

Dual Feature Augmentation Network for Generalized Zero-shot Learning

Lei Xiang¹
xl294487391@gmail.com

Yuan Zhou^{1†}
zhouyuan@nuist.edu.cn

Haoran Duan²
haoran.duan@ieee.org

Yang Long²
yang.long@ieee.org

¹ School of Artificial Intelligence
Nanjing University of Information
Science and Technology
Nanjing, China

² Department of Computer Science
Durham University
Durham, UK

Abstract

Zero-shot learning (ZSL) aims to infer novel classes without training samples by transferring knowledge from seen classes. Existing embedding-based approaches for ZSL typically employ attention mechanisms to locate attributes on an image. However, these methods often ignore the complex entanglement among different attributes' visual features in the embedding space. Additionally, these methods employ a direct attribute prediction scheme for classification, which does not account for the diversity of attributes in images of the same category. To address these issues, we propose a novel Dual Feature Augmentation Network (DFAN), which comprises two feature augmentation modules, one for visual features and the other for semantic features. The visual feature augmentation module explicitly learns attribute features and employs cosine distance to separate them, thus enhancing attribute representation. In the semantic feature augmentation module, we propose a bias learner to capture the offset that bridges the gap between actual and predicted attribute values from a dataset's perspective. Furthermore, we introduce two predictors to reconcile the conflicts between local and global features. Experimental results on three benchmarks demonstrate the marked advancement of our method compared to state-of-the-art approaches. Our code is available at <https://github.com/Sion1/DFAN>

1 Introduction

The effectiveness of deep learning algorithms in recognizing various image types heavily relies on a significant amount of annotated data. However, the process of obtaining labeled data for all categories can be a daunting and impractical task. To tackle this challenge, ZSL offers a feasible solution by transferring knowledge from seen to unseen classes. ZSL encompasses two distinct settings depending on the testing approach: conventional ZSL (CZSL) [19, 25] and generalized ZSL (GZSL) [8, 28]. CZSL only detects unseen classes, whereas GZSL accounts for both seen and unseen classes during testing.

[†]Corresponding author

Both CZSL and GZSL transfer knowledge from seen classes to unseen classes by semantic information such as manually-annotated attributes [17], sentence features [28], and word vectors [34]. Among them, attributes as shared class-level semantic features have gained significant popularity. However, a class’s attribute vector is computed by averaging the probability of the attribute occurrence across all images within that class [26, 36, 57], which results in inaccurate predictions due to the diversity of attributes in different images belonging to the same category. To address this limitation, we propose a bias learner that augments the class-level semantic features for every image.

On the other hand, relying solely on the seen classes during the training process makes models more prone to bias toward these seen classes. Therefore, generative-based methods [6, 30, 33, 39, 42] are becoming increasingly popular as they complement the information on the unseen classes. These approaches generate visual features or images of the unseen classes, thus transforming the ZSL task into a conventional classification problem and significantly reducing the bias towards the seen classes. The generative-based methods achieve considerable success but still encounter several challenges. These challenges encompass optimization difficulties stemming from the notable disparity between synthetic and authentic features, as well as extensive resource consumption during training attributed to the model’s substantial size.

In order to avoid dependence on generative models, alternative approaches [10, 9, 8, 11, 29, 41, 47] learn a single mapping function to fuse different granularity features (*i.e.*, global feature and local feature) and align them with the corresponding semantic features. However, the global features are class-oriented while the local features are attribute-oriented, and employing a single mapping function will lead to a decline in the model’s performance. In contrast to these methods, our approach employs two distinct mapping functions (local predictor and global predictor) for multilevel features to achieve visual-to-semantic mapping.

Furthermore, most embedding-based methods employ attributes as a mediator between seen and unseen classes, but these attributes are only represented in a limited number of regions. Consequently, utilizing global image features directly is not reasonable, and some methods [6, 8, 12, 16, 35] integrate attention mechanisms [9, 32] into the model to locate attribute in the image. Despite their effectiveness, they neglect the entanglement of the attribute features in the embedding space. We explicitly extract attribute features and adopt a cosine similarity loss function to overcome this limitation and enhance visual features.

Our contribution is summarized as follows: (1) We propose a semantic feature augmentation containing a bias learner, which estimates an offset to alleviate the difference between the actual and predicted attributes, leading to improved class-level semantic features for each image. (2) We employ two mapping functions to reinforce the mapping procedure of varying levels of features into the semantic space, thus avoiding inconsistencies in the mapping process of different granular features. (3) We propose a visual feature augmentation that explicitly extracts attribute features and adopts a cosine similarity loss to decouple them in the embedding space, thus enhancing the visual features. (4) We conduct comprehensive experiments on three ZSL benchmarks, demonstrating that our method achieves superior or competitive performance compared to state-of-the-art ZSL methods. Furthermore, our ablation studies on different modules support the effectiveness of our proposed approach.

