

Multi-Granularity Contrastive Zero-shot Learning Model Based on Attribute Decomposition

^a*School of Computer and Information Technology, Shanxi University, Taiyuan 030006, Shanxi, China*

^b*Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan 030006, Shanxi, China*

ARTICLE INFO

Keywords:

Zero-shot learning
Contrastive learning
Pre-trained language model
Global-local paradigm
Deep learning

ABSTRACT

Zero-shot learning aims to identify new classes by transferring semantic knowledge from seen classes to unseen classes. Image classification based on zero-shot learning usually regards different attributes as the carriers of semantic knowledge, for different attributes have different effects on the task. However, existing models lack a differentiated understanding of different attributes and ignore the impact of global context information. Therefore, this paper proposes a multi-granularity contrastive zero-shot learning model based on attribute decomposition. Specifically, different attributes are initially divided into key and common attributes i.e., attribute decomposition. Then, inspired by Navon's global-local paradigm, we work out the multi-granularity contrastive learning model, which is composed of the global learning module and the local one, to further enhance the interaction between the global and local information. Finally, zero-shot image classification is achieved by training a multi-granularity contrastive learning model. The method is experimented on three public zero-shot learning benchmark data sets (i.e., AWA2, CUB and SUN). Compared with the existing model, in the CUB dataset, the harmonic mean H is improved by 1.6%; in the SUN dataset, H increased by 7.1%; in the AWA2 dataset, H increased by 2.5%, further verifying the effectiveness of this model.

1. Introduction

ZSL (Zero-Shot Learning) (Farhadi et al., 2009), inspired by human perception, aims to reason and learn through prior knowledge. The early ZSL methods (Frome et al., 2013) generalize unseen classes more efficiently by introducing the concept of semantic embedding, i.e. embedding image and text data into a shared semantic space and measuring their similarities in the semantic space. The Generative ZSL methods (Xian et al., 2018) reduce the bias between unseen and seen classes by utilizing generative models to synthesize samples of unseen classes, extend the training data and transform the ZSL problem into a supervised learning problem.

Different from the current ZSL methods which identify the attributes in the seen class and the corresponding regions and map them to the attributes and regions of the unseen class using different attributes of samples to understand the commonalities and differences between classes, the Attention-based ZSL methods (Min et al., 2020) employ attention mechanism to distinguish discriminative regions of image attributes and enhance the relevant information of the classes by assigning attention weights to different regions or features in a given image. Afterwards, Chen et al. (2022a) propose an attribute-guided network to learn the embedding representations of discriminant visual semantic attributes, and Chen et al. (2023) use a cross-modal mask reconstruction strategy to enhance the model's ability to discriminate between fine-grained visual features in seen and unseen images by transferring prior semantic knowledge from language models and aligning it with the feature-aware ability of the visual transformer model. By incorporating the key information implied by the local regions of an image, these attention-based methods capture the subtle visual differences between images and accurately locate the regions corresponding to different attributes. However, they rely mainly on attention mechanism to discover more discriminating attributes (i.e., key attributes), and focus excessively on local features and ignore global contextual information. If the key attributes in an image are missing or inaccurately recognized, the performance will be less satisfactory.

Motivated by human reasoning and learning, Navon (1977) proposed the global-local paradigm (i.e., humans will first perceive the global information of a given object and then the local part). Specifically, when an observer consciously focuses on global features, local features do not disturb the perception, but when the observer concentrates the local features, the global features exert the influence on the perception of its local features. As shown in Fig. 1, the left figures illustrate the concepts of different features: the global feature is expressed by most common local attributes (such as the global pentagonal star is composed of local squares), and the key attributes usually further illustrate the

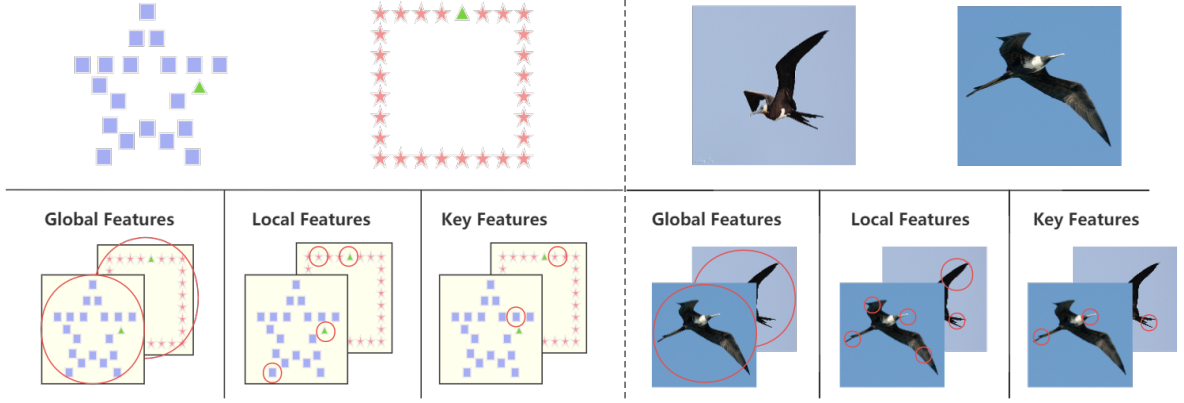


Fig. 1. The global-local paradigm illustration.

accurate class (see the squares with a triangle); the right figures explain the global-local principle in identification: the observer first perceives the object as a bird by identifying the global features through the common attributes such as wings, tail, and head, and then ascertains it as a "Frigate bird" with the key attributes like shape, beak, and the color of its tail.

Therefore, stimulated by the aforementioned global-local paradigm, and also to solve the problem, we propose a multi-granularity contrastive zero-shot learning model based on attribute decomposition. Specifically, attributes are initially categorized into key and common attributes: key attributes are the attributes that are given higher weight in the attention mechanism, and remaining attributes are common attributes. Next, the global and local learning modules are set up with different attributes to construct a multi-granularity contrastive learning module. The global contrastive learning module adopts most of the common attributes to learn the presentation of global features; the local contrastive learning module employs key attributes to learn the representation of local features. This is conducive for the model to balance the attention to the local and the global information as well as the interactive influence between the two types of information, and to capture the image features more comprehensively. The contributions are summarized as follows:

- Propose the attribute decomposition module, which decomposes attributes into key and common attributes to alleviate the problem of over-reliance on key attributes.
- Introduce the global-local paradigm into the zero-shot image classification task, design a multi-granularity contrastive learning model, and further enhance the model's ability to distinguish multi-granularity visual features in seen and unseen images.
- Validate the effectiveness and interpretation of our model through experiments on three zero-shot learning benchmark datasets (i.e., AWA2, CUB, and SUN).

2. Related Work

Zero-Shot Learning (ZSL) proposed by Lampert et al. (2009) helps simulate human reasoning and identify new classes never seen before. This study focuses on image classification in ZSL, and its focus is to enable the model to effectively classify the unseen classes utilizing the semantic information and feature knowledge learned from the seen classes (Chen et al., 2021a).

