Information Sciences 642 (2023) 119135

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Group-wise interactive region learning for zero-shot recognition

Ting Guo [a], Jiye Liang [a,*], Guo-Sen Xie [b]

[a] Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, School of Computer and Information Technology, Shanxi University, Taiyuan, Shanxi, China

[b] Nanjing University of Science and Technology, Nanjing, Jiangsu, China

ARTICLE INFO

Keywords:
Zero-shot learning
Region feature
Semantic consistency

ABSTRACT

Zero-shot learning (ZSL) methods mainly associate global or region features to semantic vectors within a single image, for transferring semantic knowledge from the seen classes to unseen ones. However, the interactive region learning among a group of images from different categories, which can enhance the discrimination of region features and thus lead to a desirable knowledge transfer between seen and unseen classes, is seldom considered. To remedy the above challenge, we propose a group-wise interactive region learning (GIRL) model to guarantee a comprehensive and explicit region interaction. Specifically, GIRL consists of an attentive region interaction (ARI) module and a holistic semantic embedding (HSE) module. ARI utilizes the semantic commonalities and differences of group regions to produce refined region features. HSE holistically maps these region features to the semantic space for a more stable semantic transfer. We also present a semantic consistency loss and a relation alignment loss that can distill the refined/original region features and introduce unseen class semantic vectors for training, respectively. Extensive experiments demonstrate the effectiveness of GIRL over other methods, achieving 68.9%, 42.9%, 75.5%, and 47.8% the Generalized ZSL (GZSL) H scores on CUB, SUN, AWA2, and APY. The code is publicly available at https://github.com/TingML/GIRL.

1. Introduction

Zero-shot learning (ZSL) [1–4] aims to classify unseen class images by leveraging the potential semantic relationships between seen and unseen ones. To achieve this, existing works introduce class semantic vectors that bridge the semantic gap between the seen and unseen classes, making ZSL feasible. The most widely used class semantic vectors are attributes [5] and word2vector [6], with attributes serving as the pivotal high-level auxiliary knowledge in ZSL. Apart from addressing the time-consuming nature of recognition tasks, ZSL can also handle the challenge of identifying new categories.

The successful ZSL paradigms learn the embedding function from the seen class images to their corresponding semantic vectors. However, earlier methods [7–11] rely on aligning global image features with ground-truth semantic vectors in the embedding space, leading to an inferior transfer of discriminative knowledge. To overcome this limitation, researchers have further proposed region-based methods for mining local discriminative region features. As in Fig. 1 (a), some methods (e.g., APN [12], AREN [13], DAZLE [14], DPPN [15], MSDN [16], TransZero [17], AREES [18]) directly identify discriminative local regions and align them with their ground-truth semantic vectors. Additionally, as in Fig. 1(b), RGEN [19] advocates for capturing the appearance relationships of variant region features. Nonetheless, all the aforementioned methods are merely based on single images to model region relationships

* Corresponding author.
E-mail addresses: gting1151@gmail.com (T. Guo), ljy@sxu.edu.cn (J. Liang), gsxiehm@gmail.com (G.-S. Xie).

https://doi.org/10.1016/j.ins.2023.119135
Received 3 October 2022; Received in revised form 25 April 2023; Accepted 7 May 2023

Available online 11 May 2023

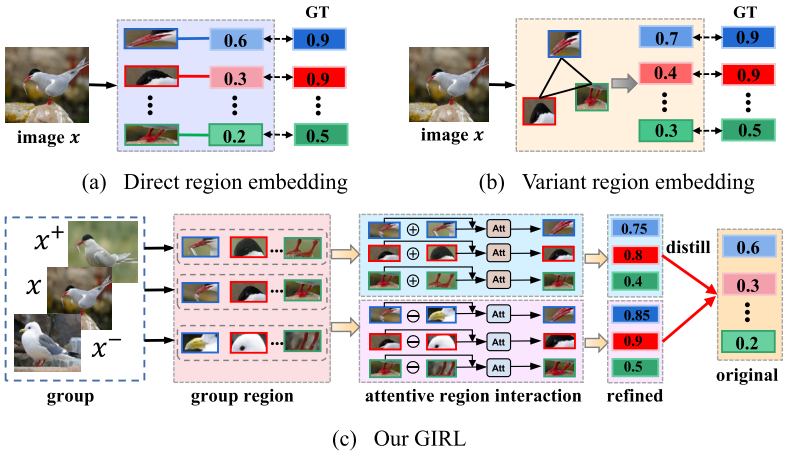


Fig. 1. Motivation illustration. The architecture of existing methods *vs.* ours. (a) The existing methods identify local regions to align directly with the ground truth attribute vector. (b) Some methods learn the variant region embedding by capturing these region appearance relationships. The methods of (a) and (b) are merely based on single images to model regions. (c) We propose to refine region features by the attentive region interaction among a group of images from different categories, and distill the semantic knowledge from the refined region feature to the original ones promoting semantic consistency between the both. "GT" means ground truth semantic vector.

and fail to capture intrinsic contrastive cues (semantic commonality and difference) for regions of different images from different classes. Considering the semantic commonalities and differences of regions can enhance the discriminative of region features, and thus the knowledge transfer from seen to unseen classes is more desirable.

To address the aforementioned challenge, we propose a group-wise interactive region learning (GIRL) model to guarantee a comprehensive and explicit region interaction among a group of images from different categories. GIRL can capture crucial contrastive cues from regions of different images with different classes, resulting in desirable knowledge transfer from seen to unseen classes, as shown in Fig. 1(c). Different from methods modeling regions in a single image, GIRL seeks more semantic information between regions based on group data. Especially, group data are achieved by a crafted group construction method to ensure that a group contains data whose semantics can complement each other. We apply the attentive region interaction (ARI) module to capture the semantic differences and commonalities of the corresponding regions by region interaction within a group data. We then perform semantic-aware region feature refinement on original region features. Further, the semantic consistency loss is utilized to distill semantic knowledge to the original region features, facilitating the model's localization of precise attribute regions. In other words, semantic consistency loss also provides supervised information to the model. Additionally, considering the absence of unseen classes in training, we design relation alignment loss to further constrain the visual-semantic alignment relation. In summary, our contributions are:

- We propose a group-wise interactive region learning (GIRL) model, which takes advantage of richer semantic information in a group of images to boost the discriminability and transferability of region features.
- The attentive region interaction (ARI) module is intended to refine the region features through region interaction in a group. A holistic semantic embedding (HSE) module is further incorporated to map these region features to the semantic space for a more stable semantic transfer.
- GIRL introduces a semantic consistency loss and a relation alignment loss that can distill the original/refined region features and can introduce unseen class semantic vectors for model training.

2. Related works

2.1. Zero-shot learning

Zero-shot learning (ZSL) aims to infer the relationship between the visual and class semantic vectors of the seen classes at the semantic level by fully integrating visual and semantic, and transferring the learned knowledge (semantic relationships) to the unseen classes. In ZSL, the direct-embedding methods [7,20,21,10,11], generation-based methods [22,9,23–25], and region-based methods have all been extensively researched. Direct-embedding methods (e.g., ESZSL [7] and LATEM [20]) establish joint embedding spaces to align image features and semantic vectors. ESZSL [7] designs two layers to model the relationships between properties, attributes, and classes. However, since these two layers consist of straightforward linear mapping, the relationships learned being linear and simple. LATEM [20] improves on this by learning the mappings set of images and text and upgrading the linear mapper to a nonlinear one. However, the above methods create simple mapping and primarily concentrate on the model, which leads to inferior zero-shot performance. The generation-based methods (e.g., f-CLSWGAN [22] and CE-GZSL [23]) are proposed to train the semantic-visual generator to synthesize visual features of unseen classes, converting zero-shot learning into traditional supervised learning. f-CLSWGAN [22] optimizes the semantic-visual generator utilizing Wasserstein generative adversarial net (WGAN) [26]