2 Related Works

2.1 Generalized Zero-shot Learning

In generalized zero-shot learning (GZSL), modern techniques can be broadly classified into generative-based and embedding-based. Among embedding-based approaches, Li *et al.* [20] proposes an episode-based training scheme to enhance the generalization ability of semantic embedding. Li *et al.* [21] introduces an augmented space incorporating class relationships, improving feature discrimination. Min *et al.* [24] incorporates a domain detector to select perceived categories, thus reducing the impact of perceived categories on visual-semantic alignment. However, these methods depend solely on global features, which poses a challenge in extracting fine-grained information to establish associations between visual and semantic features.

In contrast, generative-based methods in GZSL aim to synthesize visual features of unseen classes by leveraging the shared semantic features. Xian *et al.* [39] proposes a GAN-based approach that employs Wasserstein GAN [2] to generate visual features. Schonfeld *et al.* [30] utilizes VAE [18] to align the distributions learned from images and side-information, constructing latent features that incorporate the necessary multi-modal information linked with unseen classes. Since these methods generate visual features to complement the information of unseen classes, the issue of unseen samples being misclassified as seen classes is mitigated. However, these methods also emphasize learning global image features containing noise, constraining their transferability.

2.2 Part-based ZSL

In order to mitigate the impact of noise caused by image-level representations, recent methods have focused on investigating significant regions. Xu *et al.* [42] utilizes a region search approach to identify the regions associated with each attribute after learning a set of attribute prototypes. Liu *et al.* [22] introduces a gaze estimation module to determine visual attention regions from the perspective of the human gaze. Wang *et al.* [55] progressively adjusts prototypes based on different images and introduces class prototypes to enhance category discriminability. Ge *et al.* [12] proposes a method that incorporates all components and yields a binary mask for identifying significant object regions, facilitating the derivation of refined features that serve as the basis for generating critical weights. Chen *et al.* [8] proposes a dense attribute-based attention mechanism by calculating the similarity between visual and semantic features.

Although recent methods have shown promising results, they ignore differences in representations of the same attribute due to image variations. This limitation weakens the representativeness of the features, thus hindering accurate classification. Moreover, these methods typically employ only one mapping function for learning visual-to-semantic mapping. Since there are varying levels of visual features, using the same function leads to conflicts. We propose using two mapping functions for different visual features to address these issues. Additionally, we design a cosine similarity loss that boosts the distance between attribute features.

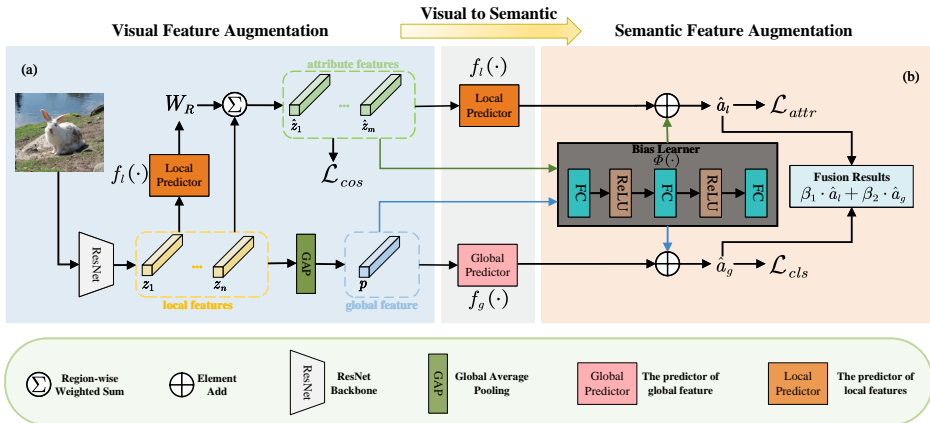


Figure 1: The architecture of DFAN includes two modules: (a) The visual feature augmentation module extracts attribute features and employs a cosine similarity loss to disentangle them. (b) The semantic feature augmentation module aims to learn an offset between actual and predicted values to enhance the semantic features for each image. In addition, two predictors are utilized to map different levels of visual features to the semantic space more effectively. Finally, we fuse different granular feature predictions as the final results to enhance the model’s generalization to different datasets.