The early embedding-based ZSL method maps visual features into a shared semantic space to achieve the interaction between vision and semantics. Zhang et al. (2017) consider visual space as an embedded space to reduce the impact of the hub problem. However, the prerequisite of such method is that the data distribution of the seen classes is similar to that of the unseen classes, but problems such as incomplete data collection in practice cause differences in distribution which further limits the model's ability to transfer knowledge.

To alleviate such problems, some researchers propose generation-based ZSL, like Variational Auto-Encoders (VAEs) (Chen et al., 2021c), Generative Adversarial Networks (GANs) (Chen et al., 2021b), and Generative

Flows (Shen et al., 2020). Such generation-based strategies help compensate for the lack of unseen class data and convert ZSL tasks into a supervised classification task. Naeem et al. (2022) utilize the VAEs to generate simple attribute values; Geng et al. (2021) employ the GANs to generate knowledge graphs; Chen & Yeh (2021) apply the generative flows to generate text descriptions. However, most generation-based ZSL methods lack stable end-to-end structure and are difficult to converge on during training (Pourpanah et al., 2022). Plus, over-reliance on visual features leads to difficulties in accurately locating the regions corresponding to different attribute features.

Recently, Attention-based ZSL (Xie et al., 2019) integrates the discriminative information implied by local regions of the image through introducing attention mechanism into ZSL tasks. Xu et al. (2020) regard semantic information as a guide to learn more discriminative regional features. Min et al. (2020) compact semantic-free representation by cross-attention. Xie et al. (2020) construct region maps through attention mechanism, conducted regional reasoning using graph neural networks, and thus accomplished the knowledge transfer between different classes. Liu et al. (2021) adopt gaze estimation to improve localization information for discriminative attributes. Xu et al. (2021) utilize low-rank semantics grouping to learned superior visual semantic encoders. Chen et al. (2022b) use a mutually semantic distillation network to refine intrinsic semantic representation between visual and semantic features. Although the attention-based ZSL is guided by semantic information, it still lacks a deep understanding of semantic information, that is, it does not learn the potential semantic information of text features.

Recently, Pre-trained Language Models (PLMs) (Min et al., 2023) has been introduced into the attention-based ZSL to address the above issues. PLMs implement knowledge representation by unsupervised learning on large-scale corpora. Zhao et al. (2023) utilize PLMs original word embeddings to find words related to class names automatically. Chen et al. (2023) integrates latent semantic knowledge from PLMs. While PLMs help learn multi-level semantic knowledge of the text when used to generate textual features, problems with the attention-based ZSL still exist.

Therefore, our multi-granularity contrastive learning model introduces the global-local paradigm into the zero-shot image classification task to balance the model's attention to global and local information and enable the model to capture image features more comprehensively.

3. Methodology

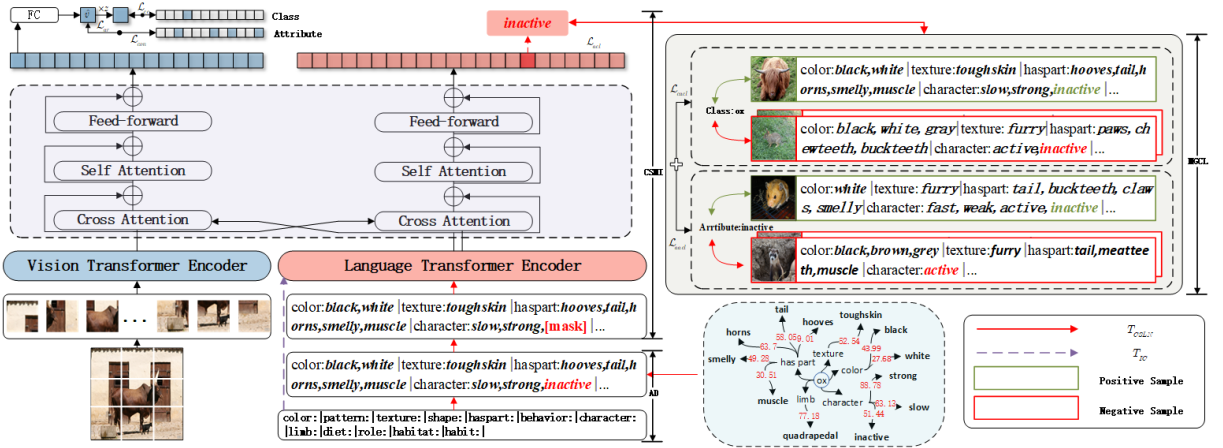


Fig. 2. Multi-granularity contrastive zero-shot learning model based on attribute decomposition.

3.1. Task Definition

ZSL aims to identify images of unseen classes through transferring the knowledge learning from seen classes using the attributes of samples. In this study, the class appearing in the train set are defined as seen classes C^s , the class not appearing in the train set are defined as unseen classes C^u , the train set is denoted as $D_s = \{(x^s, c^s) | x^s \in \mathcal{X}^s, c^s \in C^s\}$, where x^s is an image originating from the train set $\mathcal{X}^s$, and c^s is the class corresponding to x^s in the seen class C^s , and the test set is denoted as $D_u = \{(x^u, c^u) | x^u \in \mathcal{X}^u, c^u \in C^u\}$, where x^u is an image originating from the train set $\mathcal{X}^u$, and c^u is the class corresponding to x^u in the unseen class C^u , with C^s and C^u are disjoint. In addition, the classes in the dataset

are described by different attributes, and the attribute set of the dataset is denoted as $\mathcal{A} = \{a_1, a_2, \dots, a_{|\mathcal{A}|}\}$, where $|\mathcal{A}|$ is the number of attributes. The binary representation of the attributes corresponding to class c is $z^c = [z_1^c, z_2^c, \dots, z_{|\mathcal{A}|}^c]^T$, where the value of z_i^c indicates whether the corresponding attribute exists in the class or not (i.e., $z_i^c = 1$ denotes the presence of attribute a_i in class c). The attention weights of the attributes corresponding to class c is $t^c = [t_1^c, t_2^c, \dots, t_{|\mathcal{A}|}^c]^T$, denotes the degree of attention of the attribute in the class. As can be seen in the overall framework of the model in Fig. 2, the model is divided into two tasks:

(1) Image Classification Task

The image classification task is to classify images based on the attributes presented in the image. The definition of this task is shown in Eq. (1):

$$y = FC(I(g(x), h(\mathcal{A}))) \quad (1)$$

where $g(\cdot)$, $h(\cdot)$, $I(\cdot)$ and $FC(\cdot)$ denote the Visual Transformer Encoder, the Attribute Transformer Encoder, the Cross-Modal Encoder, and the Classifier. Among them, x denotes the image to be classified, and $\mathcal{A}$ denotes the set of attributes corresponding to the image x .

