \in R^{H \times W \times K}$.

Attribute Embedding Module. The attribute embedding module was designed to learn attribute-level visual features $AF = \{af_i\}_{i=1}^{K}$ from images.

3.3. Class Representation Learning

For a given image x, with a global visual feature denoted by $h(x)$ and a label represented by y, the class semantics vector $\varphi(y)$ serves as the corresponding class prototype cp_y after passing through a class semantics encoder. Cross-entropy loss based on cosine similarity can then be used as the classification loss, which yields more robust results in the visual embedding space [39]. The probability that image x is correctly classified is then given by:

$$p_x = \frac{\exp(\alpha \cdot \cos(h(x), cp_y))}{\sum_{i=1}^{M} \exp(\alpha \cdot \cos(h(x), cp_i))}, \tag{1}$$

where α is a scaling factor. The classification loss for B images in a batch can then be expressed as:

$$L_{\text{cls}} = -\frac{1}{B} \sum_{i=1}^{B} \log p_i. \tag{2}$$

3.4. Attribute Embedding Learning

Attribute Embedding Module. Attribute-level features were acquired by the attribute embedding module illustrated in Figure 2, which accepts visual representations $f(x)$ and attribute feature maps $A(x)$ as input and outputs attribute features (AF). First, the dimensions of $A(x)$ are converted to $K \times HW$ after undergoing sequential reshaping and transposing operations. Similarly, the dimensions of $f(x)$ become $HW \times C$ after a reshaping step. Matrix multiplication of $f(x)$ and $A(x)$ is then performed to acquire the attribute features $AF = \{af_i\}_{i=1}^{K}$.

Attribute Feature Alignment. In this study, solutions to the domain shift problem were pursued by learning the invariance of given attribute features for different objects. Specifically, attribute features were allowed to be distributed around corresponding attribute prototypes, such that the feature af_j of the j-th attribute was as close as possible to its attribute prototype ap_j in the embedding space. It is worth noting that attribute

features in this case should be the attributes belonging to the image x. For a given batch of images, attribute features could then be filtered out if they did not belong to any of the images, producing a new eligible set of attribute-level features $\widehat{AF} = \{\widehat{af_j}\}_{j=1}^{\hat{K}}$. For a given $\widehat{af_j}$, the objective is then to optimize the cosine similarity for a corresponding attribute prototype ap_j while ensuring a considerable distance from other attribute prototypes $ap_{j' \neq j}$ in the embedding space. Triplet loss could then be applied to align attribute features and corresponding attribute prototypes as follows:

$$L_{ass}^{att} = \sum_{j=1}^{\hat{K}} \text{Relu}(\cos(\widehat{af_j}, ap_j) - 0.5 \min_{j' \neq j} \cos(\widehat{af_j}, ap_{j'})). \tag{3}$$

Attribute-level Contrastive Learning. The domain shift problem arises when the visual differences between seen and unseen objects is obvious and their embedding spaces are heterogeneous. In this study, domain shift is addressed by improving the discriminative capacity of the embedding space by bringing the visual features of similar attributes closer together and moving the visual features of different attributes further apart. While general supervised contrastive learning [34] can be used for this purpose, implementing the algorithm without removing easy samples will increase computational complexity and produce a relatively fuzzy boundary between attributes. Specifically, we can optimize attribute representations by introducing a novel contrastive loss for hard samples, which adaptively selects difficult data and learns a highly discriminative attribute embedding space in response. As demonstrated in Figure 4, attribute-level contrastive learning aggregates visual features for the same attributes from different objects, which is beneficial for prompting the model to recognize coarse-grained images. As in attribute alignment loss, filtered attribute features $\widehat{AF} = \{\widehat{af_j}\}_{j=1}^{\hat{K}}$ are used instead of full attribute features. The hard sample selection strategy is then based on cosine similarity rankings. Specifically, for every $\widehat{af_j}$ in $\widehat{AF}$, other attribute features belonging to the same attribute $\widehat{af_j}$ are taken as positive samples, for which the cosine similarities with $\widehat{af_j}$ are sorted from large to small. The top μ ($0 \leq \mu < 1$) simple positives are then excluded and the remaining positives constitute hard positive samples $\{af_{ju}^{+}\}_{u=1}^{U}$. Attribute features belonging to different attributes $\widehat{af_j}$ are then taken as negative samples, for which the cosine similarities with $\widehat{af_j}$ are sorted from small to large. The top ε ($0 \leq \varepsilon < 1$) simple negatives are then excluded and the remaining negatives constitute hard negative samples $\{af_{jv}^{-}\}_{v=1}^{V}$. The similarity between $\widehat{af_j}$ and af_{ju}^{+} can be defined as $S(\widehat{af_j}, af_{ju}^{+}) = \exp(\cos(\widehat{af_j}, af_{ju}^{+})/\tau)$, while the similarity between $\widehat{af_j}$ and af_{jv}^{-} is given by $S(\widehat{af_j}, af_{jv}^{-}) = \exp(\cos(\widehat{af_j}, af_{jv}^{-})/\tau)$. Here, τ denotes a temperature parameter used to control the degree of attention for hard negatives. The selection operates adaptively at the attribute level: for each attribute feature, easy positives and negatives are dynamically excluded based on cosine similarity ranking rather than fixed thresholds. This enables the model to flexibly capture difficult boundary samples. Future work may explore theoretically grounded or uncertainty-driven adaptive strategies to further enhance the technical novelty.