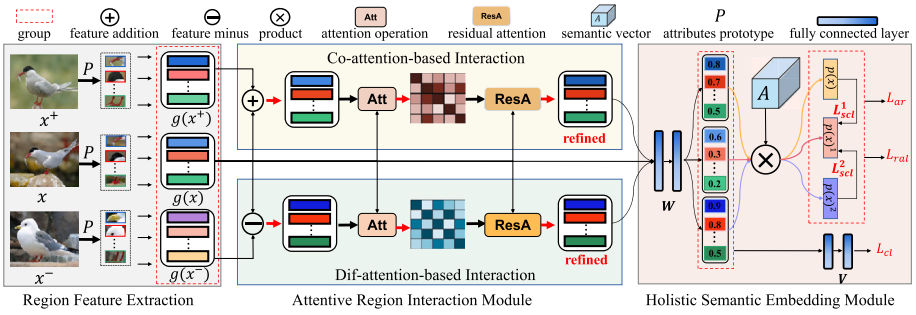


Fig. 2. The overall framework of the proposed method (GIRL). First, we conduct region feature extraction mapping image x to generate their region features leverage the attribute prototypes $\mathcal{P}$. Then an attentive region interaction module (ARI) is designed to capture and emphasize semantic differences and commonalities through the interaction of region features in a group of images to guide semantic-aware region feature refinement. Further, a holistic semantic embedding module (HSE) maps the region feature to the semantic space for a more stable semantic transfer.

and classification loss. CE-GZSL [23] trains semantic-visual generators using instance-level and class-level contrastive loss. Although generation-based methods display superior zero-shot performance, the reliability of the generated features cannot be guaranteed. To mitigate this, additional strategies have been devised to address domain bias, such as generating more visual features of unseen classes than seen ones. The direct-embedding methods and the generation-based methods focus on the learning of global features, while the region-based methods emphasize mining the fine-grained attribute knowledge of images through the learning of region features.

Region-based methods ensure the alignment of local features with semantic vectors, further achieving the effective knowledge transfer from the seen classes to the unseen ones. APN [12] learns local features of a single image using the attribute prototype network and incorporates a visual semantic embedding layer to learn global features. To improve generalization to unseen classes, APN [12] decorrelates irrelevant attribute prototypes. The DPPN [15] iteratively refines local features using a prototype progressive network. In contrast, RGEN [19] considers the relationship between different local features in a single image and employs a region graph mapping network to guarantee the learning of discriminative features. However, these methods establish region relationships based on a single image and fail to capture the intrinsic contrast cues of different image regions from different categories. In this paper, instead of learning region features in a single image as in previous studies, we investigate the region relationships in a group of images to improve the model's learning of fine-grained local features better.

2.2. Knowledge distillation

Following a teacher-student training structure, the knowledge distillation [27] is designed to utilize a complex and well-performing teacher model to transfer the "knowledge" to a simple student model, improving the performance of the student model. In knowledge distillation, the forms of "knowledge" comprise knowledge of output features, intermediate features, relational features, and structural features. The key to knowledge distillation is to effectively transfer knowledge from the teacher model to the student model. Widely used knowledge distillation methods comprise knowledge amalgamation [28], mutual distillation [29], and self-distillation [30], where self-distillation allows an individual model to enhance performance through knowledge distillation in a self-learning process. Knowledge distillation has been utilized in ZSL for distilling the intrinsic semantic knowledge between the attribute-based visual features and visual-based attribute features [16]. However, due to the inseparability of some attribute description vectors, MSDN [16] still has trouble learning the fine-grained attribute features. In contrast to MSDN which relies on external knowledge, we perform semantic distillation by mining the information in the dataset itself.

3. The proposed approach

3.1. Problem setting and notations

In ZSL problem, suppose that the training set (seen classes) is given, $D^S = \{(x_i, y_i, a_{y_i}) | x_i \in \mathcal{X}^S, y_i \in \mathcal{Y}^S, a_{y_i} \in \mathcal{A}^S\}$, where $x_i \in \mathcal{R}^{d_1 \times 1}$ is the d_1-dimensional visual feature and y_i is the label of x_i in the training set. $a_{y_i} \in \mathcal{A}^S \in \mathcal{R}^{d_2 \times 1}$ is a class semantic vector, which is constituted by d_2 attributes. The recognition of unseen classes is the ZSL's primary goal. So the effectiveness of the model is verified by a test dataset $D^U = \{(x_i, y_i, a_{y_i}) | x_i \in \mathcal{X}^U, y_i \in \mathcal{Y}^U, a_{y_i} \in \mathcal{A}^U\}$. The seen and unseen label sets are disjoint, i.e., $\mathcal{Y}^S \cap \mathcal{Y}^U = \emptyset$, thus, $\mathcal{A}^S \cap \mathcal{A}^U = \emptyset$. With $\mathcal{A} = \mathcal{A}^S \cup \mathcal{A}^U$, the goal of ZSL is to learn a prediction: $\mathcal{X}^U \rightarrow \mathcal{Y}^U$. The test samples come from the both seen and unseen classes $D^U = \{(x_i, y_i, a_{y_i}) | x_i \in \mathcal{X}, y_i \in \mathcal{Y}, a_{y_i} \in \mathcal{A}\}$ leads to generalized ZSL (GZSL). For GZSL, we learn a prediction: $\mathcal{X} \rightarrow \mathcal{Y}$, $\mathcal{X} = \mathcal{X}^S \cup \mathcal{X}^U$, $\mathcal{Y} = \mathcal{Y}^S \cup \mathcal{Y}^U$.

In this paper, we propose group-wise interactive region learning (GIRL), which takes advantage of richer semantic information in a group of images to support the model localize fine-grained attribute regions. In Fig. 2, GIRL comprises an attentive region interaction (ARI) module and a holistic semantic embedding (HSE) module. ARI refines the original region features through region interactions within a group of image data. HSE aligns visual and semantic vectors in latent space for ZSL classification. Furthermore, group construction is critical in GIRL. The data whose semantics complement each other will be beneficial to the succeeding activities.

3.2. Group construction

In this section, we describe how to construct groups that are effective for subsequent tasks. First, we explain how to construct a batch. We randomly select N_{cl} classes for a batch, and for each class, we sample N_{sam} images. Specially, we follow DPPN [15] to conduct region feature extraction, which mapping image x_i to generate their region features by leveraging the attribute prototypes $\mathcal{P}$. Define the learnable attributes prototype as $\mathcal{P} = \{p_1, \cdots, p_k, \cdots, p_{d_2}\}$, p_k denotes the k^{th} attribute, such as the visual representation of the "yellow back" shared by all birds. $\boldsymbol{r}_i^k = f(x_i, p_k) \in R^{d\times 1}$ denotes the k^{th} attribute prototype to extract the most relevant region features. Based on d_2 prototypes, follow DPPN [15] we obtain region-level visual features $R_i = \{r_i^k\}_{k=1}^{d_2}$. In DPPN, $R_i = f(x_i, P) = \phi(x_i \hbar(x_i^{\mathrm{T}} P))$, where ϕ is a prototype function implemented by two fully-connected (FC) layers, $\hbar$ is a softmax normalization along each column.

Then the group data is constructed in a batch. We compare the visual features, R_i, within a batch by calculating their Euclidean distance from one another. Based on the similarity of region features, for each image we construct a group of data (R_i, R_i^+, R_i^-) containing the region feature R_i, the feature R_i^+ from the same class most similar to R_i, and the feature R_i^- from a different class most similar to R_i. The commonality of R_i and R_i^+ reflects finer-grained information, and the semantic differences of R_i and R_i^- can play a complementary role to R_i. Utilizing information on the semantic commonality and differences among regions is critical in ZSL, as it can improve the distinguishability and transferability of regional features.

3.3. Attentive region interaction module

The central idea behind ARI is to leverage the interactions of the regions in a group to obtain refined region features. Humans can easily discern the differences or similarities between image regions and utilize these contrasting cues to strengthen their understanding of regions. Inspired by human behavior, we incorporate two branches into ARI: co-attention-based interaction (CAI) and dif-attention-based interaction (DAI). These branches simulate the way humans interact and understand visual information.

Co-attention-based Interaction. First, we perform R_i and R_i^+ regions interaction to gain the commonalities of intra-pairs in a group data,

$$C_i = (R_i + R_i^+))/2. \tag{1}$$

The vector emphasizes class-related region features (e.g. the throat region features of *yellow throat vireo* are all yellow). Then we propose to compare R_i and C_i, as the comparison can yield valuable semantic contrastive clues. By this, we can gain a more nuanced understanding of the differences between the two.

$$\Gamma_i = R_i \otimes C_i^{\mathrm{T}}, \tag{2}$$

where $\otimes$ means tensor product. Thus, Γ_i is an attention feature reflecting the semantic relationship between R_i and C_i. This attentional feature further emphasizes the importance of discriminative regions.