(2) Attribute Mask Prediction Task

The attribute mask prediction, similar to mask prediction in PLMs, involves masking certain attribute in the set of input attributes to enable the model to represent attributes. The masked attributes a_{mask} are shown in Eq. (2):

$$a_{mask} = I(g(x), h(\mathcal{A}_{[mask]})) \quad (2)$$

where $\mathcal{A}_{[mask]}$ denotes the attribute set containing the [mask] token.

3.2. Classification Models Based on Multi-granularity Contrastive Learning

In this part, Through multi-task learning (i.e., image classification and attribute mask prediction), our model accomplishes cross-modal semantic interaction, and obtains the embedded representation of images and their attributes, which was eventually input into zero-shot image classification tasks. The model is divided into Attribute Decomposition (AD), Cross-Modal Semantic Interaction (CMSI), Attribute Mask Prediction (AMP), Image Classification (IC), Multi-Granularity Contrastive Learning (MGCL) and Multi-Task Learning module.

3.2.1. Attribute Decomposition Module

This module decomposes attributes into key and common attributes through AD strategy. First, cluster discrete attributes so that the attributes of each class are unified in text format. Secondly, based on the attention weights of the attributes in each class, select the part of the attribute set with higher attention weights in the corresponding class as a key attribute set, and the remaining part is a common attribute set.

Specifically, each attribute, represented as a $\mathcal{P}$ (attribute name: attribute value) pair with the symbol "|" separating the attributes, composes the final attributes set for a given class $\mathcal{A}^c = \mathcal{P}_1 \cup \mathcal{P}_2 \cup \dots \cup \mathcal{P}_k$, $k \leq |\mathcal{A}|$. For example, the uniform representation of the class "ox" in the AWA2 (Xian et al., 2017) dataset is:

$$\dots | \underbrace{color}_{\text{attribute name}} : \underbrace{black, white}_{\text{attribute value}} | \underbrace{character}_{\text{attribute name}} : \underbrace{slow}_{\text{attribute value}} \dots | \dots$$

For a given class c , the attention weight t^c of the corresponding attributes of the class is first sorted in descending order to obtain the sorted attribute set $t^{c'} = \{t_1^{c'}, t_2^{c'}, \dots, t_{|\mathcal{A}|}^{c'}\}$, and then the first $\alpha \times |\mathcal{A}|$ attributes are selected as the key attribute set $\mathcal{A}^{c'} = \{a_i | t_i^{c'} \geq t_{\alpha}^{c'}, a_i \in \mathcal{A}^c\}$, where the hyperparameter α represents the proportion of the key attributes to the total attributes.

3.2.2. Cross-modal Semantic Interaction Module

This module encodes visual features and textual attributes respectively using the Transformer architecture, and then achieves the interaction and fusion of information between different modalities through a multi-modal encoder.

Specifically, a given image x is first segmented into equally-sized regions, input into an encoder with a positional embedding $g(\cdot)$, and encoded as a series of embedded representations $\{v_{cls}, v_1, \dots, v_n\}$, which v_{cls} is the embedded representation of [CLS]. Meanwhile, the set of attributes $\mathcal{A}_{[mask]}$ that need to be input is obtained via randomly sampling the mask attributes $a_{mask} = RSC(\mathcal{A}^{c'})$ from the set of key attributes $\mathcal{A}^{c'}$. $\mathcal{A}_{[mask]}$ is input into the encoder $h(\cdot)$ to obtain the embedding representation of the text $\{w_1, \dots, w_{mask}, \dots, w_n\}$, where w_{mask} denotes the embedding representation of the mask attribute a_{mask} .

Finally, the output $\{v_{cls}, v_1, \dots, v_n\}$ of encoder $g(\cdot)$ and the output $\{w_1, \dots, w_{mask}, \dots, w_n\}$ of encoder $h(\cdot)$ are input into the cross-modal attention layer encoder $I(\cdot)$, where the encoder $I(\cdot)$ consists of a bidirectional cross-attention module, a self-attention module, and a feed-forward module. Afterwards, the embedding representation of the image $\{v'_{cls}, v'_1, \dots, v'_n\}$ and that of the text $\{w'_1, \dots, w'_{mask}, \dots, w'_n\}$ with different modal information are obtained.

3.2.3. Attribute Mask Prediction Module

After the cross-modal semantic interaction described in 3.2.2, the predicted embedding of the mask attribute a_{mask} is denoted as w'_{mask} . The objective function $\mathcal{L}_{cmr}$ at this stage is shown in Eq. (3):

$$\mathbb{E}_{x \sim \mathcal{X}^s} [-z_{a_{mask}} \log P(a_{mask} | \mathcal{A}_{[mask]}, x)] \quad (3)$$

Where $\mathbb{E}$ and $z_{a_{mask}}$ respectively represent the expectation and the binary representation of the attribute a_{mask} in the class c . In addition, $P(a_{mask} | \mathcal{A}_{[mask]}, x)$ as shown in Eq. (4)

$$P(a_{mask} | \mathcal{A}_{[mask]}, x) = \frac{\exp(w'_{mask} \cdot w_{a_{mask}})}{\sum_{a_{mask}^- \in \mathcal{A}_{[mask]}} \exp(w'_{mask} \cdot w_{a_{mask}^-})} \quad (4)$$

where $w_{a_{mask}}$ denotes the embedding representation of a_{mask} .

3.2.4. Image classification Module

Similar to the literature [Chen et al. \(2023\)](#), this module uses the attribute regression loss function, the cross-entropy loss function, and the class label loss function to optimal the performance of the model.

Firstly, the Attribute Regression Loss $\mathcal{L}_{ar}$ is used to calculate the difference between the attribute vectors predicted by the model $\tilde{v}$ and the attribute vectors of the real class $\hat{v}$, and by minimizing this loss the model accurately transforms the image information into the corresponding attribute representations. The loss function is shown in Eq. (5):

$$\mathcal{L}_{ar} = \mathbb{E}_{x \sim \mathcal{X}^s} \|\tilde{v} - \hat{v}\|_2^2 \quad (5)$$