A value of $\tau = 0.1$ was assumed in the experiments and the attribute-level contrastive loss for every $\widehat{af_j}$ in $\widehat{AF}$ was expressed as [34]:

$$p_j = -\frac{1}{U} \sum_{u=1}^{U} \log \frac{S(\widehat{af_j}, af_{ju}^{+})}{\sum_{u=1}^{U} S(\widehat{af_j}, af_{ju}^{+}) + \sum_{v=1}^{V} S(\widehat{af_j}, af_{jv}^{-})}, \tag{4}$$

where U and V denote the number of positives and negatives for $\widehat{af_j}$, respectively. The attribute-level contrastive loss for $\widehat{AF}$ can then be expressed as:

$$L_{cl}^{att} = \frac{1}{\hat{K}} \sum_{j=1}^{\hat{K}} p_j. \tag{5}$$

As illustrated in Algorithm 1, the model adaptively selects hard samples based on cosine similarity ranking, rather than relying on a fixed distance threshold. This enables more flexible capture of feature boundaries between different attributes. Additionally, The design of DCAE, through attribute-level contrastive learning and attribute filtering mechanisms, can automatically enhance the discriminability of key attributes while suppressing the influence of redundant attributes, thereby alleviating, to some extent, the limitations of imperfect attribute definitions. In addition, the attribute embedding distributions and attention weights learned by the model provide insights for attribute design: by analyzing clustering patterns in the embedding space, one can guide attribute selection and granularity optimization, facilitating the future design of more discriminative and generalizable attribute sets.

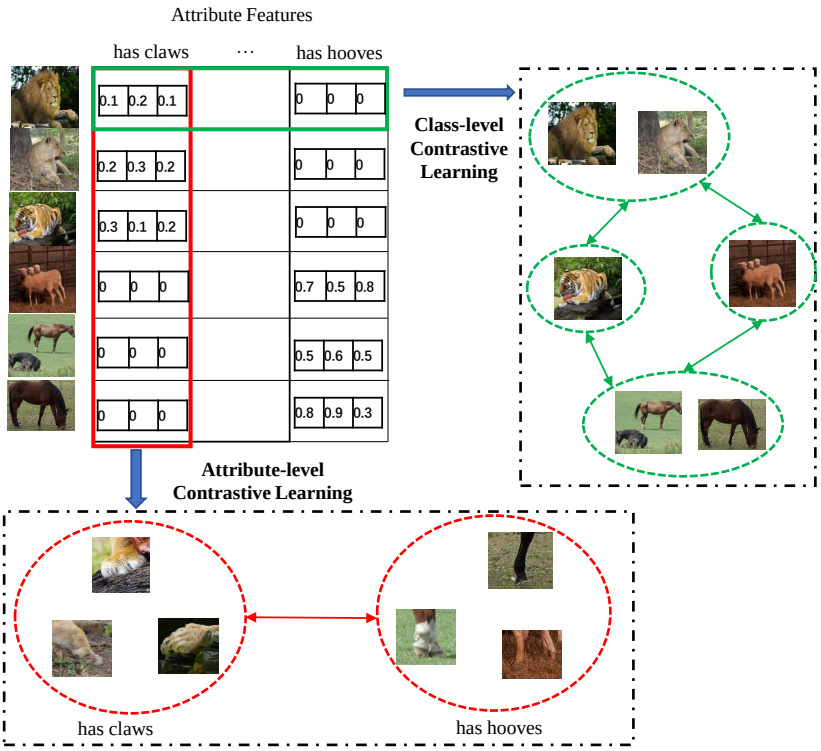

Figure 4. A description of double contrastive learning for attribute features.

Class-level Contrastive Learning. The attribute embedding space was made more discriminative for classes by introducing class-level contrastive learning. As shown in Figure 4, this process brings attribute representations across images (corresponding to the same class) closer together, which is beneficial for recognizing fine-grained images. All attribute-level features corresponding to an image x are then concatenated to acquire an attribute representation vector $ar_x \in R^{KC}$. The eligible set of attribute representation vectors for images in a batch is then given by $AR = \{ar_i\}_{i=1}^{B}$, where B indicates the batch size.

Likewise, the similarity between ar_i and ar_p is defined as $S(ar_i, ar_p) = \exp(\cos(ar_i, ar_p)/\tau)$. Class-level contrastive loss can then be defined as:

$$L_{cl}^{cla} = \frac{1}{B}\sum_{i=1}^{B}\frac{-1}{|P(i)|}\sum_{p\in P(i)}\log\frac{S(ar_i, ar_p)}{\sum_{a=1,a\neq i}^{B} S(ar_i, ar_a)}, \tag{6}$$

where $P(i) \equiv \{p \in \{1,\ldots,B\} : y_p = y_i, p \neq i\}$ is the set of indices for all positives and $|P(i)|$ is the corresponding cardinality.

Algorithm 1 Adaptive Hard Sample Selection based on Cosine Similarity Ranking

Require: Filtered attribute feature set $\widehat{AF} = \{\widehat{af_j}\}_{j=1}^{\hat{K}}$, exclusion ratios μ, ε, temperature coefficient τ
Ensure: Hard positive sample set $\{af_{ju}^{+}\}_{u=1}^{U}$ and hard negative sample set $\{af_{jv}^{-}\}_{v=1}^{V}$

```
for each attribute feature af_j ∈ AF do
    Compute cosine similarity between af_j and all other features in AF
    Positives ← features sharing the same attribute as af_j
    Negatives ← features belonging to different attributes
    Sort Positives in descending order of similarity
    Sort Negatives in ascending order of similarity
    Remove the top μ × |Positives| easy positives
    Remove the top ε × |Negatives| easy negatives
    Remaining positives → hard positive set {af_ju^+}
    Remaining negatives → hard negative set {af_jv^-}
end for
Compute attribute-level contrastive loss L_cl^att according to Equation (5)
```

The dual-contrastive optimization in DCAE consists of attribute-level and class-level contrastive learning. Attribute-level contrastive learning enhances the discriminability of local features by aggregating features of the same attribute and separating features of different attributes. Class-level contrastive learning aggregates samples of the same class to form global class embedding clusters. The combination of both forms a hierarchical embedding structure: local attribute constraints ensure fine-grained discrimination, while class-level constraints ensure coarse-grained discrimination. This structured embedding space reduces the likelihood of unseen class samples falling into incorrect clusters, thereby improving the model's generalization ability.