Finally, we perform region refined guided by the attention feature. We utilize the following formula for region refinement and refer to this operation as ***ResA***:

$$R_i^{\mathrm{ref}1} = R_i + \Gamma_i \otimes R_i. \tag{3}$$

In this regard, our objective is to derive attention features from the interaction of related regions within the group, utilize this attention feature to extract distinct semantic information, and subsequently refining the regions.

Dif-attention-based Interaction. In this branch, we perform R_i and R_i^- regions interaction,

$$D_i = (R_i - R_i^-)/2. \tag{4}$$

D_i is able to indicate the differences in region features (e.g. the distinctive region feature that differentiates the bird *warning siren* from the *yellow throat vireo* is the throat color).

Subsequently, we compare R_i and D_i to identify contrasting clues between them and determine the significance of the contrasting information.

$$\Upsilon_i = R_i \otimes D_i^{\mathrm{T}}. \tag{5}$$

Υ_i is a discriminative attention that emphasizes the different relationship between R_i and D_i, and this difference is an important complementary information for R_i. Finally, we perform region feature refinement on R_i by the operation ***ResA***:

$$R_i^{\mathrm{ref}2} = R_i + D_i + \alpha \Upsilon_i \otimes R_i, \tag{6}$$

where α is a hyper-parameter. Our objective is to derive discriminative attentive features by leveraging the interactions between inter-pairs within the group, thus guiding the refinement of semantic-aware region features.

3.4. Holistic semantic embedding module

By fully interacting with the group data, we obtain refined region features $R_i^{\text{ref}1}$ and $R_i^{\text{ref}2}$. Together with the original region features R_i, we utilize a linear layer with parameter W further maps them to the semantic space,

$$\psi(x_i) = R_i W. \tag{7}$$

Correspondingly, $R_i^{\text{ref}1}$ and $R_i^{\text{ref}2}$ are mapped to $\psi(x_i)^1$ and $\psi(x_i)^2$, respectively.

3.5. Model optimization

To effectively optimize the performance of our GIRL model, we design the attribute alignment loss, classification loss, semantic consistency loss and relation alignment loss to train GIRL.

Attribute Alignment Loss. We develop the attribute alignment loss constrained GIRL to ensure that GIRL can precise mapping visual features to their corresponding semantic vectors. The visual features are mapped to the semantic space through a linear layer with parameter W to obtain the representation $\psi(x_i)$, we next compute the similarity between $\psi(x_i)$ and the ground-truth attribute vector a_{y_i}, attend to maximize this similarity,

$$\mathcal{L}_{\text{ar}} = -\sum_{x_i \in \mathcal{X}, y_i \in \mathcal{Y}^S} \psi(x_i) \times g(a_{y_i})^{\text{T}}. \tag{8}$$

The function $g(\cdot)$ in GIRL serves as a semantic mapping tool, accomplished through a straightforward fully connected layer that maps the semantic vectors into the latent space.

Classification Loss. The classification loss is proposed to further bootstrap GIRL by getting more discriminative visual features and enlarging the class boundaries. In detail, a linear layer with parameter V maps the visual feature $\psi(x_i)$ into the class embedding space, $q_i = \psi(x_i)V$. Thus we obtain the class logit q_i of feature and subsequently utilize the cross entropy loss $\mathcal{L}_{cl}$ as a classification constraint.

$$\mathcal{L}_{\text{cl}} = -\sum_{i \in |N|} y_i' \times \ln q_i^{\text{T}}, \tag{9}$$

where $|N|$ is the sample number in training set, y_i' is the one-hot vector of y_i.

Semantic Consistency Loss. The semantic consistency loss aims to distill the distinct semantic information from the refined region features to original region features, boosting the semantic consistency between both. $R_i^{\text{ref}1}$ and $R_i^{\text{ref}2}$ are refined region features. Mapping these features into the latent space, R_i, $R_i^{\text{ref}1}$ and $R_i^{\text{ref}2}$ is represented as $\psi(x_i)$, $\psi(x_i)^1$ and $\psi(x_i)^2$, respectively. Then the class logit is generated by computing the dot product between the projected visual feature and all class semantic embeddings. For $\psi(x_i)$, the class logit is $p(x_i) = \{\psi(x_i) \times g(a_1)^{\text{T}}, \cdots, \psi(x_i) \times g(a_{|N_s|})^{\text{T}}\}$. Analogously, $\psi(x_i)^1$ and $\psi(x_i)^2$ are $p(x)^1$ and $p(x_i)^2$. $|N_s|$ is the number of seen classes. In GIRL, semantic consistency loss is intended to guide $p(x_i)$ using $p(x_i)^1$ and $p(x_i)^2$, formulated as:

$$\begin{aligned}\mathcal{L}_{\text{scl}} = \mathcal{L}_{\text{scl}}^1 + \mathcal{L}_{\text{scl}}^2 = \sum_{x_i \in \mathcal{X}} (&D_{KL}(p(x_i)||p(x_i)^1) \\ &+ D_{KL}(p(x_i)||p(x_i)^2)),\end{aligned} \tag{10}$$

where $D_{KL}(p||p') = \sum_{c=1}^{|N_s|} p_c \log(\frac{p_c}{p_c'})$.

Relation Alignment Loss. In ZSL and GZSL, a major concern is the bias problem that arises since only the seen classes images are utilized during training. To mitigate this issue, we propose a relation alignment loss. First, we utilize the product to calculate the relation $S = \mathcal{A}^S \times \mathcal{A}^{U^{\text{T}}}, S \in R^{|N_s| \times |N_u|}$ between the attribute semantic vectors of seen classes $\mathcal{A}^S$ and the unseen classes $\mathcal{A}^U$. Semantic relations are utilized as ground truth values to guide the predicted probability of visual features further. The relation alignment loss is formulated as:

$$\mathcal{L}_{\text{ral}} = \sum_{x_i \in \mathcal{X}} \sum_{j=1}^{|N_u|} s_{y_i j} \log \widetilde{\xi_{ij}} + (1 - s_{y_i j}) \log(1 - \widetilde{\xi_{ij}}), \tag{11}$$

where $\xi_{ij} = \psi(x_i) a_j^{u\text{T}}$, $a_j^u \in \mathcal{A}^U$, $|N_u|$ is the number of unseen classes. $\widetilde{\xi_{ij}}$ is the softmax-layer normalization of ξ_{ij}, y_i is the label of x_i, also is the column location in S, $s_{y_i j}$ tends to reflect the relationship between x_i and a_j^u. The relation alignment loss is similar to RGEN [13] but differs from it: for the class relation S, RGEN utilizes least square regression (LSR) whereas we employ a straightforward product computation, this makes our method more simpler.

Finally, the total loss function of GIRL for end-to-end model training with four losses is as follows:

$$\mathcal{L} = \mathcal{L}_{\text{ar}} + \mathcal{L}_{\text{ce}} + \lambda_1 \mathcal{L}_{\text{scl}} + \lambda_2 \mathcal{L}_{\text{ral}}, \tag{12}$$

where λ_1 and λ_2 are the hyperparameters.

Table 1
GZSL results (%) on CUB, SUN, AWA2, and APY. The best and the second best results are marked in red and blue. † denotes the GZSL results of DPPN without finetune the backbone. (For interpretation of the colors in the tables, the reader is referred to the web version of this article.)