Secondly, the cross-entropy loss is used to maximize the match between the images and the semantic vectors of the real class labels. The loss function is shown in Eq. (6):

$$\mathcal{L}_{cc} = \mathbb{E}_{x \sim \mathcal{X}^s} [-\log \frac{\exp(\tilde{v} \cdot \hat{v})}{\sum_{\hat{c} \in \mathcal{C}^s} \exp(\tilde{v} \cdot \hat{v}^{\hat{c}})}] \quad (6)$$

where $\hat{c}$ denotes a seen class that excludes the class to which the to-be-classified image belongs. Finally, a supervised class label loss is used to further enhance the model's ability to classify different classes with limited samples, and the loss function is shown in Eq. (7):

$$\mathcal{L}_{con} = \mathbb{E}_{x \sim \mathcal{X}^s} [-\log f_{\theta}(\tilde{v} | \mathcal{A}, x)] \quad (7)$$

where $f_{\theta}(\tilde{v} | \mathcal{A}, x)$ is calculated as shown in Eq. (8):

$$f_{\theta} = \frac{\exp(\text{sim}(\tilde{v}, \tilde{v}^+)/\tau)}{\exp(\text{sim}(\tilde{v}, \tilde{v}^+)/\tau) + \sum_{\tilde{v}' \in \mathcal{N}(\tilde{v})} \exp(\text{sim}(\tilde{v}, \tilde{v}')/\tau)} \quad (8)$$

where τ , $\tilde{v}^+$ and $\tilde{v}$ respectively denote the temperature hyperparameter, positive samples, and the representation of different images of the same class, and $\tilde{v}^-$, $\text{sim}(\cdot)$ and $\mathcal{N}(\tilde{v})$ represent the negative samples, the similarity metric (cosine similarity) between the samples respectively, and the set of negative samples consist of the representations of different classes of images within the same batch.

3.2.5. Multi-granularity Contrastive Learning Module

Based on the global-local paradigm in psychology, we at this stage construct a contrastive learning module that conforms to human perception. Entirely taking into account the effective interactions between global and local information, the module is divided into two parts: global contrastive learning and local contrastive learning, and plays a complementary role in representation learning. Specific contrastive learning is achieved by masking different attributes to construct positive and negative samples.

Global Contrastive Learning: Global contrastive learning learns the similarities and differences between classes by comparing samples from different classes, helps the model better understand and differentiate different classes and improves the module's classification performance. Specifically, positive and negative samples are constructed as follows:

(1) Positive sample construction

Samples from the dataset that first belong to the same class c and then have more than 70% of the same attribute values as the original sample x besides the mask attribute are selected as positive samples.

(2) Negative sample construction

Negative sample selected from the datasets do not belong to class c , and two prerequisites to construct them are as follows: i. The key attributes of the masks are different and attributes value of negative sample and that of the original sample x is more than 70% similar. ii. The key attributes of the masks are the same and attributes value of negative sample and that of the original sample x is less than 30% similar; The module loss function representation is shown in Eq. (9):

$$\mathcal{L}_{caci} = \mathbb{E}_{x \sim X^s} [-\text{Min}(z_a, z_{a+}) \log f_{\phi}(\tilde{\omega}' | \mathcal{A}, x)] \quad (9)$$

where $f_{\phi}(\tilde{\omega}' | \mathcal{A}, x)$ is consistent with the contrastive loss in Eq. (8), z_a, z_a^+ and $\tilde{\omega}'$ denote the binary representation of the original sample attribute a , the binary representation of the positive sample attribute a^+ , and the average pooling representation on the language side of the encoder $I(\cdot)$.

Local Contrastive Learning: Local contrastive learning enables the model to learn subtle attribute differences, which helps better capture the samples' semantic features. When constructing the samples of local contrastive learning, we only need to focus on the differences and similarities of attributes without distinguishing whether they belong to the same class. Positive and negative samples are specifically constructed as follows:

(1) Positive sample construction

The key attributes of the masks are the same and attributes value of negative sample and that of the original sample x is no more than 30% similar;

(2) Negative sample construction

The key attributes of the masks are different and attributes value of negative sample and that of the original sample x is more than 70% similar. The module loss function representation is shown in Eq. (10):

$$\mathcal{L}_{aaci} = \mathbb{E}_{x \sim X^s} [-\text{Min}(z_a, z_{a+}) \log f_{\psi}(\tilde{\omega}' | \mathcal{A}, x)] \quad (10)$$

where $f_{\psi}(\tilde{\omega}' | \mathcal{A}, x)$ is consistent with the contrastive loss in Eq. (8), z_a denotes the binary representation of the original sample attribute a , z_a^+ denotes the binary representation of the positive sample attribute a^+ . The complete multi-granularity contrastive loss function representation is shown in Eq. (11):

$$\mathcal{L}_{acl} = (1 - \mu)\mathcal{L}_{caci} + \mu\mathcal{L}_{aaci} \quad (11)$$

where $1 - \mu$ denotes the probability of an global contrastive loss $\mathcal{L}_{caci}$, and similarly, μ denotes the probability of local contrastive loss $\mathcal{L}_{aaci}$.

3.2.6. Multi-Task Learning Module

Multi-task learning adopts task-switching strategies, forcing the model to spread attribute information between visual and language tasks. That is, the model performs the CSMI by visually patching and $\mathcal{A}_{[mask]}$ at this stage, and accomplishes a simple IC task in the T_{IC} phase.

Table 1

Datasets' statistical information.

Dataset	Semantic Dimension	#Seen Class	#Unseen Class	#Training Image	#Testing Image
AWA2	85	40	10	23527	13795
CUB	312	150	50	7057	4731
SUN	102	645	72	10320	4020

Note that L_{IC} and L_{CSMI} are the loss functions of IC and CSMI tasks respectively, as shown in Eq. (12) and Eq. (13). Here the probability of L_{IC} is denoted as ρ , and similarly, the probability of L_{CSMI} is expressed as $1 - \rho$:

$$L_{IC} = \mathcal{L}_{zsl} + \lambda_{con}\mathcal{L}_{con} \quad (12)$$

$$L_{CSMI} = \mathcal{L}_{zsl} + \lambda_{cmr}\mathcal{L}_{cmr} + \mathcal{L}_{acl} \quad (13)$$

where the ZSL loss $\mathcal{L}_{zsl}$ is denoted by $\mathcal{L}_{zsl} = \mathcal{L}_{cc} + \lambda_{ar}\mathcal{L}_{ar}$.

3.3. Zero-Shot Prediction

After model training, our model performs zero-shot recognition in the learned cosine similarity space. Two evaluation settings are Conventional Zero-Shot Learning (CZSL), which classifies the test samples from C^u , and Generalized Zero-Shot Learning (GZSL), which classifies the test samples from $C^u \cup C^s$. Specifically, the attribute set $\mathcal{A}_{[mask]}$ and the predicted image x are treated as input. Following the setup of Min et al. (2023), the prediction c^* is shown in Eq. (14):

$$c^* = \arg \max_{c \in C^u/C} (\tilde{v} \cdot \hat{v}^c) - \gamma \mathbb{I}[c \in C^s] \quad (14)$$

where $\mathbb{I} = 1$, if c is a seen class and $\mathbb{I} = 0$ otherwise. γ is the calibration factor tuned on a held-out validation set, and $c \in C^u/C$ corresponds to the CZSL/GZSL respectively.

4. Experiment and Analysis

In order to verify the effectiveness of AD as well as MGCL, we conduct four types of experiments, which are the contrast experiments with the mainstream ZSL methods, the ablation experiments, the performance analyses and the interpretability analyses.