3 Method

Problem Definition. This paper focuses on GZSL, which has two disjoint sets. Let $S = \{x_i^s, y_i^s\}_{i=1}^{N_s}$ denotes the training set (seen classes) consisting of C_s classes, where x_i^s is the i -th seen sample and y_i^s is the corresponding label. The testing set (unseen classes) $U = \{x_i^u, y_i^u\}_{i=1}^{N_u}$ consists of C_u classes, where x_i^u is the i -th unseen sample and y_i^u is the corresponding label. Besides, we define the shared semantic features as matrix $A = \{a_j\}_{j=1}^C \in \mathcal{R}^{C \times M}$ where $C = C_s + C_u$ is the total number of classes and M is the number of attributes. In matrix A , each row $a_j \in \mathcal{R}^M$ represents the semantic feature for j -th class. Each element in vector a_j , composed of M attributes, represents the likelihood of the corresponding attribute’s occurrence within the j -th class.

Overview. In this paper, we propose a Dual Feature Augmentation Network (DFAN) that leverages two feature augmentation modules to facilitate knowledge transfer from seen to unseen classes. Specifically, DFAN consists of two modules, the visual feature augmentation module, and the semantic feature augmentation module, as depicted in Figure 1. The visual feature augmentation module extracts the attribute features explicitly by an attention mechanism. To enable the mapping of different levels of visual features to the semantic space, two predictors are employed to map visual features of different levels. Furthermore, a cosine similarity loss is integrated to disentangle visual features of different attributes. Concurrently, the semantic feature augmentation module acquires an offset to mitigate the difference between the actual and predicted attribute values.

3.1 Visual Feature Augmentation

In our approach, ResNet [14] is used for feature extraction. Given an image x , the local features $Z = \{z_i\}_{i=1}^{N=H \times W}$ are extracted from the second last layer of ResNet, where $z_i \in \mathcal{R}^D$ represents the feature of i -th region and D is the embedding dimension. After that, the global feature $p \in \mathcal{R}^D$ is obtained through a global average pooling layer. To extract attribute features, we follow Ke Zhu and Jianxin Wu [45] to adopt an attention mechanism. Instead of modifying the logit directly, we impose constraints on the attribute features and combine the prediction results using two predictors. Specifically, we define two distinct linear layers $f_l(\cdot)$ and $f_g(\cdot)$ without bias term as the local predictor and global predictor, respectively. This architecture enables the efficient mapping of visual features at different levels to the semantic space and avoids potential conflicts between them.

We first use the local predictor to get the local features predict score as matrix $A_l = f_l(Z)$ where $A_l = \{a_j^l\}_{j=1}^N \in \mathcal{R}^{N \times M}$. In matrix A_l , each row a_j^l corresponds to the predicted score for the j -th region. Specifically, the element $a_{i,j}^l$ in the vector a_j^l represents the probability of attribute j appearing at the i -th region. Subsequently, we obtain the weight of each attribute for each region by applying a softmax operation along the rows, given by:

$$w_{i,j}^r = \frac{\exp(a_{i,j}^l)}{\sum_{k=1}^M \exp(a_{k,j}^l)} \quad (1 \leq i \leq N, 1 \leq j \leq M), \quad (1)$$

where $w_{i,j}^r$ is the i -th region weight for j -th attribute. These weights collectively form a matrix $W_R \in \mathcal{R}^{N \times M}$, containing the weights of each region for every attribute. Subsequently, the j -th attribute feature $\hat{z}_j \in \mathcal{R}^D$ is calculated by region-wise weighted sum as:

$$\hat{z}_j = \sum_{k=1}^N w_{k,j}^r \cdot z_k, \quad (2)$$

Similarly, we can get all attribute features as $\hat{Z} = [\hat{z}_1, \dots, \hat{z}_M]$. In order to enhance visual features, we address the entanglement of attributes in the embedding space by reducing their cosine similarities. Specifically, we compute cosine similarities among attribute features and obtain a similarity matrix $\hat{Z}^T \hat{Z}$. The diagonal values in this matrix represent the similarities between the feature and itself, while the other values represent the similarity between different attributes. To decouple attribute features, we make the diagonal values one and the other zero. The ideal similarity matrix should be an identity matrix. Then, we introduce \mathcal{L}_{cos} loss to minimize the L2 norm between the similarity and identity matrix as follows:

$$\mathcal{L}_{cos} = \|\hat{Z}^T \hat{Z} - I\|_2, \quad (3)$$

where $I \in \mathcal{R}^M$ is the identity matrix.