Methods	CUB			SUN			AWA2			APY		
	Au	As	H	Au	As	H	Au	As	H	Au	As	H
Direct-Embedding Methods												
ESZSL (ICML 2015) [7]	12.6	63.8	21.0	11.0	27.9	15.8	5.9	77.8	11.0	2.4	70.1	4.6
SJE (CVPR 2015) [34]	23.5	59.2	33.6	14.7	30.5	19.8	8.0	73.9	14.4	3.7	55.7	6.9
SYNC (CVPR 2016) [8]	11.5	70.9	19.8	7.9	43.3	13.4	10.0	90.5	18.0	7.4	66.3	13.3
LATEM (CVPR 2016) [20]	15.2	57.3	24.0	14.7	28.8	19.5	11.5	77.3	20.0	0.1	73.0	0.2
ALE (TPAMI 2016) [21]	23.7	62.8	34.4	21.8	33.1	26.3	14.0	81.8	23.9	4.6	73.7	8.7
FGN+WL (WACV 2022) [35]	49.4	51.8	50.6	46.3	28.7	35.5	45.8	83.4	59.2	17.9	73.3	28.7
GatingAE (TIP 2022) [10]	58.1	54.9	56.4	38.1	45.4	41.4	81.3	57.3	67.2	-	-	-
BGSNet (TMM 2023) [11]	60.9	73.6	66.7	45.2	34.3	39.0	61.0	81.8	69.9	31.0	54.9	39.7
Generation-Based Methods												
f-CLSWGAN (CVPR 2018) [22]	43.7	57.7	49.7	42.6	36.6	39.4	57.9	61.4	59.6	32.9	61.7	42.9
cycle-CLSGAN (ECCV 2018) [36]	45.7	61.0	52.3	49.4	33.6	40.0	56.9	64.0	60.2	29.3	51.2	37.3
CADA-VAE (CVPR 2019) [9]	51.6	53.5	52.4	47.2	35.7	40.6	55.8	75.0	63.9	30.3	61.3	40.6
f-VAEGAN-D2 (CVPR 2019) [37]	48.4	60.1	53.6	45.1	38.0	41.3	57.6	70.6	63.5	-	-	-
CE-GZSL+SDFA2 (AAAI 2022) [38]	59.2	59.6	54.0	46.2	32.6	38.2	59.3	75.0	66.2	21.5	85.1	34.3
TDCSS (CVPR 2022) [39]	44.2	62.8	51.9	-	-	-	59.2	74.9	66.1	-	-	-
ICCE (CVPR 2022) [25]	67.3	65.5	66.4	-	-	-	65.3	82.3	**72.8**	45.2	46.3	45.7
ABA-GAN (TMM 2023) [40]	59.8	53.8	56.6	39.4	44.2	41.7	66.8	62.9	64.8	65.0	35.9	**46.3**
Region-Based Methods												
AREN (CVPR 2019) [13]	38.9	78.7	52.1	19.0	38.8	25.5	15.6	92.9	26.7	30.0	47.9	36.9
LFGAA (ICCV 2019) [41]	36.2	80.9	50.0	18.5	40.0	25.3	27.0	93.4	41.9	-	-	-
DAZLE (CVPR 2020) [14]	56.7	59.6	58.1	52.3	24.3	33.2	60.3	75.7	67.1	-	-	-
RGEN (ECCV 2020) [19]	60.0	73.5	66.1	44.0	31.7	36.8	67.1	76.5	71.5	48.1	30.4	37.2
APN (NeurIPS 2020) [12]	65.3	69.3	67.2	41.9	34.0	37.6	57.1	72.4	63.9	-	-	-
† DPPN (NeurIPS 2021) [15]	60.4	66.1	63.1	45.0	36.2	40.2	61.8	86.8	72.2	35.6	58.7	44.3
I2DFormer (NeurIPS 2022) [42]	35.3	57.6	43.8	-	-	-	66.8	76.8	71.5	-	-	-
ERPCNet (TCSVT 2022) [43]	67.1	69.6	68.4	47.2	31.9	38.1	59.1	82.0	68.7	32.7	49.3	39.3
AREES (TNNLS 2022) [18]	53.6	56.9	55.2	51.3	35.9	**42.2**	57.9	77.0	66.1	34.3	69.3	46.0
MSDN (CVPR 2022) [16]	68.7	67.5	68.1	52.2	34.2	41.3	62	74.5	67.7	-	-	-
TransZero (AAAI 2022) [17]	69.3	68.3	**68.8**	52.6	33.4	40.8	61.3	82.3	70.2	-	-	-
GIRL (Ours)	66.2	71.8	**68.9**	47.9	38.8	**42.9**	68.5	84.3	**75.5**	37.0	67.6	**47.8**

3.6. Zero-shot prediction

Considering that attentive region interaction module is a plug-and-play module and that the group construction cannot be done since the label is unknown, the module should not be included in test stage. During the test phase, the embedding $\psi(x_i)$ of x_i in the latent space is obtained by the network. We predict its label by following:

$$y_i^* = \arg \max_{y_i \in \mathcal{Y}^U / \mathcal{Y}} \psi(x_i) \times g(a_{y_i}), \tag{13}$$

$y_i \in \mathcal{Y}^U / \mathcal{Y}$ corresponds to ZSL/GZSL.

4. Experiment

4.1. Datasets

We conducted extensive experiments on four ZSL benchmarks, i.e., Caltech-USCD Birds-200-2011 (CUB) [31], SUN [32], Animals with Attributes2 (AWA2) [1], and Attribute Pascal and Yahoo (APY) [33] to verify the validity of GIRL. CUB contains 11788 images of 200 fine-grained birds classes, of which 150 are seen and 50 are unseen classes, and the class semantic vectors are 312 attributes. SUN contains 14340 images of 717 fine-grained scene classes, of which 645 are seen and 72 are unseen classes, with the class semantic vectors are 102 attributes. AWA2 is a coarse-grained dataset compared to CUB and SUN, which contains 37322 images of 50 animal classes, of which 40 are seen and 10 are unseen classes, and the class semantic vectors are 85 attributes. APY consists of 15339 images categorized into 32 classes, with 20 being seen and 12 being unseen classes. The dataset is coarse-grained and comprises 64 attribute-based class semantic vectors.

4.2. Evaluation

In this paper, we validate the proposed method under both ZSL and GZSL settings. For ZSL, the test dataset contains only unseen classes, so the evaluation criterion is the average class accuracy of unseen classes Acc. For GZSL, the test dataset contains both seen

Table 2
ZSL results (%) on CUB, SUN, AWA2, and APY. The best and the second best results are marked in red and blue. † denotes the ZSL results of DPPN obtained by using the code released by the authors.

Methods	CUB	SUN	AWA2	APY
	Acc	Acc	Acc	Acc
Direct-Embedding Methods				
ESZSL (ICML 2015) [7]	53.9	54.5	58.6	38.3
SJE (CVPR 2015) [34]	53.9	53.7	61.9	32.9
SYNC (CVPR 2016) [8]	55.6	56.3	46.6	23.9
LATEM (CVPR 2016) [20]	49.3	55.3	55.8	35.2
ALE (TPAMI 2016) [21]	54.9	58.1	62.5	39.7
FGN+WL (WACV 2022) [35]	62.3	64.4	69.1	39.9
BGSNet (TMM 2023) [11]	73.3	63.9	69.1	43.8
Generation-Based Methods				
f-CLSWGAN (CVPR 2018) [22]	57.3	60.8	68.2	-
cycle-CLSGAN (ECCV 2018) [36]	58.4	60.1	66.3	38.6
CADA-VAE (CVPR 2019) [9]	59.8	61.7	63.0	35.7
f-VAEGAN-D2 (CVPR 2019) [37]	61.0	64.7	71.1	
ICCE (CVPR 2022) [25]	**78.4**	-	72.7	**49.5**
ABA-GAN (TMM 2023) [40]	62.0	62.5	72.0	44.6
Region-Based Methods				
AREN (CVPR 2019) [13]	71.8	60.6	67.9	36.9
LFGAA (ICCV 2019) [41]	67.6	61.5	68.1	-
DAZLE (CVPR 2020) [14]	66.0	59.4	67.9	-
RGEN (ECCV 2020) [19]	76.1	63.8	73.6	37.2
APN (NeurIPS 2020) [12]	72.0	61.6	68.4	-
† DPPN (NeurIPS 2021) [15]	72.6	64.6	68.9	-
I2DFormer (NeurIPS 2022) [42]	45.4	-	**76.4**	-
ERPCNet (TCSVT 2022) [43]	72.5	63.1	71.8	43.5
AREES (TNNLS 2022) [18]	65.7	64.3	73.6	44.2
MSDN (CVPR 2022) [16]	76.1	**65.8**	70.0	-
TransZero (AAAI 2022) [17]	**76.8**	65.6	70.1	-
GIRL (Ours)	75.2	**66.5**	**73.9**	**48.5**

and unseen classes, and the evaluation metrics comprise Au, As, and H. As/Au is the class average accuracy of seen/unseen classes, H is defined as $H = 2 \times (As \times Au)/(As + Au)$. A higher H means a higher class average accuracy of seen and unseen classes, the more balanced the accuracy of the two, i.e., the better the model is able to handle the domain bias problem.

4.3. Implementation details

GIRL utilizes ResNet-101 [44], pre-trained on ImageNet, as its backbone for extracting visual features without fine-tuning. We utilize adam optimizer [45] with hyperparameters (momentum = 0.9, weight decay = $1e{-}4$). For CUB, SUN, AWA2, and APY, the learning rates are $2e{-}4$, $1e{-}4$, $4e{-}4$, and $2e{-}4$, the batchsize b are 32, 64, 64, and 20, respectively. In our model, the input image is 448×448 following [19,15]. More detailed information about the hyperparameters and implementation details can be found in Section 4.5.