Attribute Localization. The strength $a(x)$ of all attributes was predicted using global max pooling applied to H and W for the attribute feature maps $A(x)$ [15]. The $a(x)$ term was then optimized by minimizing the mean squared error (MSE) using a class semantics vector $\varphi(y)$ as follows:

$$L_{mse} = \|a(x) - \varphi(y)\|_2^2, \tag{7}$$

where y is the label for image x. This loss is intended to improve the localizability of the attributes.

3.5. Optimization

The overall loss function for the proposed end-to-end model can be defined as:

$$L = L_{cls} + \lambda_1 L_{mse} + \lambda_2 L_{ass}^{att} + \lambda_3 L_{cl}^{att} + \lambda_4 L_{cl}^{cla}, \tag{8}$$

where λ_1, λ_2, λ_3, and λ_4 are hyper-parameters used for MSE loss, attribute-level alignment loss, attribute-level contrastive loss, and class-level contrastive loss, respectively. The joint training of class representation and attribute embedding is critical for GZSL.

3.6. Zero-Shot Recognition

Once the full model is learned, a mapping function from low-level visual features x to class representations $h(x)$ can be directly used for zero-shot learning inference in CZSL. Given an image x, the classifier searches visual embedding spaces with the highest compatibility as follows:

$$\hat{y} = \arg\max_{\tilde{y} \in Y^u} \alpha \cdot \cos(h(x), cp_{\tilde{y}}) . \tag{9}$$

In the case of GZSL, images must be tested from both seen and unseen classes. However, only seen classes are available in the training phase, which leads to significant bias in the predicted results for seen classes [40]. To resolve this issue, a calibrated stacking (CS) [41], which subtracts a calibration factor γ from the scores of seen classes, giving unseen classes a fairer evaluation. DCAE provides highly discriminative attribute- and class-level embeddings, enabling CS to adjust scores more effectively in the embedding space. Together, they ensure reliable recognition performance for unseen classes in GZSL. The GZSL classifier used in this study can be defined as:

$$\hat{y} = \arg\max_{\tilde{y} \in Y^u \cup Y^s} \alpha \cdot \cos(h(x), cp_{\tilde{y}}) - \gamma \mathbb{I}[\tilde{y} \in Y^s], \tag{10}$$

where the α parameter appeared previously in Equation (1) and $\mathbb{I} = 1$ if $\tilde{y}$ belongs to the seen classes ($\mathbb{I} = 0$ otherwise).

4. Experiments

4.1. Datasets

Experiments were conducted with three widely used datasets, including CUB-200-2011 (CUB) [17], SUN [19], and Animals with Attributes 2 (AWA2) [18]. The CUB dataset consists of 11,788 images featuring 200 species of birds. This includes 150 seen classes and 50 unseen classes, each of which exhibits 312 attributes. SUN comprises 14,340 images from 717 scenario categories, including 645 seen samples and 72 unseen samples, with each class containing 102 attributes. AWA2 is an animal dataset comprising 37,322 images from 40 seen classes and 10 unseen classes, with 85 attributes in total.

4.2. Metrics

In CZSL tasks, the top-1 accuracy was evaluated for unseen classes (i.e., *acc*). The harmonic mean $H = (2 \times S \times U)/(S + U)$ was then used to validate GZSL model performance [18]. In this case, U and S denote the top-1 accuracy for unseen and seen classes, respectively.

4.3. Implementation Details

The image encoder consisted of a ResNet101 [42] backbone pretrained on ImageNet [43]. The included class semantics and attribute semantics encoders consisted of multilayer perceptrons with two 1024-unit hidden layers and a 2048-unit output layer. A ReLU activation function was then adopted in all fully connected layers and contrastive learning (AF) was implemented using a liner layer, in which the output size was 1024. An SGD [44] optimizer was also adopted with a momentum of 0.9 and a weight decay of 10^{-5}. The initial learning rate of 10^{-3} decayed every 10 epochs with a decay factor of 0.5 and hyperparameters in the model were acquired using a grid search applied to

the validation set. The temperature parameter τ for both attribute-level and class-level contrastive loss steps was set to 0.1. When selecting hard samples, μ was set to 0.32 and ε was set to 0.42 for all datasets, with a scale factor (α) of 25. The calibration factor γ was 1.0 for AWA2 and 0.7 for CUB and SUN. The M-way N-shot episode-based training method was also applied with $M = 16$ and $N = 2$ for CUB and AWA2. Similarly, values of $M = 8$ and $N = 2$ were used for SUN.

4.4. Comparisons with State-of-the-Art Methods

The superiority of the proposed DCAE technique was demonstrated through comparisons with representative state-of-the-art ZSL methods, which can be categorized into embedding-based and generative-based zero-shot learning models.

CZSL Performance. DCAE was compared with recent state-of-the-art models, which can be divided into generative and embedding methods. Table 1 shows CZSL results for all three datasets. In the case of CZSL, DCAE consistently achieved the highest accuracies, including values of 77.0%, 67.2%, and 74.4% for CUB, SUN, and AWA2, representing improvements of 0.9%, 1.2%, and 0.8%, respectively. These results suggest that our model can effectively transfer knowledge learned from seen classes to unseen classes. The obvious increases in accuracy for the SUN dataset are likely the result of improvements in the mining of commonalities between features from the same attribute, facilitating the accurate recognition of attributes from unseen classes. The excellent performance for CUB confirms that our method can effectively achieve fine-grained recognition due to its capacity to learn discriminative attribute embedding spaces.

GZSL Performance. Table 1 displays the results of applying the proposed method and other state-of-the-art algorithms for GZSL. It is evident from the table that DCAE achieved the best performance with H values of 74.0%, 43.6%, and 76.9% for CUB, SUN, and AWA2, respectively. This represents an improvement of 4.1% for AWA2, which is particularly significant. It is also worth noting the top-1 accuracies for unseen classes in both CUB and AWA2 are the highest among all methods, which confirms that DCAE can effectively overcome the domain shift problem and is more discriminative for object classification tasks. In addition, the accuracy for unseen classes in CUB was 2.8% higher than the accuracy for seen classes produced by the MSDN method [28], indicating that existing models are more prone to overfitting with seen classes. These benefits are a result of double contrastive optimization for attribute features, which enables the proposed model to recognize attributes in target classes with the help of learned knowledge.