3.2 Semantic Feature Augmentation

To obtain attribute offset for each sample, we design a shared multilayer perceptron $\phi(\cdot)$ to learn such offset. Specifically, we use a 3-layer fully connected layer with ReLU as the activation function after the first two layers. This architecture is chosen because of its capability to capture complex non-linear relationships between input features and attribute offset. Then, we obtain the final prediction result of attribute features \hat{Z} as:

$$\hat{a}_l = \langle W_l, \hat{Z} \rangle + \frac{1}{M} \sum_{k=1}^M \phi(\hat{z}_k), \quad (4)$$

where $W_l \in \mathcal{R}^{D \times M}$ is the linear layer weight of the local predictor and $\langle \cdot, \cdot \rangle$ denotes the column-wise inner product. The second term in Eq. 4 corresponds to an offset between predicted and actual attribute values. Notably, the inner product of a D-dimensional vector and the columns in W_l yield attribute appearance probabilities. Leveraging the attribute features we have, our task is simplified to compare these features with the matrix. The probability of attribute occurrence is consequently obtained through inner product operations on corresponding columns. Then, the model is updated using the cross-entropy loss as:

$$\mathcal{L}_{attr} = -\log \frac{\exp \langle \hat{a}_l, a_y \rangle}{\sum_{k=1}^{C_s} \exp \langle \hat{a}_l, a_k \rangle}, \quad (5)$$

where a_y denotes the ground truth semantic feature of x .

Analogously, we use the same approach to get the prediction result of global feature p :

$$\hat{a}_g = f_g(p) + \phi(p), \quad (6)$$

it is worth noting that the global predictor $f_g(\cdot)$ is used here instead of the local predictor, which can effectively avoid the conflict of different granular features during the mapping process. Subsequently, a cross-entropy loss is adopted to update the model as:

$$\mathcal{L}_{cls} = -\log \frac{\exp \langle \hat{a}_g, a_y \rangle}{\sum_{k=1}^{C_s} \exp \langle \hat{a}_g, a_k \rangle}. \quad (7)$$

Finally, we define the overall loss function of our model as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{attr} + \mathcal{L}_{cls} + \lambda \mathcal{L}_{cos}, \quad (8)$$

where λ is the hyperparameter to control the cosine similarity loss.

3.3 Zero-shot Prediction

CZSL prediction. After training the model, given an image x , we first extract its global feature p and local features Z using ResNet. Subsequently, we derive attribute features \hat{Z} from the above local features using the attention mechanism. After that, the prediction results \hat{a}_g and \hat{a}_l are obtained by two predictors, respectively. Finally, we combine these scores, controlled by coefficients (β_1, β_2) , to get a prediction score and predict the category of the input image by:

$$\hat{y} = \arg \max_{j \in C_u} a_j^T (\beta_1 \cdot \hat{a}_l + \beta_2 \cdot \hat{a}_g) \quad (9)$$

GZSL prediction. For GZSL, we need to consider both seen and unseen classes. We introduce Calibrated Stacking (CS) [9] to reduce the prediction bias towards seen classes. The prediction category is:

$$\hat{y} = \arg \max_{j \in C_s \cup C_u} a_j^T (\beta_1 \cdot \hat{a}_l + \beta_2 \cdot \hat{a}_g) - \gamma \cdot \mathcal{I}(x \in S) \quad (10)$$

where the indicator $\mathcal{I}(\cdot) \in \{0, 1\}$ indicates whether or not x is a seen class and γ is a calibration factor.

4 Experiments and Analysis

4.1 Experimental Settings

Dataset. We evaluated our proposed method on three benchmark datasets for GZSL: Animals with Attributes 2 (AWA2[63]), Caltech-UCSD Birds-200-2011 (CUB)[64], and SUN Attribute (SUN)[27]. AWA2 consists of 37,322 images of 85 different animal species, CUB contains 11,788 images of 200 bird species, and SUN comprises 14,340 images of 717 scenes, each class annotated with 102 attributes. In order to serve as semantic descriptors, we utilized the pre-defined attributes for each dataset. Additionally, we adopted the Proposed Split[63] to divide all classes into seen and unseen classes on each dataset, which allowed for a more comprehensive evaluation of our method’s effectiveness in handling the GZSL problem.

Evaluation. The evaluation methodology proposed in [63] is adopted in our work. For the CZSL scenario, we only measure the per-class Top-1 accuracy on the unseen classes. For the GZSL scenario, we evaluate the model’s performance on both seen and unseen classes, represented as S and U , respectively. To provide a comprehensive evaluation of the model’s GZSL performance, we use the harmonic mean, denoted as H , which is a suitable performance metric and calculated as $H = (2 \times S \times U) / (S + U)$.