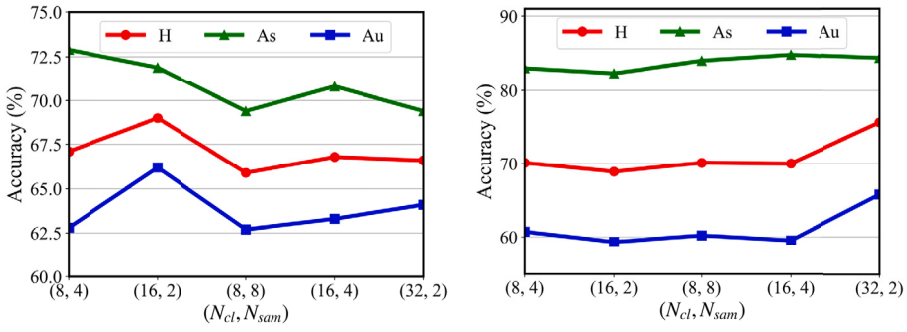
4.4. Comparison with state-of-the-art methods

Generalized Zero-Shot Learning We first compare GIRL with the current state-of-the-art methods, which comprise direct-embedding methods, generation-based methods, and region-based methods in the GZSL setting. From Table 1, we can conclude the following:

i) GIRL outperforms the state-of-the-art methods on four datasets. H score is marginally superior to TransZero [17] on CUB, with a mere 0.1% increase. However the latter utilizes external information, i.e., additional word embedding. Compared to the direct-embedding methods and generation-based methods, our method improves by at least 2.2%. Without external information, GIRL constructs group data and designs intra-group data interactions to improve the model's localization of fine-grained attribute features.

ii) For SUN, GIRL obtains the top-1 accuracy of $H = 42.9\%$, which exceeds the best current result $H = 42.2\%$ of AREES [18] by 0.7%. RGEN [19], DPPN [15], TransZero [17], and MSDN [16] are most relevant to our approach, which are all devoted to mining discriminative region knowledge. Compared to these methods, our method has 6.1%, 2.7%, 2.1%, and 1.6% improvements, respectively. The results fully illustrate that GIRL uses its own dataset information to design semantic consistency loss and relational alignment loss, further boosting the discriminability and transferability of region features.

iii) GIRL obtains the top-1 accuracy of $H = 75.5\%$, which exceeds the best current result $H = 72.8\%$ by 2.7% on AWA2. When compared to the approaches that are comprised in the region-based methods, our H score outperforms TransZero [17] by 5.3%,



(a) CUB (b) AWA2

Fig. 3. The effects of N_{cl} and N_{sam} on CUB and AWA2, in terms of H, As, Au.

RGEN [19] by 4.0%, and DPPN [15] by 3.3%. In contrast to these methods, GIRL guarantees the discriminability and transferability of regional features through the interaction of regional features in group data.

iv) GIRL obtains the top-1 accuracy of H = 47.8%, which exceeds the best current result H = 46.3% by 1.5% on APY. RGEN [19] and DPPN [15] are most relevant to our approach, which are all devoted to mining discriminative region knowledge. Our method has 10.6% and 3.5% improvements, respectively. GIRL improves the localization of fine-grained features through the interaction of regions between different images.

v) For a fair comparison with the APN [12] and DPPN [15], we present the results without fine-tuning the ResNet-101. On all four datasets, our H scores outperform methods that do not rely on external information (e.g. APN [12], DPPN [15], AREN [13], and RGEN [19]). This result demonstrates that the mining of information from group data can also facilitate the localization of fine-grained attribute features without external information.

vi) RGEN [19] constructs region interactions in a single image. Our H score is superior to RGEN [19] on all four datasets. This can be attributed to the fact that GIRL has a tendency to capture additional and crucial semantic contrastive cues by utilizing region interaction in group data compared to RGEN. [19]. Such cues are important for ZSL to boost the discriminability and transferability of local features.

Zero-Shot Learning In Table 2, we also record the comparison results of GIRL with other state-of-the-art methods in the ZSL setting. We can conclude the following from Table 2:

i) For the coarse-grained dataset AWA2, compared to RGEN [19], GIRL obtains a subtle gain of over 0.3%. DPPN [15], TransZero [17], and MSDN [16] are relevant to GIRL, compared to these methods, GIRL demonstrates respective improvements of 5.0%, 3.8%, and 3.9%. GIRL focus more on the interaction between regions in different samples, and this interaction can provide more discriminative information to the model.

ii) For the fine-grained dataset SUN, our method obtains a competitive performance, yielding the best performance of 66.5%, outperforming the current best method MSDN [16] by 0.7%. This result demonstrates that our method can learn discriminative region features and ensures that the learned knowledge can be transferred to unseen classes.

iii) On CUB, GIRL achieves slightly worse performance than TransZero, which employs attribute description vectors. However, GIRL leverages its own information to extract discriminative features. On APY, GIRL outperforms existing region-based methods, demonstrating its effectiveness in learning discriminative local regions.

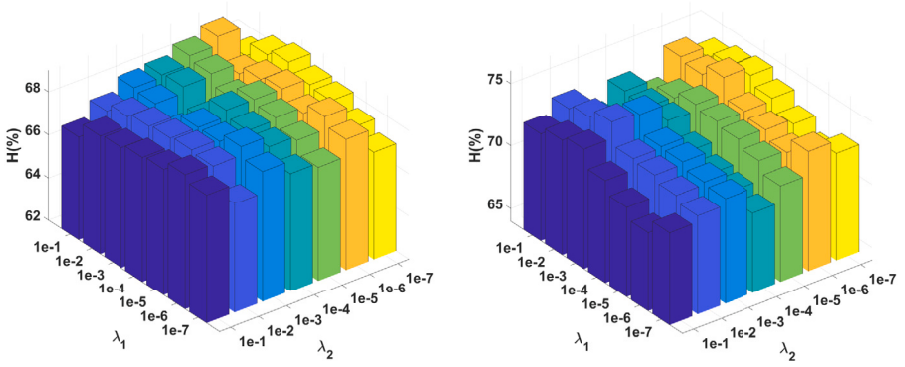
4.5. Ablation study

Effects of N_{cl} and N_{sam}. In GIRL, we randomly select N_{cl} classes for a batch, and each class is sampled with N_{sam} images. In Fig. 3, $(N_{cl}, N_{sam}) = (16, 4)$ represents that we randomly selected 16 classes, and each class randomly selected 4 samples in a batch. The best H is obtained at $(16, 2)$ and $(32, 2)$ for CUB and AWA2, respectively.

Effects of λ_1 and λ_2. In GIRL, λ_1 and λ_2 adjust the parameters of $\mathcal{L}_{\text{scl}}$ and $\mathcal{L}_{\text{ral}}$. For CUB and AWA2, the λ_1 is within $\{1e{-}7, 1e{-}6, 1e{-}5, 1e{-}4, 1e{-}3, 1e{-}2, 1e{-}1\}$, and the λ_2 is within $\{1e{-}7, 1e{-}6, 1e{-}5, 1e{-}4, 1e{-}3, 1e{-}2, 1e{-}1\}$. $\mathcal{L}_{\text{ar}}$ and $\mathcal{L}_{\text{ce}}$ are more important in the overall loss function, so for $\mathcal{L}_{\text{scl}}$ and $\mathcal{L}_{\text{ral}}$, we adjust the parameters starting from $\lambda_1 = 1e{-}7$. In Fig. 4, for CUB and AWA2, the overall trend is relatively stable. GIRL gets the best results at $\lambda_1 = 1e{-}1$, $\lambda_2 = 1e{-}6$ on CUB. This shows that the loss $\mathcal{L}_{\text{scl}}$ is important with respect to $\mathcal{L}_{\text{ral}}$. The $\mathcal{L}_{\text{scl}}$ distill the distinct semantic information from the refined region features to the original region features, boosting the model's localization of precise attribute regions. For AWA2, GIRL gets the best results at $\lambda_1 = 1e{-}3$, $\lambda_2 = 1e{-}6$. Similarly, this result proves that $\mathcal{L}_{\text{scl}}$ is more important and the loss of $\mathcal{L}_{\text{ral}}$ alleviates the bias problem to some extent.