DCAE demonstrates relatively robust performance on coarse-grained attributes because these attributes reflect prominent differences between classes, and the corresponding visual features are easy to capture, allowing contrastive learning to directly align the features. However, for fine-grained attributes (SUN), the inter-class differences are subtle, requiring precise local feature extraction and accurate alignment. For highly abstract attributes or those requiring complex reasoning, the model may struggle to fully capture discriminative information, leading to potential confusion in the embedding space. To address the challenges posed by coarse- and fine-grained attributes, future work could explore cross-modal attribute enhancement, attention mechanisms, multi-step reasoning modules, and adaptive feature alignment strategies to further improve the model's generalization across varying attribute granularities and heterogeneous data scenarios.

Table 1. CZSL and GZSL results for the CUB, SUN, and AWA2 datasets. The best and second-best results are marked in red and blue, respectively. The - symbol indicates that no results were reported. Methods are categorized into embedding-based and generative-based zero-shot learning techniques. Validation metrics are expressed as percentages.

Type	Methods	CUB				SUN				AWA2			
		CZSL	GZSL			CZSL	GZSL			CZSL	GZSL		
		acc	*U*	*S*	*H*	*acc*	*U*	*S*	*H*	*acc*	*U*	*S*	*H*
Generative	f-CLSWGAN [45]	57.3	43.7	57.7	49.7	60.8	42.6	36.6	39.4	68.2	57.9	61.4	59.6
	f-VAEGAN-D2 [46]	61.0	48.4	60.1	53.6	64.7	45.1	38.0	41.3	71.1	57.6	70.6	63.5
	LisGAN [47]	58.8	46.5	57.9	51.6	61.7	42.9	37.8	40.2	-	-	-	-
	TF-VAEGAN [48]	64.9	52.8	64.7	58.1	66.0	45.6	40.7	43.0	72.2	59.8	75.1	66.6
	HSVA [49]	-	52.7	58.3	55.3	-	48.6	39.0	43.3	-	56.7	79.8	66.3
	ICCE [50]	-	67.3	65.5	66.4	-	-	-	-	-	65.3	82.3	72.8
Embedding	TCN [51]	59.5	52.6	52.0	52.3	61.5	31.2	37.3	34.0	71.2	61.2	65.8	63.4
	DAZLE [11]	66.0	56.7	59.6	58.1	59.4	52.3	24.3	33.2	67.9	60.3	75.7	67.1
	RGEN [13]	76.1	60.0	73.5	66.1	63.8	44.0	31.7	36.8	73.6	67.1	76.5	71.5
	APN [15]	72.0	65.3	69.3	67.2	61.6	41.9	34.0	37.6	68.4	57.1	72.4	63.9
	DCEN [52]	-	63.8	78.4	70.4	-	43.7	39.8	41.7	-	62.4	81.7	70.8
	MSDN [28]	76.1	68.7	67.5	68.1	65.8	52.2	34.2	41.3	70.1	62.0	74.5	67.7
	DCAE(Ours)	77.0	70.3	78.1	74.0	67.2	46.2	41.2	43.6	74.4	69.3	86.4	76.9

4.5. Ablation Studies

Component Analysis. This section evaluates the effectiveness of various components included in the proposed method by conducting a series of ablation studies with three datasets. The effects of each DCAE component are presented in Table 2. The extraction of attribute-related discriminative features is crucial for zero-shot learning, as it is key to bridging semantic gaps and overcoming the domain shift problem. We first defined $L_{cls} + \lambda_1 L_{mse}$, which represents both classification loss in class representation learning and MSE loss in attribute localization. Triplet loss was then added, followed sequentially by attribute-level contrastive loss without hard sample selection, attribute-level contrastive loss with hard sample selection, and class-level contrastive loss. The results clearly indicated that L_{ass}^{att}, L_{cl}^{att}, and L_{cl}^{cla} improved initial model performance. Specifically, the L_{ass}^{att} term could also be used to learn clear attribute prototypes and to align attribute features with these corresponding prototypes. The learning of attribute features was further strengthened by using L_{cl}^{att} to make features from the same attribute more compact in the embedding space, with clearer boundaries between the features of different attributes. Improvements in H were observed to be 3.6%, 1.9%, and 2.3% for CUB, SUN, and AWA2 after the addition of L_{ass}^{att} and L_{cl}^{att}, respectively, indicating the combined importance of attribute prototype generation and attribute-level feature optimization.

Performance differences were evaluated for methods both using and not using hard sample selection for attribute-level contrastive learning. These results demonstrated that hard sample selection improved outcomes for CUB, SUN, and AWA2 samples by 2.6%, 0.1%, and 1.6%, respectively. Decreases in accuracy for unseen classes in the SUN dataset may be attributable to the more abstract nature of attribute-level features in scene images. Specifically, since features of this type typically require global image consideration, contrastive learning mechanisms based on negative sample selection may not effectively address this issue. However, accuracy for seen classes improved across all datasets. As such, selecting hard samples effectively improved the discriminative capacity of the embedding space. An L_{cl}^{cla} term was also added, further enabling our model to recognize different types of objects and taking into account the overall condition of all attributes. Improvements in H

were 0.3%, 0.3%, and 1.1% for CUB, SUN, and AWA2 after the addition of L_{cl}^{cla}, indicating that class-level contrastive loss can further improve performance.

Table 2. The results of an ablation study involving three datasets. The best results in each case are marked in **bold**, while HS denotes hard sample selection. Validation metrics are expressed as percentages.