Implement Details. We use ResNet101 as the backbone of our model, which is pre-trained on ImageNet1k. It is worth noting that the classes in ImageNet1k do not overlap with the unseen classes in the benchmark datasets. The Adam optimizer is utilized for finetuning the ResNet with a learning rate of 0.00001 and weight decay of 0.0001. We use a learning rate of 0.001 and weight decay of 0.00001 to optimize the linear layer. All experiments are conducted on an NVIDIA RTX 3090 GPU, Intel processor, a memory-sized 32GB, and trained with 80 epochs. Regarding the combined coefficients, we set (0.5, 0.5) for CUB and SUN and (0.0, 1.0) for AWA2 since class-level features are more crucial in coarse-grained datasets like AWA2. We implement all our methods using the PyTorch framework.

4.2 Comparison with State-of-the-Art Methods

Conventional Zero-Shot Learning. We compare the performance of our proposed method with the state-of-the-art methods in the CZSL setting. Table 1 presents the results of CZSL on various datasets. Our proposed method achieves the third-best accuracy of 77.3% and best accuracy of 67.9% on CUB and SUN, respectively. We are only 0.5% away from the best performance on CUB and 0.3% from the second-best performance on this dataset. Our method differs from the second-best result on the SUN dataset by 2.1%. These results prove that our method learns better attribute features and achieves superior results on fine-grained datasets. However, compared to other non-generative methods, we do not obtain better results on coarse-grained datasets such as AWA2. The primary reason is AWA2’s attribute descriptions, which focus on global image content rather than local details. This poses a challenge to part-based methods. Furthermore, unlike other part-based methods like MSDN [7] and TransZero [6], which leverage additional semantic features such as word2vec[23] or GloVe[27] for locating attribute, our DFAN only uses the attribute description given by the dataset.

Methods	CUB				SUN				AWA2			
	GZSL			CZSL	GZSL			CZSL	GZSL			CZSL
	U	S	H	acc	U	S	H	acc	U	S	H	acc
Generative Methods												
f-VAEGAN-D2 [10]	48.4	60.1	53.6	61.0	45.1	38.0	41.3	64.7	57.6	70.6	63.5	71.1
E-PGN [10]	52.0	61.1	56.2	72.4	–	–	–	–	52.6	83.5	64.6	73.4
Composer [10]	56.4	63.8	59.9	69.4	55.1	22.0	31.4	62.6	62.1	77.3	68.8	71.5
GCM-CF [10]	61.0	59.7	60.3	–	47.9	37.8	42.2	–	60.4	75.1	67.0	–
CE-GZSL [10]	63.9	66.8	65.3	77.5	48.8	38.6	43.1	63.3	63.1	78.6	70.0	70.4
FREE [8]	55.7	59.9	57.7	–	47.4	37.2	41.7	–	60.4	75.4	67.1	–
Non-generative Methods												
DAZLE [10]	56.7	59.6	58.1	–	52.3	24.3	33.2	–	60.3	75.3	67.1	–
APN [10]	65.3	69.3	67.2	72.0	41.9	34.0	37.6	61.6	57.1	72.4	63.9	68.4
GEM-ZSL [10]	64.8	77.1	70.4	77.8	38.1	35.7	36.9	62.8	64.8	77.5	70.6	67.3
SR2E [10]	61.6	70.6	65.8	–	43.1	36.8	39.7	–	58.0	80.7	67.5	–
MSDN [8]	65.3	69.3	67.2	72.0	52.2	34.2	41.3	65.8	62.0	74.5	67.7	70.1
TransZero [8]	69.3	68.3	68.8	76.8	52.6	33.4	40.8	65.6	61.3	82.3	70.2	70.1
ours	65.4	79.7	71.8	77.3	51.0	36.4	42.5	67.9	58.9	88.0	70.5	67.4

Table 1: Results (%) of the state-of-the-art CZSL and GZSL modes on CUB, SUN and AWA2, including generative methods and non-generative methods. The best, second-best and third-best results are marked in **Red**, **Blue** and **Green**, respectively. The symbol “–” indicates no results.

Methods	CUB				SUN				AWA2			
	GZSL			CZSL	GZSL			CZSL	GZSL			CZSL
	U	S	H	acc	U	S	H	acc	U	S	H	acc
single predictor	54.3	72.2	62.0	65.0	43.3	22.8	29.8	53.6	56.8	76.4	65.2	60.2
two predictors	60.8	80.3	69.2	76.0	39.7	32.2	35.6	58.5	55.5	87.3	67.8	62.7
two predictors + bias learner	63.2	81.2	71.1	76.7	38.8	33.2	35.7	59.3	55.1	88.9	68.0	63.4

Table 2: Ablation study of different modules. ‘single predictor’ uses a shared predictor, and ‘two predictors’ uses two different predictors for different level features.