4.1. Datasets

In this study, the proposed model is validated on three ZSL benchmark datasets: AWA2 (Xian et al., 2017), CUB (Welinder et al., 2010), and SUN (Patterson & Hays, 2012). Following the splitting rule Xian et al. (2017) which is widely used in current methods, the dataset is divided into train and test sets as shown in Table 1, where AWA2 is composed of animals with similar morphologies but different habits and similar habits but different morphologies, CUB birds of different classes but visually very close, and SUN images of scenes taken from different angles.

4.2. Evaluating Metrics

The Top-1 classification accuracy is adopted as the evaluation metric of CZSL, and is the proportion of correctly classified images on a given test dataset, denoted as T1. The calculation formula is shown in Eq. (15)

$$\text{Top} - 1 = \left(\frac{\text{predict the correct sample}}{\text{total sample count}} \right) \times 100\% \quad (15)$$

GZSL has the data of both seen and unseen classes in the test stage. Top-1 classification accuracy is the evaluation metric for the two classes, S is marked as the accuracy of the seen class, and U is the accuracy of the unseen class. The

Table 2

Datasets' statistical information.

Dataset	λ_{ar}	λ_{con}	λ_{acl}	ρ	γ	μ	α
AWA2	0.01	0.05	0.01	0.4	0.8	0.7	0.5
CUB	0.01	0.05	0.001	0.4	0.7	0.7	0.5
SUN	0.01	0.05	0.01	0.4	0.4	0.5	0.5

calculation formula follows the basic formula of Eq. (15). At the same time, the harmonic mean accuracy (H) is the evaluation metric to comprehensively measure the comprehensive performance of our model in the seen and unseen classes. The formula for calculating the H is shown in Eq. (16):

$$H = \frac{2 \times S \times U}{S + U} \quad (16)$$

4.3. Implementation Details

In this study, the experiments run on a server with Ubuntu 23.04 operating system and NVIDIA 4090Ti graphics card. Python language and Pytorch 1.9.0 deep learning framework are used in Experiments. The original image and attributes are input, and after encoding them using transformer architecture, cross-modal attention is employed to transfer the knowledge of the two features interaction. The various parameters used in the dataset and their specific values are shown in Table 2, where λ_{ar} , λ_{con} , λ_{acl} , ρ , γ , μ and α respectively denote attribute regression loss weight, class label loss weight, loss weight of MGCL, mask ratio, correction overlay weight, the ratio of local contrastive learning to global contrastive learning, and the ratio of key attribute's to all attributes.

4.4. Experimental Results Compared with Mainstream Methods

We compare our model with the representative or state-of-the-art (SOTA) methods proposed over the past four years on three ZSL datasets. The comparison methods include two categories: non-generative formula (CE-GZSL (Xian et al., 2018), GCM-CF (Yue et al., 2021), DCRGAN (Radford et al., 2016), DisVAE (Li et al., 2021), BSeGN (Xie et al., 2022)) and generative methods (DVBE (Min et al., 2020), TransZero (Chen et al., 2022a), DUET (Chen et al., 2023), APN (Xu et al., 2020), RGEN (Xie et al., 2020), GEM-ZSL (Liu et al., 2021), LSG (Xu et al., 2021), MSDN (Chen et al., 2022b)).

The experimental results are shown in Table 3. Underlines mark the best performance of the comparison method or bold when we obtain SOTA. The experimental results show that in the CUB dataset, our model performs well, surpassing all the generative and non-generative methods in the H metric and performing the above methods in the S metric. In the SUN dataset, our model achieves accuracy (51.5%), outperforming existing representative methods, and exceed DUET by 7.1% on the H metric and is better than all representative methods on the T1 metric because our model increases the importance of key attributes in classification. In the AWA2 dataset, our model improves its performance by 2.1% compared to the SOTA method and outperforms DUET on all metrics. This study argues that the attributes in the CUB dataset are all more critical, and the AMP strategy hardly works on this dataset, leading to poor results on CZSL. In contrast to GEM-ZSL which uses a goal-oriented human attention estimation module, the attention mechanism of our model does not fully mimic human attention, and gives poor performance.

In the three datasets, our model also achieves better performance on the S and H metrics, with respective improvements of 1.4%, 7.6%, and 2.9% on the S metric over CUB, SUN, and AWA2, and with respective improvements of 1.6%, 7.1%, and 2.1% on the H metric over CUB, SUN, and AWA2, when compared to the latest model DUET. The experiment demonstrates that our model can retain the good performance of predicting the seen class while classifying the unseen class.

4.5. Ablation Research

In order to evaluate the performance of each module in the model and the complementary advantages among the modules, the (S, U, H) performance gains brought by different modules are verified on the AWA2 dataset. The specific methods are as follows: (1) removing the AD module, (2) removing the cross-modal semantic interaction module

Table 3

The results of our method compare with the performance of current zero-shot learning methods. † and ◇ represent the generative and non-generative methods.

	Method	CUB				SUN				AWA2			
		CZSL	GZSL			CZSL	GZSL			CZSL	GZSL		
		T1	U	S	H	T1	U	S	H	T1	U	S	H
†	CE-GZSL(CVPR)(2021)	77.5	63.1	66.8	65.3	63.3	48.8	38.6	43.1	70.4	63.1	78.6	70.0
	GCM-CF(CVPR)(2021)	-	61.0	59.7	60.3	-	47.9	37.8	42.2	-	60.4	75.1	67.0
	DCRGAN(TMM)(2021)	61.0	55.8	66.8	60.8	63.7	47.1	38.5	42.4	-	-	-	-
	DisVAE(AAAI)(2021)	-	51.1	58.2	54.4	-	36.6	<u>47.6</u>	41.4	-	56.9	80.2	66.6
	HSVA(NeurIP)(2021)	62.8	52.7	58.3	55.3	63.8	48.6	39.0	43.3	-	56.7	79.8	66.3
	BSeGN(TNNLS)(2022)	65.3	55.3	60.8	58.0	<u>66.4</u>	48.9	38.3	42.9	71.5	59.3	78.0	67.4
◇	APN(NeurIPS)(2020)	72.0	65.3	69.3	67.2	61.6	41.9	34.0	37.6	68.4	57.1	72.4	63.9
	DVBE(CVPR)(2020b)	-	53.2	60.2	56.5	-	45.0	37.2	40.7	-	63.6	70.8	67.0
	RGENT(EVVC)(2020)	76.1	60.0	<u>73.5</u>	66.1	63.8	44.0	31.7	36.8	<u>73.6</u>	<u>67.1</u>	76.5	71.5
	GEM-ZSL(CVPR)(2021a)	<u>77.8</u>	64.8	69.3	67.2	62.8	38.1	35.7	36.9	67.3	64.8	77.5	70.6
	LSG(TIP)(2021)	52.9	49.6	50.4	50.0	53.4	<u>52.8</u>	23.1	32.2	61.1	60.4	<u>84.9</u>	70.6
	MSDN(CVPR)(2022b)	76.1	68.7	67.5	68.1	65.8	52.2	34.2	41.3	70.1	62.0	74.5	67.7
	TransZero(AAAI)(2022a)	76.8	<u>69.3</u>	68.3	<u>68.8</u>	65.6	52.6	33.4	40.8	70.1	61.3	82.3	70.2
	DUET(TIP)(2023)	72.3	62.9	72.8	67.5	64.4	45.7	<u>45.8</u>	45.8	69.9	63.7	84.7	<u>72.7</u>
	Ours	70.3	64.6	74.2	69.1	69.8	50.8	53.4	52.1	72.1	67.3	85.1	75.2

Table 4

Results of the ablation experiments.