Method	CUB			SUN			AWA2		
	U	*S*	*H*	*U*	*S*	*H*	*U*	*S*	*H*
$L_{cls} + \lambda_1 L_{mse}$	63.7	77.9	70.1	40.9	41.9	41.4	63.1	87.9	73.5
$+\lambda_2 L_{ass}^{att}$	67.8	74.5	71.0	43.6	42.0	42.8	62.8	**88.4**	73.4
$+\lambda_3 L_{cl}^{att}$ w/o HS	65.5	77.8	71.1	**47.2**	39.8	43.2	65.4	85.9	74.2
$+\lambda_3 L_{cl}^{att}$ w HS	69.8	78.1	73.7	43.8	**42.8**	43.3	67.5	86.5	75.8
$+\lambda_4 L_{cl}^{cla}$	**70.3**	**78.1**	**74.0**	46.2	41.2	**43.6**	**69.3**	86.4	**76.9**

Training Method Analysis. The validation experiments utilized an episode-based training approach to enhance model generalizability. Specifically, M categories and N images were sampled from each mini-batch, in which M varied between {4,8,16} and N was fixed at 2. In addition, the impact of these values on model performance was recorded and compared with a random sampling approach utilizing a mini-batch size of 32. Table 3 indicates this episode-based training approach achieved better performance than random sampling for all datasets. This is likely because the number of training categories is fixed, which can result in the formation of more reliable positive and negative sample pairs. For example, our model produced the highest accuracy for CUB and AWA2 with $M = 16$ and $N = 2$ and for SUN with $M = 8$ and $N = 2$.

Table 3. The influence of training methodology on GZSL results. The best outcomes are marked in **bold** and validation metrics are expressed as percentages.

	Training Method	M-Way	N-shot	U	S	H
CUB	Random Sampling	-	-	69.8	68.5	69.1
	Episode-Based	4	2	65.9	74.3	69.8
		8	2	67.5	**78.4**	72.6
		16	2	**70.3**	78.1	**74.0**
SUN	Random Sampling	-	-	**55.7**	34.0	42.2
	Episode-Based	4	2	47.2	39.5	43.0
		8	2	46.2	41.2	**43.6**
		16	2	42.1	**43.8**	42.9
AWA2	Random Sampling	-	-	61.4	**88.1**	72.4
	Episode-Based	4	2	60.2	85.7	70.7
		8	2	66.4	83.5	74.0
		16	2	**69.3**	86.4	**76.9**

4.6. Computational Complexity Analysis

To evaluate the computational efficiency of the proposed DCAE model, we measure the computational cost using FLOPs, which represent the number of floating-point operations required for a single forward pass. We compare DCAE with several representative embedding-based zero-shot learning methods, including AREN [12], APN [15], DCEN [52], DPPN [29], and GEM-ZSL [39]. All experiments are conducted on the CUB dataset with a batch size of 1, and the backbone ResNet-101 is fine-tuned during training, meaning that the parameters of the feature extraction network are updated in each iteration. Table 4 presents the comparison of FLOPs among different methods. As shown, all methods are within the

same order of magnitude (10^{10}). GEM-ZSL achieves the lowest FLOPs (3.13×10^{10}), indicating the highest computational efficiency, while DCEN has the highest FLOPs (8.46×10^{10}) due to its dual ResNet-101 architecture in feature extraction. In comparison, DCAE requires 4.05×10^{10} FLOPs, which is lower than that of DCEN and DPPN, demonstrating better computational efficiency. Furthermore, DCAE introduces only a lightweight covariance-adaptive modeling module in the feature enhancement stage, which does not significantly increase the number of parameters or memory usage. As a result, both training and inference remain computationally feasible and efficient. Although DCAE is not the most computationally efficient among all compared methods, it achieves a favorable balance between computational complexity and recognition performance, exhibiting strong competitiveness in both conventional and generalized zero-shot learning tasks.

Table 4. Comparison of FLOPs among different ZSL methods on the CUB dataset. The bolded FLOPs values indicate the minimum computational cost among all compared models.

Method	FLOPs
AREN	3.15×10^{10}
APN	3.18×10^{10}
DCEN	8.46×10^{10}
DPPN	4.13×10^{10}
GEM-ZSL	$\mathbf{3.13 \times 10^{10}}$
DCAE (Ours)	4.05×10^{10}

4.7. Hyperparameter Analysis

The Importance of Hard Sample Selection. The selection of positive and negative samples is critical for contrastive learning [53]. Samples from the same class exhibiting lower correlation are defined as hard positives, while samples from different classes exhibiting higher correlation are defined as hard negatives. In this case, hard samples were beneficial for improving the discriminative capacity of the attribute embedding space. CZSL and GZSL results for CUB were then recorded by varying μ and ε from {0.32, 0.35, 0.42, 0.45}, while fixing the other parameters as defaults. As shown in Figure 5, our model performed poorly when μ was set too large or ε was set too small, due to the number of positives being smaller than the number of negatives. Removing too many easy positives and too few easy negatives can also limit the robustness and discriminative capacity of the attribute embedding space. These results also indicated that our model achieved the best GZSL performance for CUB when μ and ε were set to 0.32 and 0.42, respectively. The same values were used for SUN and AWA2 for convenience.

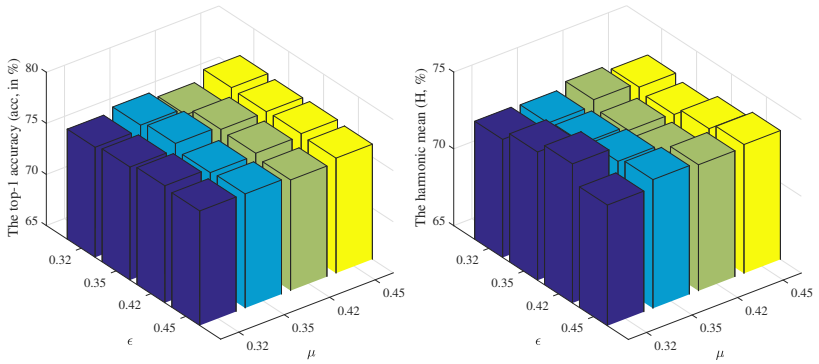

Figure 5. CUB recognition results using different values of μ and ε, expressed as a percentage.