Generalized Zero-Shot Learning. Table 1 displays the results of different methods in the GZSL setting. Our method attained the best performance on CUB and the second-best performance on both SUN and AWA2 when compared to other approaches. In the AWA2 dataset, we only differed from the second-best approach by a margin of 0.1%. On the other hand, the generative methods outperform non-generative methods on the SUN dataset. The reason is that SUN has a large number of categories and uneven image distribution, making it difficult for non-generative methods to learn a robust semantic space. Conversely, these generative methods effectively enhance SUN performance by generating images or visual features to complement information. In contrast, our DFAN achieves the best results in the non-generative approaches. These results suggest that our proposed method delivers superior performance in GZSL and CZSL settings.

4.3 Ablation Study

Effect of Two predictors. DFAN employs two different predictors to avoid conflicts between features of different levels. Ablation studies are conducted to validate their effectiveness, and the results are presented in Table 2. Our findings show that using two predictors significantly improves recognition accuracy, particularly in CZSL setting on the CUB dataset, where our method outperforms the single predictor approach by up to 11%. Furthermore, in the GZSL setting, compared to utilizing a single predictor, employing two predictors achieves a significant gain of 10% on average in seen class accuracy while maintaining unseen class accuracy.

Methods	CUB				SUN				AWA2			
	GZSL			CZSL	GZSL			CZSL	GZSL			CZSL
	U	S	H	acc	U	S	H	acc	U	S	H	acc
\mathcal{L}_{cls}	54.3	72.2	62.0	65.0	43.3	22.8	29.8	53.6	56.8	76.4	65.2	60.2
\mathcal{L}_{attr}	51.8	69.4	59.3	61.5	45.3	26.5	33.5	62.0	50.6	88.4	64.4	63.7
$\mathcal{L}_{cls}+\mathcal{L}_{attr}$	60.8	80.3	69.2	76.0	39.7	32.2	35.6	58.5	55.5	87.3	67.8	62.7
$\mathcal{L}_{cls}+\mathcal{L}_{attr}+\mathcal{L}_{cos}$	65.4	79.7	71.8	77.3	51.0	36.4	42.5	67.9	58.9	88.0	70.5	67.4

Table 3: Ablation study of loss terms in Eq. 8. We accumulate each term from top to bottom of the table.

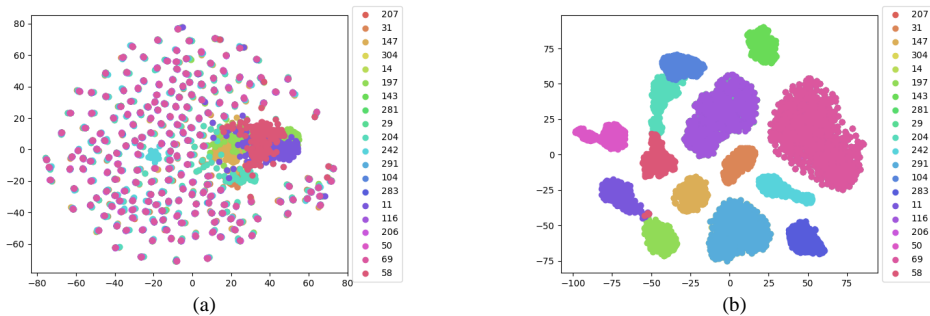


Figure 2: T-SNE visualizations of attribute features and the numbers to the right of each legend represent randomly selected attributes. (a) \mathcal{L}_{cos} is not used. (b) \mathcal{L}_{cos} is used.

Effect of Bias Learner. The bias learner effectively corrects attribute prediction differences by learning offsets similar to residual connections. Its performance is evaluated in Table 2. In the CZSL setting, the bias learner improves seen class accuracy by an average of 1% compared to using two predictors. In the GZSL setting, our module improves harmonic mean by 1.9% on CUB, where the dataset is exclusively composed of bird species, and the distinction between classes relies mainly on subtle differences between attributes. In contrast, we did not achieve significant improvements on the AWA2 and SUN datasets due to their diverse categories and irregular attribute distributions, which makes it difficult for fully connected layers to accurately capture the relationship between attributes.