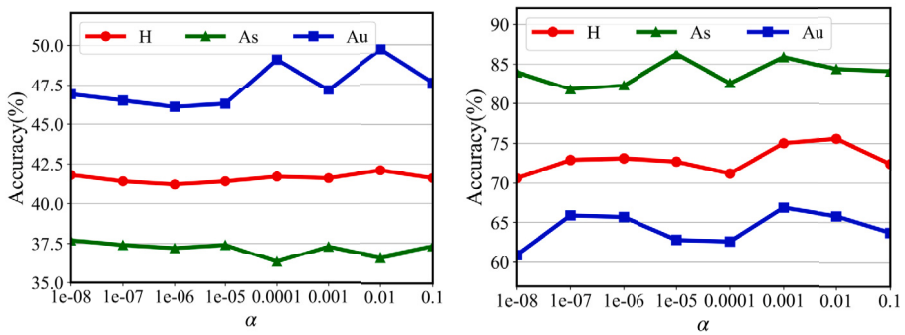
Effects of α. In the dif-attention-based Interaction (DAI) module, α reflects the importance of the feature $\Upsilon(x_i) \otimes R(x_i)$ in region feature refinement. We show the As, Au, and H results when the value is within $\{1e{-}8, 1e{-}7, 1e{-}6, 1e{-}5, 1e{-}4, 1e{-}3, 1e{-}2, 1e{-}1\}$ under GZSL setting on SUN and AWA2. Fig. 5 shows that a relatively large α can get improved accuracy. For SUN and AWA2, $\alpha = 1e{-}2$ is suitable to yield satisfactory results and the overall change trend is stable.

Component Analysis. To make it easier to grasp the mechanism of GIRL, we performed a component analysis to verify the importance of each individual component. The ARI module, which consists of the DAI and CAI modules, is a part of GIRL. We also



(a) CUB (b) AWA2

Fig. 4. Effects of λ_1 and λ_2 on CUB and AWA2 in terms of H.



(a) SUN (b) AWA2

Fig. 5. Effects of α on SUN and AWA2, in terms of H, As, Au.

Table 3
Component analysis on SUN, AWA2 and APY.

Methods	SUN			AWA2			APY		
	Au	As	H	Au	As	H	Au	As	H
baseline	44.0	33.9	38.3	60.9	84.0	70.6	36.7	58.6	45.2
GIRL w/o ARI	45.1	33.9	38.7	62.3	83.4	71.3	37.3	59.0	45.7
GIRL w/o DAI	44.9	35.4	39.6	61.5	86.5	71.9	38.1	59.0	46.3
GIRL w/o CAI	48.3	36.1	41.7	63.8	83.4	72.3	35.5	65.7	46.1
GIRL w/o L_{scl}	49.5	35.6	41.4	64.6	83.9	73.0	38.2	57.3	45.9
GIRL w/o L^2_{scl}	47.6	36.2	41.1	67.1	81.7	73.6	35.5	65.7	46.1
GIRL w/o L^1_{scl}	49.5	36.0	41.7	65.7	85.1	74.1	34.5	65.7	46.0
GIRL w/o L_{ral}	47.6	37.3	41.8	66.9	85.1	74.8	35.6	67.6	46.6
GIRL (full)	47.9	38.8	**42.9**	68.5	84.3	**75.5**	37.0	67.6	**47.8**

confirm the significance of L_{scl}, L^1_{scl}, L^2_{scl}, and L_{ral} in the loss function. The baseline is the model of GIRL without all components mentioned in Table 3, the H value decreases by 4.6%, 4.9%, and 2.6% compared to GIRL (full) on SUN, AWA2, and APY. GIRL without ARI drops by 4.2%, 4.2%, and 2.1%. ARI can improve the model's localization of fine-grained discriminative region features through the interaction of the region in the group. The DAI and CAI modules can enhance feature discriminative by learning semantic commonalities and differences among regions. L_{scl} plays a key role in GIRL, aiming to distill discriminative semantic knowledge from refined region features to original ones. Assuming no L_{scl}, the H value would drop by 1.5%, 2.5%, and 1.9% on SUN, AWA2, and APY. L^2_{scl} has a greater impact on the overall model relative to L^1_{scl}, due to the importance of different region knowledge for the localization of discriminative region features. L_{ral} has a greater impact on the overall model, GIRL without L_{ral} decreases by 1.1%, 0.7%, and 1.9% compared to its full model on SUN, AWA2, and APY.

Group Construction. In this section, we discuss how to perform group construction in a batch. A group should contain samples that semantically complement each other, and for the selection of these samples, we explore different strategies:

Table 4
Effects of group construction on SUN, AWA2 and APY.

Methods	CUB			AWA2			APY		
	Au	As	H	Au	As	H	Au	As	H
RS	64.7	70.4	67.3	63.1	84.1	72.1	34.5	65.7	45.3
CRS	64.5	71.7	67.9	67.1	84.3	73.7	36.2	65.7	46.7
SCSS	66.2	71.9	**69.0**	68.5	84.3	**75.6**	37.0	67.6	**47.8**

Table 5
Results of GZSL with input size 224×224 on CUB, SUN, and AWA2. The best result is bolded.

Methods	CUB			SUN			AWA2		
	Au	As	H	Au	As	H	Au	As	H
DAZLE [14]	56.7	59.6	58.1	52.3	24.3	33.2	60.3	75.7	67.1
AREN [13]	38.9	78.7	52.1	19.0	38.8	25.5	15.6	92.9	26.7
LFGAA [41]	36.2	80.9	50.0	18.5	40.0	25.3	27.0	93.4	41.9
RGEN [19]	60.0	73.5	66.1	44.0	31.7	36.8	67.1	76.5	**71.5**
APN [12]	65.3	69.3	67.2	41.9	34.0	37.6	57.1	72.4	63.9
GIRL	65.8	69.1	**67.4**	49.5	32.4	**39.2**	59.0	80.4	68.0

Table 6
Results of GZSL under the setting of without post-processing on CUB, SUN, and AWA2.

Methods	CUB			SUN			AWA2		
	Au	As	H	Au	As	H	Au	As	H
SJE [34]	23.5	59.2	33.6	14.7	30.5	19.8	8.0	73.9	14.4
SYNC [8]	11.5	70.9	19.8	7.9	43.3	13.4	10.0	90.5	18.0
DAZLE [14]	42.0	65.3	51.1	21.7	31.9	25.8	25.7	82.5	39.2
AREN [13]	38.9	78.7	52.1	19.0	38.8	25.5	15.6	92.9	26.7
LFGAA [41]	36.2	80.9	50.0	18.5	40.0	25.3	27.0	93.4	41.9
GIRL	47.9	80.5	**60.1**	18.9	43.4	**26.3**	31.8	95.1	**47.7**

i) Random selection within a batch (RS). We randomly select three samples as a group.

ii) Class-wise random selection (CRS). For sample $R(x)$, samples are randomly selected among intra/inter classes to form the group.

iii) Selection by class and sample similarity (SCSS). For sample $R(x)$, the most similar samples among intra/inter classes are randomly selected to form a group. We have performed experimental validation on CUB and AWA2.

As can be seen from the Table 4, first, the group construction using the third strategy can achieve better results. The common information between the most similar samples is more obvious, which contribute to feature refinement. The differences between the most similar samples of different classes can train GIRL to get obvious subtle differences and leverage these differences to refine the original region features. Second, when constructing groups within a batch, the time complexity of group construction with the third strategy is $O(b^2)$, and the time complexity of class-wise random selection is $O(b)$, b is the batch size. Considering the time complexity and accuracy together, we choose the third strategy as our group construction technique.

Results of Input Size 224×224**.** In GIRL, we follow the settings of current state-of-the-art methods DPPN [15], TransZero [17], MSDN [16], etc., with an input image size of 448×448. For a fair comparison, the input size is modified to 224×224 in GIRL. In Table 5, GIRL outperforms the current methods by 0.2% and 1.6% on the two fine-grained datasets CUB and SUN, and achieves the second best result on AWA2. The experimental results demonstrate the effectiveness of our method.

Comparison without Post-Processing. Some current methods utilize post-processing, e.g. calibration stacking [46] or domain detector [47] to alleviate the domain bias problem. In this paper, we follow DPPN [15] to alleviate the domain bias problem using post-processing. For a fair comparison with other methods, in this section, we further do some experiments, i.e., GIRL compares with other methods without the post-processing. In Table 6, our method outperforms the current state-of-the-art methods by 8%, 0.5%, and 5.8% on CUB, SUN, and AWA2. This result shows that our method can enhance fine-grained attribute localization in zero-shot learning through inter-region interactions, and enhance the transferability of semantic knowledge.