Methods	U(%)	S(%)	H(%)	$\Delta H(\%)$
(1) $Ours_{w/o AD}$	62.7	85.5	72.3	2.9↓
(2) $Ours_{w/o MGCL}$	64.8	83.6	73.1	2.1↓
(3) $Ours_{w/o Laacl}$	65.3	84.7	73.8	1.4↓
(4) $Ours_{w/o Lcacl}$	64.6	86.0	73.7	1.5↓
Ours (Full model)	67.3	85.1	75.2	-

after contrastive learning, (3) removing the local-level contrastive learning module, and (4) removing the global-level contrastive learning module.

In Table 4, ΔH represents the percentage of the performance decrease of the H metric compared to the overall model. Table 4 shows the performance of the U metric steadily improves after the continuous integration of each module, indicating that each module can help the model to learn a better representation. When adding the AD module, the performance of the S metric decreases, and the model mistakenly regards some vital information as redundant information. When (1) deletes the AD module, the sharp decline of the performance indicates that the AD strategy plays a vital role in the model. The introduction of local(3) and global contrastive learning(4) further improves the performance over (2) by 0.7% and 0.6% respectively. After combining the two contrastive learning, the performance of the overall model is improved by 2.1%. This suggests that both modules are beneficial and that contrastive learning can help the model learn better visual representations by narrowing within-class distances in the latent space. The experimental results show global contrastive learning is equally essential as local contrastive learning, which conforms to the global-local paradigm in psychology.

4.6. Performance Analyses

In this study, the parameters that affect the classification results include the probability of local contrastive learning and global contrastive learning μ , and the proportion of key attributes to all attributes α . Experiments are carried out in three ZSL datasets to verify the effect of and on the classification results, μ and α evaluation metrics use the harmonic

mean accuracy (H). Where the range of values as $\mu = \{0, 0.2, 0.4, 0.6, 0.7, 0.8, 1.0\}$ and the range of values as $\alpha = \{0, 0.2, 0.4, 0.5, 0.6, 0.8, 1.0\}$, and the results of the experiment are shown in Fig. 3.

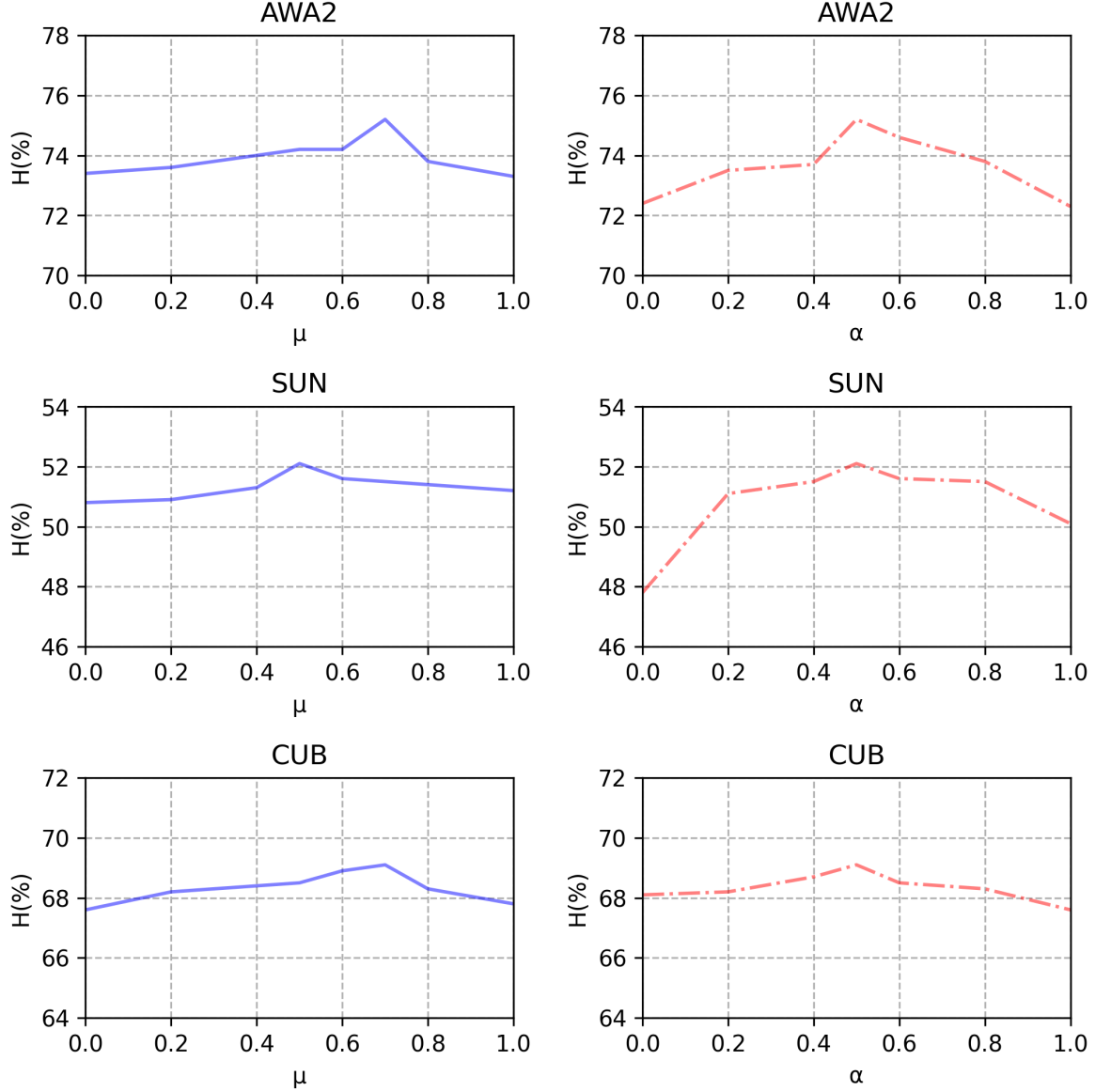


Fig. 3. Parameter analysis diagram.