Effect of Loss Function. In order to investigate the effectiveness of each of our loss functions for attribute representation learning, we conduct an ablation study by adding losses one by one, starting from the baseline where no attribute features are included in the framework, denoted as \mathcal{L}_{cls} . The results are presented in Table 3, where it is evident that incorporating \mathcal{L}_{attr} and \mathcal{L}_{cos} both improve the performance and their effects are complementary. Notably, the cosine similarity loss \mathcal{L}_{cos} has the greatest impact on the overall performance improvement.

t-SNE Visualizations. In order to assess the representativeness of our proposed attribute features, we conducted a t-SNE visualization analysis on 20 randomly selected attribute features extracted from multiple images, with and without \mathcal{L}_{cos} . As depicted in Figure 2, the attribute features learned with \mathcal{L}_{cos} are distinctly clustered based on the corresponding attributes and exhibit a clear boundary, whereas those without \mathcal{L}_{cos} are relatively mixed. This

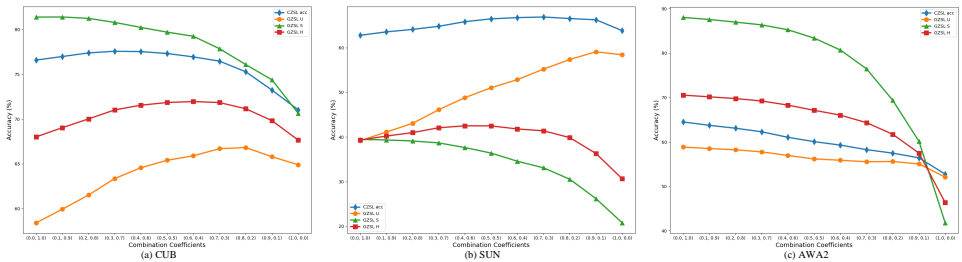


Figure 3: Effect of coefficients (β_1, β_2) in the CZSL and GZSL setting.

observation suggests that incorporating \mathcal{L}_{cos} in our DFAN effectively enhances the discrimination ability of attribute-level features, thereby promoting effective attribute representation learning.

4.4 Discussion on β_1 and β_2

We evaluate the impact of the ratio β on the final prediction results, which controls the relative contribution of attribute-level features and class-level features in the proposed model. The results are presented in Figure 3. In fine-grained datasets, such as CUB and SUN, the CZSL accuracy and the harmonic mean increase first and then decrease as the attribute features play a more dominant role in these datasets. Our DFAN performs best with coefficients (0.3,0.7) on CUB and (0.7,0.3) on SUN in CZSL. In GZSL, our method’s peak performance is at (0.6, 0.4) for CUB, while SUN reaches its best result at (0.5, 0.5). In contrast, for the coarse-grained dataset AWA2, as the value of β_1 gradually increases, the CZSL accuracy and the harmonic mean decrease and reach a minimum when β_1 equals 1. Our DFAN performs best in CZSL and GZSL when the coefficient is (0.0, 1.0). This outcome can be attributed to the coarse-grained dataset’s attribute descriptions, which encompass the entirety of the image, leading to the dominant influence of class-level features on the ultimate prediction.

5 Conclusion

In conclusion, our proposed DFAN effectively addresses the limitations of existing methods in GZSL. The visual feature augmentation module enhances attribute feature extraction and disentangles them through a cosine similarity loss. This module enables better capturing of attribute-specific visual features. With the aid of a bias learner, the semantic feature augmentation module effectively handles discrepancies between predicted and actual attribute values, leading to improved shared semantic features for each sample. Moreover, DFAN utilizes two mapping functions to bridge the visual and semantic features, facilitating effective fusion and linkage of information at different levels. The results validate the effectiveness of DFAN’s visual and semantic feature augmentation modules and the utilization of two mapping functions.

Acknowledgement

This project is supported by International Exchanges 2022 IEC\NSFC\223523 and Securing the Energy/Transport Interface EP/X037401/1.

References

- [1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(7):1425–1438, 2015.
- [2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.
- [3] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *European conference on computer vision*, pages 52–68. Springer, 2016.
- [4] Long Chen, Hanwang Zhang, Jun Xiao, Wei Liu, and Shih-Fu Chang. Zero-shot visual recognition using semantics-preserving adversarial embedding networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1043–1052, 2018.
- [5] Shiming Chen, Wenjie Wang, Beihao Xia, Qinmu Peng, Xinge You, Feng Zheng, and Ling Shao. Free: Feature refinement for generalized zero-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 122–131, 2021.
- [6] Shiming Chen, Ziming Hong, Yang Liu, Guo-Sen Xie, Baigui Sun, Hao Li, Qinmu Peng, Ke Lu, and Xinge You. Transzero: Attribute-guided transformer for zero-shot learning. In *AAAI*, volume 2, page 3, 2022.
- [7] Shiming Chen, Ziming Hong, Guo-Sen Xie, Wenhan Yang, Qinmu Peng, Kai Wang, Jian Zhao, and Xinge You. Msdn: Mutually semantic distillation network for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7612–7621, 2022.
- [8] Yu Du, Miaojing Shi, Fangyun Wei, and Guoqi Li. Boosting zero-shot learning via contrastive optimization of attribute representations. *arXiv preprint arXiv:2207.03824*, 2022.
- [9] Haoran Duan, Yang Long, Shidong Wang, Haofeng Zhang, Chris G Willcocks, and Ling Shao. Dynamic unary convolution in transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
- [10] Rui Gao, Fan Wan, Daniel Organisciak, Jiyao Pu, Haoran Duan, Peng Zhang, Xingsong Hou, and Yang Long. Privacy-enhanced zero-shot learning via data-free knowledge transfer. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 432–437. IEEE, 2023.