Results of with Finetuning. In ZSL, finetuning ResNet-101 [44] is a solution to alleviate the domain bias problem. DPPN has achieved better results with finetuning ResNet-101 [44]. We give the results of GIRL without fine-tuning in Table 1. The GZSL results of SUN and AWA2 are still superior to the DPPN with fine-tuning by 1.9% and 2.5%. In Table 7, the GIRL results with finetuned ResNet-101 superior to DPPN on SUN and AWA2. It is superior to the existing method by 2.1% on SUN and by 2.6% on AWA2 than DPPN. Our algorithm is slightly lower than DPPN on CUB, but from the Table 1, our algorithm is significantly superior to DPPN without finetuning the ResNet-101 on CUB.

Table 7
The results of GIRL with Finetuning on CUB, SUN, and AWA2. The best result is bolded.

Methods	CUB			SUN			AWA2		
	Au	As	H	Au	As	H	Au	As	H
RGEN [19]	60.0	73.5	66.1	44.0	31.7	36.8	67.1	76.5	71.5
DPPN [15]	70.2	77.1	**73.5**	47.9	35.8	41.0	63.1	86.8	73.1
GIRL	69.8	76.6	73.0	46.9	39.8	**43.1**	68.0	85.3	**75.7**

Table 8
GZSL results using soft spatial attention for region feature extraction on CUB, SUN, and AWA2.

Methods	CUB			SUN			AWA2		
	Au	As	H	Au	As	H	Au	As	H
SJE [34]	23.5	59.2	33.6	14.7	30.5	19.8	8.0	73.9	14.4
SYNC [8]	11.5	70.9	19.8	7.9	43.3	13.4	10.0	90.5	18.0
AREN [13]	38.9	78.7	52.1	19.0	38.8	25.5	15.6	92.9	26.7
GIRL	47.9	80.5	**60.1**	18.9	43.4	**26.3**	31.8	95.1	**31.9**

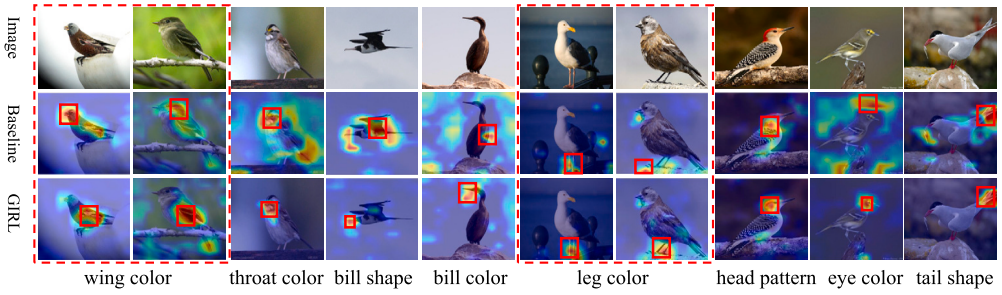


Fig. 6. Visualization of attribute localization for the baseline and GIRL to prove the necessity of region interaction in the group on CUB.

The region feature extraction methods. We extract the region feature attribute features following the DPPN [15]. We follow AREN [13] to extract attribute features using soft spatial attention. In Table 8, the GIRL result is superior to AREN. On the fine-grained datasets CUB and SUN, GIRL exceeds AREN by 8% and 0.8%, respectively. On the coarse-grained dataset AWA2, it outperforms AREN by 5.2%. The above results show that comprehensive and explicit region interaction among a group of images from different categories can enhance the discrimination of region features and thus lead to a desirable knowledge transfer between seen and unseen classes.

4.6. Qualitative results

To intuitively exhibit the advantages of region interaction in groups for learning fine-grained attribute features, we visualize the results of attribute localization at different models. In Fig. 6, the comparison between baseline (mentioned in the **Component Analysis**) and GIRL demonstrate the importance of region interaction for attribute localization. Our method (GIRL) is accurate and fine for attribute localization compared to baseline, further boosting the discriminability of local features.

To further demonstrate GIRL can boost the discriminability of local features, we visualize the results of fine-grained attribute localization under different models. In Fig. 7, A (GIRL without DAI) and B (GIRL without CAI) show the importance of differential attention-based and common attention-based regional interactions on attribute localization, respectively. GIRL is accurate in terms of fine-grained attribute localization compared to models A and B.

In Fig. 8, we display the t-SNE [48] visualization of unseen class features for APN [12], TransZero [17], DPPN [15], and our GIRL on AWA2. The distribution of unseen classes derived by GIRL has more distinct class boundaries than the other three methods. It demonstrates GIRL can transfer semantic knowledge to unseen classes and guarantee the class separability of unseen classes through the interaction of regions in a group.

5. Conclusion

In this paper, we propose group-wise interactive region learning (GIRL) for tackling ZSL and GZSL. GIRL consists of attentive region interaction (ARI) module and holistic semantic embedding (HSE) module. ARI captures the semantic differences and commonalities of the corresponding regions by region interaction in a group data, and then refines original region features. HSE holistically maps these region features to the semantic space for a more stable semantic transfer. We introduce semantic consistency loss to promote semantic distillation from refined region features to original region features. The relational alignment loss

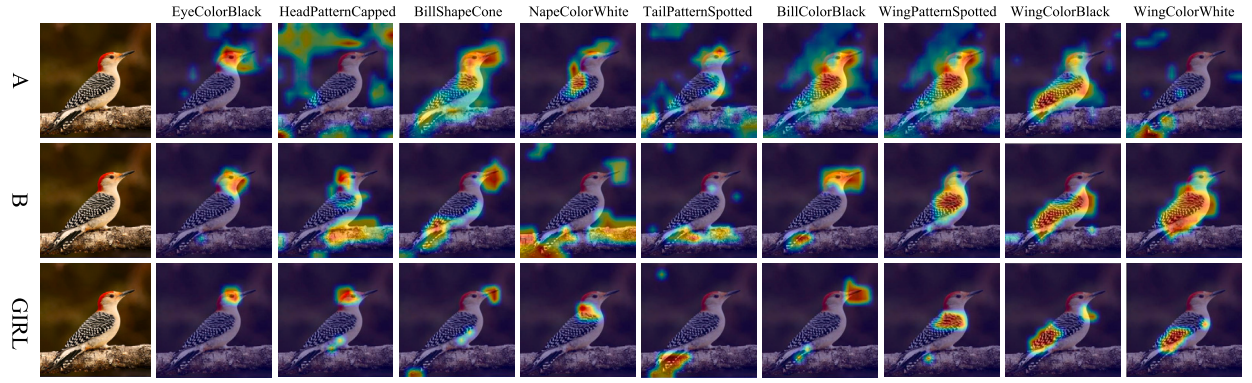
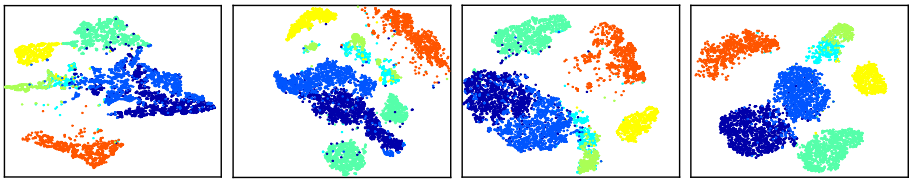


Fig. 7. Visualization of attribute localization on A (GIRL without DAI), B (GIRL without CAI), and GIRL to verify the importance of each module in ARI, further prove the necessity of the interactive region learning among a group of images from different categories.



(a) APN (b) TransZero (c) DPPN (d) **GIRL (Ours)**

Fig. 8. t-SNE of unseen test images for APN [12], TransZero [17], DPPN [15] and GIRL on AWA2.

further introduces unseen class semantic vectors into the training. For ZSL and GZSL, GIRL yields some state-of-the-art results. ARI being a plug-and-play module, in future work, we will introduce this idea of group data interaction to other tasks e.g. fine-grained classification and zero-shot segmentation.

CRediT authorship contribution statement

Ting Guo: Conceptualization, Methodology, Software, Visualization, Writing – original draft. **Jiye Liang:** Methodology, Supervision. **Guo-Sen Xie:** Methodology, Project administration, Writing – review & editing.

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.

Data availability

I have shared the link to my code at the Attach File step.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. U21A20473, No. 62276134, No. 62276158).