Balance Factors for λ_3 and λ_4. The effects of λ_3 and λ_4 on DCAE performance were evaluated using the CUB and AWA2 datasets. Figure 6 shows U, S, and H results for both CUB and AWA2 samples when varying λ_3 and λ_4 over {0.01, 0.05, 0.1, 0.5, 1}. The left subplot shows GZSL performance for the CUB dataset when λ_4 was set to 0.05 and λ_3 was varied. The right subplot shows GZSL performance when λ_3 was set to 0.8 and λ_4 was varied. Similar results are shown for AWA2 data, with λ_4 set to 0.1 and λ_3 varied (left subplot) and λ_3 set to 1 with λ_4 varied (right subplot). As seen in the figure, H was incremented with gradual increases in λ_3 and λ_4, indicating gradual improvements in network performance. This is because the incorporation of attribute-level contrastive loss and class-level contrastive loss force DCAE to mine commonalities between the same attributes from different objects. However, as λ_3 and λ_4 increase further, the accuracy decreases slightly since overly large coefficient values hinder the learning of discriminative class representations. Also, CUB is a fine-grained dataset in which different types of birds are not easy to distinguish visually. Thus, in order to improve the discriminative capacity of the proposed model, the value of λ_4 should not be too small. In contrast, AWA2 is a coarse-grained dataset, in which the same attributes for different types of animals are visually distinct. As such, λ_3 must be set to a larger value to constrain the distribution of attribute-level features in the embedding space.

As shown in Figures 4 and 5, the proposed model exhibits stable performance across a wide range of hyperparameter values, indicating strong robustness to moderate variations in μ, ε, λ_3 and λ_4. When these hyperparameters slightly deviate from their optimal values, the overall performance remains relatively consistent, suggesting that the optimization process is not sensitive to small perturbations. This stability mainly stems from the adaptive balancing mechanism between the attribute-level and class-level contrastive learning objectives, which effectively prevents any single loss component from dominating the optimization during training.

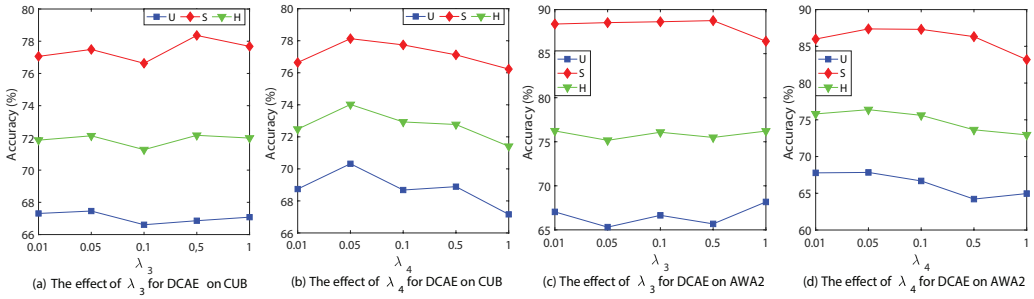

Figure 6. GZSL results for CUB and AWA2 data using different values of λ_3 and λ_4.

4.8. Qualitative Results

A t-SNE Visualization of Class Representations. Figure 7 presents t-SNE visualization results [54] for class representations (i.e., global visual features $h(x)$) from both seen and unseen classes in the CUB and AWA2 datasets. The experiments involved the first 10 seen classes and 10 unseen classes from CUB and the first 10 seen classes and all unseen classes from AWA2. It is evident from these visualization results that the distributions of representations for a single given class were compact, while the representations for different classes offered good separability, which confirms the proposed model was able to learn a highly discriminative visual space. Visualization results for unseen classes also illustrated that our model could effectively overcome the domain shift problem.

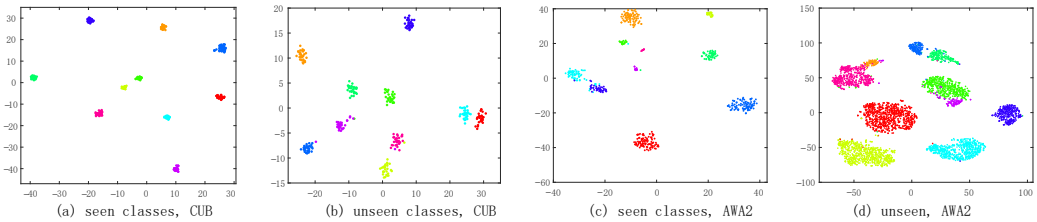

Figure 7. The t-SNE visualization of class representations for seen and unseen classes from CUB and AWA2. Different colors represent different categories.

The t-SNE Visualization of Attribute Features. The discriminative capacity of the attribute embedding space was further validated using a t-SNE visualization of attribute-level features from the first 15 attributes in multiple CUB and AWA2 images. Figure 8 demonstrates that attribute features from different attributes were easily separated, while features from the same attributes were clustered together. This outcome confirms the proposed model can mine common traits from features across images belonging to the same attribute, offering the ability to accurately identify attributes, which is also conducive to overcoming the domain shift problem.

Although the t-SNE visualization results in Figures 6 and 7 demonstrate the proposed model's advantages in feature discriminability, we also observed several challenging cases. Specifically, when fine-grained categories exhibit highly similar visual characteristics (e.g., bird species with nearly identical colors or shapes) or when semantic attribute descriptions lack sufficient distinctiveness, the model may generate less separable embeddings, leading to feature alignment confusion and recognition errors for unseen classes. In future work, we plan to introduce semantic disambiguation mechanisms and uncertainty-aware feature optimization strategies to further alleviate such failure cases and enhance the model's robustness in complex scenarios.