This paper draws the following conclusions: (1) When μ is too large or too small, the performance is degraded because when $\mu = 1$ represents using only global contrastive learning, focusing on the global information, and when $\mu = 0$ represents using only local contrastive learning, focusing on the local information. The result reveals that the global and local information cannot be fused effectively; (2) When α is nearly 1, the key attribute is close to 100% of all attributes, which is equivalent to a random sampling of all attributes, i.e., this AD module loses its effect, resulting in a sharp decline in model performance. When α is nearly 0, the key attribute is close to 0% of all attributes, and because the masked attributes does not exist in some classes positive or negative samples cannot be constructed. In other words, the contrastive learning module loses its effect, resulting in a sharp decline in model performance; (3) After experiments, we find that the best results are obtained in the CUB dataset when $\mu = 0.6$ and $\alpha = 0.4$, in the SUN and AWA2 datasets when $\mu = 0.7$, and when $\alpha = 0.5$.

Table 5

Results of the ablation experiments.

Image Sample		Attribute			Mistake Example
Zebrn(S) 	Baseline	<i>stripes</i> (70.01%) black(2.03%)	white(7.57%) grazer(1.72%)	oldworld(3.85%)	
	Ours	stripes(47.93%) black(6.42%)	white(12.43%) grazer(4.38%)	oldworld(7.85%)	
german+shepherd(S) 	Baseline	<i>domestic</i> (32.18%) patches(7.83%)	<i>meatteeth</i> (11.55%) fierce(8.50%)	fierce(8.50%) muscle(3.49%)	
	Ours	domestic(19.74%) patches(12.89%)	meatteeth(7.64%) agility(12.33%)	fierce(10.41%) muscle(7.55%)	
Bobcat(U) 	Baseline	<i>stalker</i> (14.69%) oldworld(4.91%)	<i>jungle</i> (14.09%) agility(3.08%)	oldworld(3.85%)	
	Ours	stalker(18.61%) oldworld(8.54%)	jungle(18.20%) agility(7.81%)	meat(8.84%)	
Bat(U) 	Baseline	<i>lean</i> (39.65%) agility(5.31%)	<i>small</i> (25.87%) spots(1.04%)	gray(7.51%)	
	Ours	lean(17.69%) agility(10.88%)	small(18.05%) spots(4.36%)	gray(27.41%)	

Our method compares with the performance of current zero-point learning methods in Fig. 3. This paper draws the following conclusions: (1) When μ is too large or too small, the performance is degraded because when $\mu = 1$ represents using only global contrastive learning, focusing on the global information, and when $\mu = 0$ represents using only local contrastive learning, focusing on the local information. The result is that the global and local information cannot be fused effectively. (2) When α is nearly 1, the key attribute is close to 100% of all attributes, equivalent to a random sampling of all attributes, i.e., this AD module loses its effect, resulting in a sharp decline in model performance. When α is nearly 0, the key attribute is close to 0% of all attributes, and attributes for which no mask exists in some classes cannot be constructed as positive or negative samples. The contrastive learning module loses its effect, resulting in a sharp decline in model performance. (3) After experiments, it is found that the best results are obtained in the CUB dataset when $\mu = 0.6$ and $\alpha = 0.4$, in the SUN and AWA2 datasets when $\mu = 0.7$, and when α is close to 0.5.

4.7. Interpretation Study

This part demonstrates that incorporating AD and global-local contrastive strategies can mitigate the over-reliance of key attributes. The model's predicted results of the attributes of the seen and unseen classes in the AWA2 dataset is shown in Table 5, where the percentage of attributes denotes the importance score of the attribute and the italics denote the key attributes. Table 5 shows that compared with the Baseline (DUET model), the role of key attributes in our model is weakening. In contrast, the importance of all other common attributes increases, which verifies that the global-local paradigm can effectively utilize the global context information to alleviate the problem of over-focusing on local key attributes. For example, in the "Zebra" image, the baseline model gives higher importance to the "stripes" attribute, which leads to the misclassification of the image into the "tiger" class. Although both have the "stripes" attribute, their other attributes differ. Through the global-local contrastive strategy, the importance of common attributes such as "white" and "oldworld" is enhanced, and the model can accurately identify the image as "Zebra".

In addition, in Baseline, the attribute importance is sometimes incorrectly predicted, such as in the example image of "german+shepherd", the key attribute "meatteeth" does not exist. However, it is given a high importance score. The strategy proposed in this study can effectively alleviate the problem of incorrect recognition of key attributes.