- [11] Jiannan Ge, Hongtao Xie, Shaobo Min, and Yongdong Zhang. Semantic-guided reinforced region embedding for generalized zero-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1406–1414, 2021.
- [12] Jiannan Ge, Hongtao Xie, Shaobo Min, Pandeng Li, and Yongdong Zhang. Dual part discovery network for zero-shot learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3244–3252, 2022.
- [13] Zongyan Han, Zhenyong Fu, Shuo Chen, and Jian Yang. Contrastive embedding for generalized zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2371–2381, 2021.
- [14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [15] Dat Huynh and Ehsan Elhamifar. Compositional zero-shot learning via fine-grained dense feature composition. *Advances in Neural Information Processing Systems*, 33: 19849–19860, 2020.
- [16] Dat Huynh and Ehsan Elhamifar. Fine-grained generalized zero-shot learning via dense attribute-based attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4483–4493, 2020.
- [17] Dinesh Jayaraman and Kristen Grauman. Zero-shot recognition with unreliable attributes. *Advances in neural information processing systems*, 27, 2014.
- [18] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [19] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE conference on computer vision and pattern recognition*, pages 951–958. IEEE, 2009.
- [20] Kai Li, Martin Renqiang Min, and Yun Fu. Rethinking zero-shot learning: A conditional visual classification perspective. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3583–3592, 2019.
- [21] Yan Li, Junge Zhang, Jianguo Zhang, and Kaiqi Huang. Discriminative learning of latent features for zero-shot recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7463–7471, 2018.
- [22] Yang Liu, Lei Zhou, Xiao Bai, Yifei Huang, Lin Gu, Jun Zhou, and Tatsuya Harada. Goal-oriented gaze estimation for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3794–3803, 2021.
- [23] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
- [24] Shaobo Min, Hantao Yao, Hongtao Xie, Chaoqun Wang, Zheng-Jun Zha, and Yongdong Zhang. Domain-aware visual bias eliminating for generalized zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12664–12673, 2020.

- [25] Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. Zero-shot learning with semantic output codes. *Advances in neural information processing systems*, 22, 2009.
- [26] Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2751–2758. IEEE, 2012.
- [27] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.
- [28] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 49–58, 2016.
- [29] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *International conference on machine learning*, pages 2152–2161. PMLR, 2015.
- [30] Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero-and few-shot learning via aligned variational autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8247–8255, 2019.
- [31] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. *Advances in neural information processing systems*, 26, 2013.
- [32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [33] Vinay Kumar Verma, Gundeep Arora, Ashish Mishra, and Piyush Rai. Generalized zero-shot learning via synthesized examples. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4281–4289, 2018.
- [34] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.
- [35] Chaqun Wang, Shaobo Min, Xuejin Chen, Xiaoyan Sun, and Houqiang Li. Dual progressive prototype network for generalized zero-shot learning. *Advances in Neural Information Processing Systems*, 34:2936–2948, 2021.
- [36] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010.
- [37] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4582–4591, 2017.

- [38] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018.
- [39] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5542–5551, 2018.
- [40] Yongqin Xian, Saurabh Sharma, Bernt Schiele, and Zeynep Akata. f-vaegan-d2: A feature generating framework for any-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10275–10284, 2019.
- [41] Guo-Sen Xie, Li Liu, Xiaobo Jin, Fan Zhu, Zheng Zhang, Jie Qin, Yazhou Yao, and Ling Shao. Attentive region embedding network for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9384–9393, 2019.
- [42] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. Attribute prototype network for zero-shot learning. *Advances in Neural Information Processing Systems*, 33:21969–21980, 2020.
- [43] Yunlong Yu, Zhong Ji, Jungong Han, and Zhongfei Zhang. Episode-based prototype generating network for zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14035–14044, 2020.
- [44] Zhongqi Yue, Tan Wang, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. Counterfactual zero-shot and open-set visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15404–15414, 2021.
- [45] Ke Zhu and Jianxin Wu. Residual attention: A simple but effective method for multi-label recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 184–193, 2021.