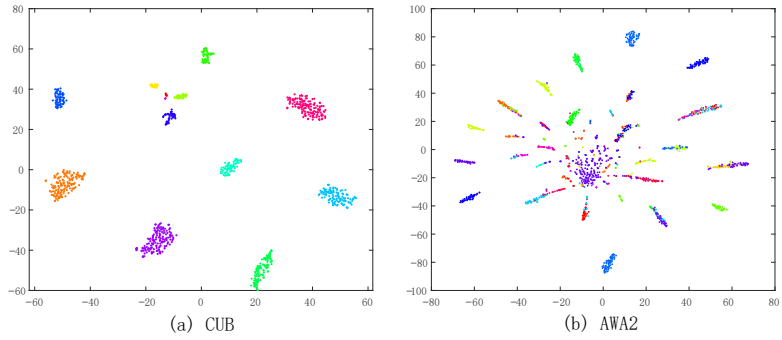

Figure 8. A t-SNE visualization of attribute features from CUB and AWA2. Different colors represent different categories.

5. Conclusions

This study introduced a novel embedding-based ZSL framework, termed dual-contrastive attribute embedding (DCAE), for generalized zero-shot learning. This approach overcomes the domain shift problem by explicitly learning discriminative attribute-level features and well-separated attribute prototypes, both of which are optimized by attribute-level contrastive loss and class-level contrastive loss. In addition to the optimization of attribute features, DCAE also learns class representations and class prototypes in visual space. Extensive experiments on three popular benchmark datasets demonstrated the superiority of DCAE over existing state-of-the-art ZSL methods.

Author Contributions: Conceptualization, Q.L.; methodology, K.J. and Q.L.; writing—original draft preparation, Q.L.; writing—review and editing, K.J., Z.Z., and Y.L.; supervision, K.J.; funding acquisition, K.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Department of Science and Technology of Guangdong Province through the Guangdong Basic and Applied Basic Research Foundation, grant number 2022A1515110667.

Data Availability Statement: Publicly available datasets analyzed in this study can be found here: CUB-200-2011 (CUB), http://www.vision.caltech.edu/datasets/cub_200_2011 (accessed on 3 October 2025); SUN, https://cs.brown.edu/~gmpatter/sunattributes.html (accessed on 3 October 2025); Animals with Attributes 2 (AWA2), https://cvml.ist.ac.at/AwA2/ (accessed on 3 October 2025).

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

Abbreviation	Full Term
ZSL	Zero-shot Learning
DCAE	Dual-contrastive Attribute Embedding
CZSL	Conventional ZSL
GZSL	Generalized ZSL
GANs	Generative Adversarial Networks
VAEs	Variational Autoencoders
MSE	Mean Squared Error
CS	Calibrated Stacking
CUB	CUB-200-2011
AWA2	Animals with Attributes 2

References

1. Xie, G.S.; Zhang, X.Y.; Shu, X.; Yan, S.; Liu, C.L. Task-driven feature pooling for image classification. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1179–1187.
2. Zhang, Z.; Xu, Y.; Shao, L.; Yang, J. Discriminative block-diagonal representation learning for image recognition. *IEEE Trans. Neural Networks Learn. Syst.* **2017**, *29*, 3111–3125. [CrossRef]
3. Xian, Y.; Schiele, B.; Akata, Z. Zero-shot learning-the good, the bad and the ugly. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4582–4591.
4. Palatucci, M.; Pomerleau, D.; Hinton, G.E.; Mitchell, T.M. Zero-shot learning with semantic output codes. *Adv. Neural Inf. Process. Syst.* **2009**, *22*.. [CrossRef]
5. Li, J.; Jing, M.; Lu, K.; Zhu, L.; Shen, H.T. Investigating the bilateral connections in generative zero-shot learning. *IEEE Trans. Cybern.* **2021**, *52*, 8167–8178. [CrossRef] [PubMed]
6. Li, Z.; Chen, Q.; Liu, Q. Augmented semantic feature based generative network for generalized zero-shot learning. *Neural Networks* **2021**, *143*, 1–11. [CrossRef] [PubMed]
7. Lampert, C.H.; Nickisch, H.; Harmeling, S. Attribute-based classification for zero-shot visual object categorization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *36*, 453–465. [CrossRef]
8. Xu, X.; Tsang, I.W.; Liu, C. Complementary attributes: A new clue to zero-shot learning. *IEEE Trans. Cybern.* **2019**, *51*, 1519–1530. [CrossRef]
9. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. *Adv. Neural Inf. Process. Syst.* **2013**, *26, 3111–3119*.
10. Chen, S.; Hong, Z.; Liu, Y.; Xie, G.S.; Sun, B.; Li, H.; Peng, Q.; Lu, K.; You, X. Transzero: Attribute-guided transformer for zero-shot learning. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 330–338. [CrossRef]
11. 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, 14–19 June 2020; pp. 4483–4493.