5. Conclusion

For the zero-shot image classification task, this study focuses on how to effectively retain and make full use of the overall information, as well as reduce the model's over-reliance on key attributes while focusing on fine-grained features. Based on the global-local paradigm, we propose an Attribute Decomposition strategy and an Multi-Granularity Contrastive Learning strategy in this study. The experiments show that our model effectively alleviates the problem of obscure key attributes of pictures in the zero-shot image classification task, and enhances the model's ability to differentiate multi-granularity visual features in both seen and unseen images. Extensive ablation experiments and analyses on three ZSL benchmarks demonstrate our model significantly outperforms existing state-of-the-art methods which further prove our model's validity and interpretability.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- Chen, J., Geng, Y., Chen, Z., Horrocks, I., Z. Pan, J., & Chen, H. (2021a). Knowledge-aware zero-shot learning: Survey and perspective. In *Proceedings of the 30th Thirtieth International Joint Conference on Artificial Intelligence (IJCAI '21)* (pp. 4366–4373). <https://doi.org/10.24963/ijcai.2021/597>.
- Chen, S., Hong, Z., Liu, Y., Xie, G.-S., Sun, B., Li, H., Peng, Q., Lu, K., & You, X. (2022a). Transzero: Attribute-guided transformer for zero-shot learning. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI '22)*, 36, 330–338. <https://doi.org/10.1609/aaai.v36i1.19909>.
- Chen, S., Hong, Z., Xie, G.-S., Yang, W., Peng, Q., Wang, K., Zhao, J., & You, X. (2022b). Msdn: Mutually semantic distillation network for zero-shot learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR '22)* (pp. 7612–7621). <https://doi.org/10.1109/cvpr52688.2022.00746>.
- Chen, S., Wang, W., Xia, B., Peng, Q., You, X., Zheng, F., & Shao, L. (2021b). Free: Feature refinement for generalized zero-shot learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV '21)* (pp. 122–131). <https://doi.org/10.1109/iccv48922.2021.00019>.
- Chen, S., Xie, G., Liu, Y., Peng, Q., Sun, B., Li, H., You, X., & Shao, L. (2021c). Hsva: Hierarchical semantic-visual adaptation for zero-shot learning. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NIPS '21)*, 34, 16622–16634. <https://doi.org/10.48550/arXiv.2109.15163>.
- Chen, Y.-H., & Yeh, M.-C. (2021). Text-enhanced attribute-based attention for generalized zero-shot fine-grained image classification. In *Proceedings of the 2021 International Conference on Multimedia Retrieval (ICMR '21)* (pp. 447–450). <https://doi.org/10.1145/3460426.3463592>.
- Chen, Z., Huang, Y., Chen, J., Geng, Y., Zhang, W., Fang, Y., Z. Pan, J., & Chen, H. (2023). Duet: Cross-modal semantic grounding for contrastive zero-shot learning. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI '23)*, 37, 405–413. <https://doi.org/10.1609/aaai.v37i1.25114>.
- Farhadi, A., Endres, I., Hoiem, D., & Forsyth, D. (2009). Describing objects by their attributes. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '09)* (pp. 1778–1785). <https://doi.org/10.1109/CVPR.2009.5206772>.
- Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., & Mikolov, T. (2013). Devise: A deep visual-semantic embedding model. In *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS '13)*, 26, 2121–2129. <https://dl.acm.org/doi/10.5555/2999792.2999849>.
- Geng, Y., Chen, J., Chen, Z., Pan, J. Z., Ye, Z., Yuan, Z., Jia, Y., & Chen, H. (2021). Ontozsl: Ontology-enhanced zero-shot learning. In *Association for Computing Machinery (ACM '2021)* (pp. 3325–3336). volume 12. <https://doi.org/10.1145/3442381.3450042>.
- Lampert, C. H., Nickisch, H., & Harmeling, S. (2009). Learning to detect unseen object classes by between-class attribute transfer (cvpr '09). In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 951–958). <https://doi.org/10.1109/CVPR.2009.5206594>.
- Li, X., Xu, Z., Wei, K., & Deng, C. (2021). Generalized zero-shot learning via disentangled representation. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI '21)* (pp. 1966–1974). volume 35. <https://doi.org/10.1609/aaai.v35i3.16292>.
- Liu, Y., Zhou, L., Bai, X., Huang, Y., Gu, L., Zhou, J., & Harada, T. (2021). Goal-oriented gaze estimation for zero-shot learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR '21)* (pp. 3794–3803). <https://doi.org/10.1109/cvpr46437.2021.00379>.
- Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heintz, I., & Roth, D. (2023). Recent advances in natural language processing via large pre-trained language models: A survey. *Association for Computing Machinery Computing Surveys (ACM '23)*, 56, 1–40. <https://doi.org/10.1145/3605943>.
- Min, S., Yao, H., Xie, H., Wang, C., Zha, Z.-J., & Zhang, Y. (2020). Domain-aware visual bias eliminating for generalized zero-shot learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR '20)* (pp. 12661–12670). <https://doi.org/10.1109/CVPR42600.2020.01268>.

- Naeem, M. F., Xian, Y., Gool, L. V., & Tombari, F. (2022). I2dformer: Learning image to document attention for zero-shot image classification. In *Proceedings of the 36th Annual Conference on Neural Information Processing Systems (NIPS '22)*, 35, 12283–12294. <https://doi.org/10.3929/ethz-b-000593721>.
- Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, 9, 353–383. [https://doi.org/10.1016/0010-0285\(77\)90012-3](https://doi.org/10.1016/0010-0285(77)90012-3).
- Patterson, G., & Hays, J. (2012). Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'12)* (pp. 2751–2758). <https://doi.org/10.1109/cvpr.2012.6247998>.
- Pourpanah, F., Abdar, M., Luo, Y., Zhou, X., Wang, R., Lim, C. P., Wang, X.-Z., & Wu, Q. J. (2022). A review of generalized zero-shot learning methods. *2022 IEEE transactions on pattern analysis and machine intelligence (TPAMI '22)*, 45, 4051–4070. <https://doi.org/10.1109/TPAMI.2022.3191696>.
- Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. *2016 International Conference on Learning Representations (ICLR '16)*, . <https://doi.org/10.48550/arXiv.1511.06434>.
- Shen, Y., Qin, J., Huang, L., Liu, L., Zhu, F., & Shao, L. (2020). Invertible zero-shot recognition flows. In *2020 European Conference on Computer Vision (ECCV '21)* (pp. 614–631). https://doi.org/10.1007/978-3-030-58517-4_36.
- Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., & Perona, P. (2010). Caltech-ucsd birds 200, .
- Xian, Y., Lorenz, T., Schiele, B., & Akata, Z. (2018). Feature generating networks for zero-shot learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'18)* (pp. 5542–5551). <https://doi.org/10.1109/CVPR.2018.00581>.
- Xian, Y., Schiele, B., & Akata, Z. (2017). Zero-shot learning-the good, the bad and the ugly. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '17)* (pp. 4582–4591). <https://doi.org/10.1109/cvpr.2017.328>.
- Xie, G.-S., Liu, L., Jin, X., Zhu, F., Zhang, Z., Qin, J., Yao, Y., & Shao, L. (2019). Attentive region embedding network for zero-shot learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR '19)* (pp. 9384–9393). <https://doi.org/10.1109/cvpr.2019.00961>.
- Xie, G.-S., Liu, L., Zhu, F., Zhao, F., Zhang, Z., Yao, Y., Qin, J., & Shao, L. (2020). Region graph embedding network for zero-shot learning. In *2020 European Conference on Computer Vision (ECCV '20)* (pp. 562–580). https://doi.org/10.1007/978-3-030-58548-8_33.
- Xie, G.-S., Zhang, X.-Y., Xiang, T.-Z., Zhao, F., Zhang, Z., Shao, L., & Li, X. (2022). Leveraging balanced semantic embedding for generative zero-shot learning. *2022 IEEE Transactions on Neural Networks and Learning Systems (TNNLS '22)*, 34, 9575–9582. <https://doi.org/10.1109/tnnls.2022.3208525>.
- Xu, B., Zeng, Z., Lian, C., & Ding, Z. (2021). Semi-supervised low-rank semantics grouping for zero-shot learning. *2021 IEEE Transactions on Image Processing (TIP '21)*, 30, 2207–2219. <https://doi.org/10.1109/tip.2021.3050677>.
- Xu, W., Xian, Y., Wang, J., Schiele, B., & Akata, Z. (2020). Attribute prototype network for zero-shot learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20)*, 33, 21969–21980. <https://doi.org/10.5555/3495724.3497567>.
- Yue, Z., Wang, T., Sun, Q., Hua, X.-S., & Zhang, H. (2021). Counterfactual zero-shot and open-set visual recognition. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR '21)* (pp. 15404–15414). <https://doi.org/10.1109/cvpr46437.2021.01515>.
- Zhang, L., Xiang, T., & Gong, S. (2017). Learning a deep embedding model for zero-shot learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '17)* (pp. 2021–2030). <https://doi.org/10.1109/cvpr.2017.321>.
- Zhao, X., Ouyang, S., Yu, Z., Wu, M., & Li, L. (2023). Pre-trained language models can be fully zero-shot learners. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL '23)*, (p. 15590–15606). <https://doi.org/10.18653/v1/2023.acl-long.869>.