References

[1] C.H. Lampert, H. Nickisch, S. Harmeling, Learning to detect unseen object classes by between-class attribute transfer, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 951–958.
[2] G.-S. Xie, X.-Y. Zhang, Y. Yao, Z. Zhang, F. Zhao, L. Shao, Vman: a virtual mainstay alignment network for transductive zero-shot learning, IEEE Trans. Image Process. 30 (2021) 4316–4329.
[3] M.R. Zarei, M. Taheri, Y. Long, Kernelized distance learning for zero-shot recognition, Inf. Sci. 580 (2021) 801–818.
[4] P. Zhao, S. Zhang, J. Liu, H. Liu, Zero-shot learning via the fusion of generation and embedding for image recognition, Inf. Sci. 578 (2021) 831–847.
[5] C. Lampert, H. Nickisch, S. Harmeling, Attribute-based classification for zero-shot visual object categorization, IEEE Trans. Pattern Anal. Mach. Intell. 36 (3) (2014) 453–465.
[6] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems, 2013, pp. 3111–3119.

[7] B. Romera-Paredes, P.H.S. Torr, An embarrassingly simple approach to zero-shot learning, in: International Conference on Machine Learning, 2015, pp. 2152–2161.
[8] S. Changpinyo, W.-L. Chao, B. Gong, F. Sha, Synthesized classifiers for zero-shot learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5327–5336.
[9] E. Schonfeld, S. Ebrahimi, S. Sinha, T. Darrell, Z. Akata, Generalized zero- and few-shot learning via aligned variational autoencoders, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 8247–8255.
[10] K. Gukyeong, R.G. Al, A gating model for bias calibration in generalized zero-shot learning, IEEE Trans. Image Process. (2022) 1–12.
[11] Y. Li, Z. Liu, X. Chang, J. McAuley, L. Yao, Diversity-boosted generalization-specialization balancing for zero-shot learning, IEEE Trans. Multimed. (2023) 1–11.
[12] W. Xu, Y. Xian, J. Wang, B. Schiele, Z. Akata, Attribute prototype network for zero-shot learning, in: Advances in Neural Information Processing Systems, 2020.
[13] G. Xie, L. Liu, X. Jin, F. Zhu, Z. Zhang, J. Qin, Y. Yao, L. Shao, Attentive region embedding network for zero-shot learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 9384–9393.
[14] D. Huynh, E. Elhamifar, Fine-grained generalized zero-shot learning via dense attribute-based attention, in: IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 4482–4492.
[15] C. Wang, S. Min, X. Chen, X. Sun, H. Li, Dual progressive prototype network for generalized zero-shot learning, in: Advances in Neural Information Processing Systems, 2021, pp. 2936–2948.
[16] S. Chen, Z. Hong, G. Xie, W. Wang, Q. Peng, K. Wang, J. Zhao, X. You, MSDN: mutually semantic distillation network for zero-shot learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2022, pp. 7602–7611.
[17] S. Chen, Z. Hong, Y. Liu, G. Xie, B. Sun, H. Li, Q. Peng, K. Lu, X. You, Transzero: attribute-guided transformer for zero-shot learning, in: The AAAI Conference on Artificial Intelligence, 2022, pp. 330–338.
[18] Y. Liu, Y. Dang, X. Gao, J. Han, L. Shao, Zero-shot learning with attentive region embedding and enhanced semantics, IEEE Trans. Neural Netw. Learn. Syst. (2022) 1–12.
[19] G. Xie, L. Liu, F. Zhu, F. Zhao, Z. Zhang, Y. Yao, J. Qin, L. Shao, Region graph embedding network for zero-shot learning, in: European Conference on Computer Vision, 2020, pp. 562–580.
[20] Y. Xian, Z. Akata, G. Sharma, Q.N. Nguyen, M. Hein, B. Schiele, Latent embeddings for zero-shot classification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 69–77.
[21] Z. Akata, F. Perronnin, Z. Harchaoui, C. Schmid, Label-embedding for image classification, IEEE Trans. Pattern Anal. Mach. Intell. 38 (7) (2016) 1425–1438.
[22] Y. Xian, T. Lorenz, B. Schiele, Z. Akata, Feature generating networks for zero-shot learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5542–5551.
[23] Z. Han, Z. Fu, S. Chen, J. Yang, Contrastive embedding for generalized zero-shot learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 2371–2381.
[24] H. Liu, L. Yao, Q. Zheng, M. Luo, H. Zhao, Y. Lyu, Dual-stream generative adversarial networks for distributionally robust zero-shot learning, Inf. Sci. 519 (2020) 407–422.
[25] X. Kong, Z. Gao, X. Li, M. Hong, J. Liu, C. Wang, Y. Xie, Y. Qu, En-compactness: self-distillation embedding & contrastive generation for generalized zero-shot learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2022, pp. 9296–9305.
[26] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A.C. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.
[27] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, arXiv:1503.02531, 2015.
[28] T. Xu, C. Liu, Data-distortion guided self-distillation for deep neural networks, in: The AAAI Conference on Artificial Intelligence, AAAI Press, 2019, pp. 5565–5572.
[29] Y. Zhang, T. Xiang, T.M. Hospedales, H. Lu, Deep mutual learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4320–4328.
[30] Z. Zhang, M.R. Sabuncu, Self-distillation as instance-specific label smoothing, in: Advances in Neural Information Processing Systems, 2020.
[31] C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie, The Caltech-UCSD Birds-200-2011 Dataset, Tech. rep., 2011.
[32] G. Patterson, J. Hays, SUN attribute database: discovering, annotating, and recognizing scene attributes, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2751–2758.
[33] A. Farhadi, I. Endres, D. Hoiem, D.A. Forsyth, Describing objects by their attributes, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1778–1785.
[34] Z. Akata, S.E. Reed, D. Walter, H. Lee, B. Schiele, Evaluation of output embeddings for fine-grained image classification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2927–2936.
[35] J. Sánchez, M. Molina, Trading-off information modalities in zero-shot classification, in: IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 1677–1685.
[36] R. Felix, B.G.V. Kumar, I.D. Reid, G. Carneiro, Multi-modal cycle-consistent generalized zero-shot learning, in: European Conference on Computer Vision, 2018, pp. 21–37.
[37] Y. Xian, S. Sharma, B. Schiele, Z. Akata, F-VAEGAN-D2: a feature generating framework for any-shot learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10275–10284.
[38] X. Zhao, Y. Shen, S. Wang, H. Zhang, Boosting generative zero-shot learning by synthesizing diverse features with attribute augmentation, in: The AAAI Conference on Artificial Intelligence, 2022, pp. 3454–3462.
[39] Y. Feng, X. Huang, P. Yang, J. Yu, J. Sang, Non-generative generalized zero-shot learning via task-correlated disentanglement and controllable samples synthesis, in: IEEE Conference on Computer Vision and Pattern Recognition, 2022, pp. 9336–9345.
[40] Y. Yang, X. Zhang, M. Yang, C. Deng, Adaptive bias-aware feature generation for generalized zero-shot learning, IEEE Trans. Multimed. 25 (2023) 280–290.
[41] Y. Liu, J. Guo, D. Cai, X. He, Attribute attention for semantic disambiguation in zero-shot learning, in: International Conference on Computer Vision, 2019, pp. 6697–6706.
[42] M.F. Naeem, Y. Xian, L.V. Gool, F. Tombari, I2dformer: learning image to document attention for zero-shot image classification, in: Advances in Neural Information Processing Systems, 2022.
[43] Y. Li, Z. Liu, L. Yao, X. Wang, J.J. McAuley, X. Chang, An entropy-guided reinforced partial convolutional network for zero-shot learning, IEEE Trans. Circuits Syst. Video Technol. 32 (8) (2022) 5175–5186.
[44] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[45] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, in: International Conference on Learning Representations, 2015.
[46] W. Chao, S. Changpinyo, B. Gong, F. Sha, An empirical study and analysis of generalized zero-shot learning for object recognition in the wild, in: European Conference on Computer Vision, vol. 9906, 2016, pp. 52–68.
[47] S. Min, H. Yao, H. Xie, C. Wang, Z. Zha, Y. Zhang, Domain-aware visual bias eliminating for generalized zero-shot learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 12661–12670.
[48] V.D.M. Laurens, G. Hinton, Visualizing data using t-sne, J. Mach. Learn. Res. 9 (2605) (2008) 2579–2605.