12. 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.
13. Xie, G.S.; Liu, L.; Zhu, F.; Zhao, F.; Zhang, Z.; Yao, Y.; Qin, J.; Shao, L. Region graph embedding network for zero-shot learning. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 562–580.
14. Zhu, Y.; Xie, J.; Tang, Z.; Peng, X.; Elgammal, A. Semantic-guided multi-attention localization for zero-shot learning. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, *14943–14953*.
15. Xu, W.; Xian, Y.; Wang, J.; Schiele, B.; Akata, Z. Attribute prototype network for zero-shot learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21969–21980.
16. Fu, Y.; Hospedales, T.M.; Xiang, T.; Gong, S. Transductive multi-view zero-shot learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 2332–2345. [CrossRef] [PubMed]
17. Welinder, P.; Branson, S.; Mita, T.; Wah, C.; Schroff, F.; Belongie, S.; Perona, P. Caltech-UCSD Birds 200 (CUB-200); Technical Report CNS-TR-2010-001, California Institute of Technology, Pasadena, CA, USA, 29 September 2010. Available online: https://authors.library.caltech.edu/records/cyyh7-dkg06 (accessed on 3 November 2025).
18. 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]
19. Patterson, G.; Hays, J. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2751–2758.
20. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
21. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114.
22. Liu, Y.; Gao, X.; Han, J.; Shao, L. A discriminative cross-aligned variational autoencoder for zero-shot learning. *IEEE Trans. Cybern.* **2022**, *53*, 3794–3805. [CrossRef]
23. Frome, A.; Corrado, G.S.; Shlens, J.; Bengio, S.; Dean, J.; Ranzato, M.; Mikolov, T. Devise: A deep visual-semantic embedding model. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, *2121–2129*.
24. Romera-Paredes, B.; Torr, P. An embarrassingly simple approach to zero-shot learning. In Proceedings of the International Conference on Machine Learning, PMLR, Lille, France, 7–9 July 2015; pp. 2152–2161.
25. Yun, Y.; Wang, S.; Hou, M.; Gao, Q. Attributes learning network for generalized zero-shot learning. *Neural Networks* **2022**, *150*, 112–118. [CrossRef] [PubMed]
26. Li, Q.; Hou, M.; Lai, H.; Yang, M. Cross-modal distribution alignment embedding network for generalized zero-shot learning. *Neural Networks* **2022**, *148*, 176–182. [CrossRef]
27. Ji, Z.; Yu, X.; Yu, Y.; Pang, Y.; Zhang, Z. Semantic-guided class-imbalance learning model for zero-shot image classification. *IEEE Trans. Cybern.* **2021**, *52*, 6543–6554. [CrossRef]
28. Chen, S.; Hong, Z.; Xie, G.S.; Yang, W.; Peng, Q.; Wang, K.; Zhao, J.; You, X. MSDN: Mutually Semantic Distillation Network for Zero-Shot Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 7612–7621.
29. Wang, C.; Min, S.; Chen, X.; Sun, X.; Li, H. Dual Progressive Prototype Network for Generalized Zero-Shot Learning. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 2936–2948.
30. Cunegatto, E.H.T.; Zinani, F.S.F.; Rigo, S.J. Multi-objective optimisation of micromixer design using genetic algorithms and multi-criteria decision-making algorithms. *Int. J. Hydromechatronics* **2024**, *7*, 224–249. [CrossRef]
31. Yazdani, K.; Fardindoost, S.; Frencken, A.L.; Hoorfar, M. Multi-objective optimization of expansion-contraction micromixer using response surface methodology: A comprehensive study. *Int. J. Heat Mass Transf.* **2024**, *227*, 125570. [CrossRef]
32. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 13–18 July 2020; pp. 1597–1607.
33. Jeon, S.; Min, D.; Kim, S.; Sohn, K. Mining better samples for contrastive learning of temporal correspondence. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2021; pp. 1034–1044.
34. Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; Krishnan, D. Supervised contrastive learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 18661–18673.
35. Le-Khac, P.H.; Healy, G.; Smeaton, A.F. Contrastive representation learning: A framework and review. *IEEE Access* **2020**, *8*, 193907–193934. [CrossRef]
36. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 9729–9738.
37. Lim, J.Y.; Lim, K.M.; Lee, C.P.; Tan, Y.X. SCL: Self-supervised contrastive learning for few-shot image classification. *Neural Networks* **2023**, *165*, 19–30. [CrossRef]

38. Han, Z.; Fu, Z.; Chen, S.; Yang, J. Contrastive embedding for generalized zero-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2021; pp. 2371–2381.
39. Liu, Y.; Zhou, L.; Bai, X.; Huang, Y.; Gu, L.; Zhou, J.; Harada, T. Goal-oriented gaze estimation for zero-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2021; pp. 3794–3803.
40. Liu, Z.; Li, Y.; Yao, L.; Wang, X.; Long, G. Task aligned generative meta-learning for zero-shot learning. *Proc. AAAI Conf. OnArtificial Intell.* **2021**, *35*, 8723–8731. [CrossRef]
41. Chao, W.L.; Changpinyo, S.; Gong, B.; Sha, F. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 52–68.
42. 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.
43. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
44. Bottou, L. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 177–186.
45. Xian, Y.; Lorenz, T.; Schiele, B.; Akata, Z. Feature generating networks for zero-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5542–5551.
46. Xian, Y.; Sharma, S.; Schiele, B.; Akata, Z. f-vaegan-d2: A feature generating framework for any-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10275–10284.
47. Li, J.; Jing, M.; Lu, K.; Ding, Z.; Zhu, L.; Huang, Z. Leveraging the invariant side of generative 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. 7402–7411.
48. Narayan, S.; Gupta, A.; Khan, F.S.; Snoek, C.G.; Shao, L. Latent embedding feedback and discriminative features for zero-shot classification. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 479–495.
49. Chen, S.; Xie, G.; Liu, Y.; Peng, Q.; Sun, B.; Li, H.; You, X.; Shao, L. Hsva: Hierarchical semantic-visual adaptation for zero-shot learning. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 16622–16634.
50. Kong, X.; Gao, Z.; Li, X.; Hong, M.; Liu, J.; Wang, C.; Xie, Y.; Qu, Y. En-Compactness: Self-Distillation Embedding & Contrastive Generation for Generalized Zero-Shot Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 9306–9315.
51. Jiang, H.; Wang, R.; Shan, S.; Chen, X. Transferable contrastive network for generalized zero-shot learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9765–9774.
52. Wang, C.; Chen, X.; Min, S.; Sun, X.; Li, H. Task-independent knowledge makes for transferable representations for generalized zero-shot learning. *Proc. AAAI Conf. OnArtificial Intell.* **2021**, *35*, 2710–2718. [CrossRef]
53. Robinson, J.; Chuang, C.Y.; Sra, S.; Jegelka, S. Contrastive learning with hard negative samples. *arXiv* **2020**, arXiv:2010.04592.
54. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.