SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

July 2024, Vol. 67, Iss. 7, 172103:1–172103:13
https://doi.org/10.1007/s11432-023-4051-9

Rethinking attribute localization for zero-shot learning

Shuhuang CHEN[1], Shiming CHEN[1*], Guo-Sen XIE[2], Xiangbo SHU[2], Xinge YOU[1] & Xuelong LI[3]

[1]*School of Electronic Information and Communication, Huazhong University of Science and Technology, Wuhan* 430074*, China;*
[2]*School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing* 210094*, China;*
[3]*School of Artificial Intelligence, Optics and Electronics, Northwestern Polytechnical University, Xi'an* 710072*, China*

Received 28 March 2023/Revised 31 August 2023/Accepted 5 December 2023/Published online 25 June 2024

Abstract Recent advancements in attribute localization have showcased its potential in discovering the intrinsic semantic knowledge for visual feature representations, thereby facilitating significant visual-semantic interactions essential for zero-shot learning (ZSL). However, the majority of existing attribute localization methods heavily rely on classification constraints, resulting in accurate localization of only a few attributes while neglecting the rest important attributes associated with other classes. This limitation hinders the discovery of the intrinsic semantic relationships between attributes and visual features across all classes. To address this problem, we propose a novel attribute localization refinement (ALR) module designed to enhance the model's ability to accurately localize all attributes. Essentially, we enhance weak discriminant attributes by grouping them and introduce weighted attribute regression to standardize the mapping values of semantic attributes. This module can be flexibly combined with existing attribute localization methods. Our experiments show that when combined with the ALR module, the localization errors in existing methods are corrected, and state-of-the-art classification performance is achieved.

Keywords zero-shot learning, attention mechanism, attribute localization, image classification

1 Introduction

Humans can identify unseen classes by leveraging prior knowledge gained from seen classes and using attributes shared across the two disjoint class sets. Zero-shot learning (ZSL), inspired by human cognition, aims to identify unseen classes through shared semantic knowledge between seen and unseen classes [1–3]. Depending on the class space explored in the testing phase, ZSL can be categorized into conventional and generalized settings [4–7]. In conventional ZSL (CZSL), the test phase exclusively involves samples from unseen classes. Meanwhile, in generalized ZSL (GZSL), the test samples encompass both seen and unseen classes.

Traditional approaches to associating semantic knowledge and visual features typically involve learning the mapping relationship between global visual features and semantic vectors in seen classes. However, these methods have major drawbacks. For instance, a semantic vector composed of attributes can be decoupled, and each individual attribute can be associated with a specific visual region. Similarly, discriminant information exists across different local regions, which global visual features cannot capture adequately, failing to reflect discriminant local information. Although methods based on local representation [8–11] aim to identify discriminative regions within images, they struggle to establish an association between semantic knowledge and visual features at the attribute level, resulting in inferior performances in ZSL.

To address the need for an accurate association between visual features and semantic knowledge at the attribute level, attribute localization has been proposed [12–17]. Unlike traditional methods that align global visual feature vectors with corresponding class semantic vectors, attribute localization approaches

* Corresponding author (email: gchenshiming@gmail.com)

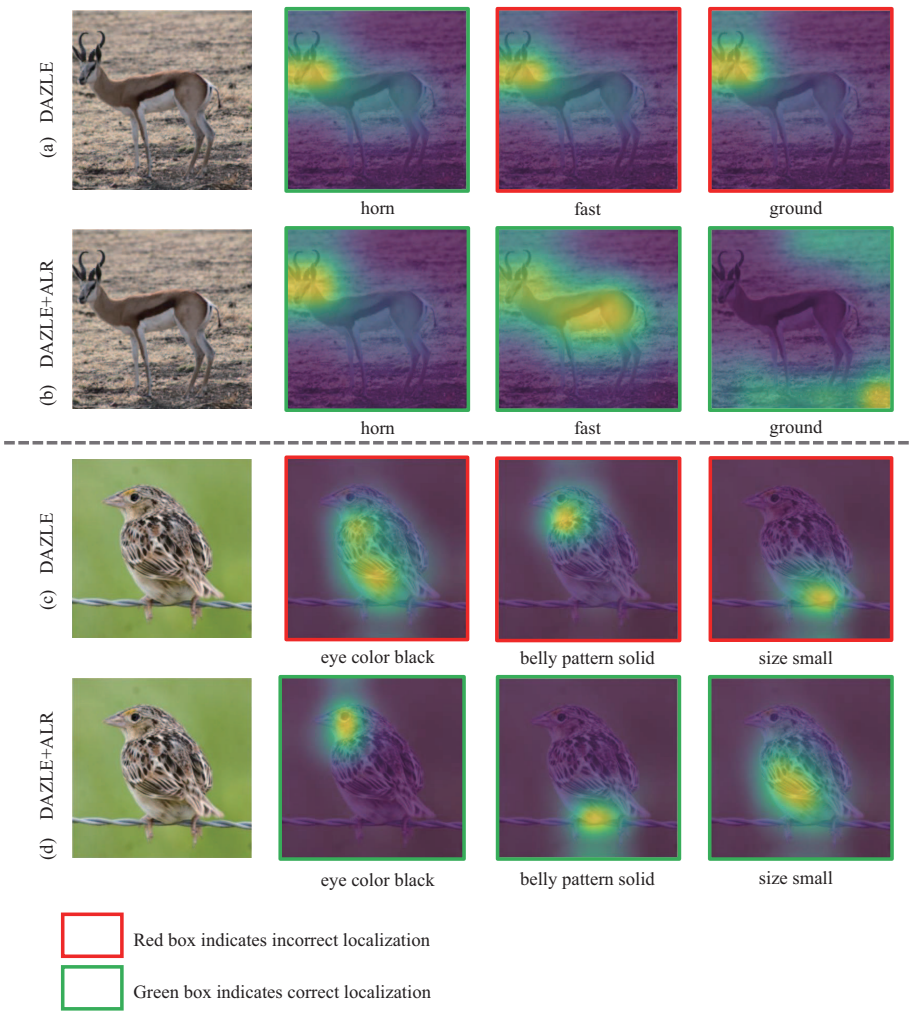

Figure 1 (Color online) Motivation illustration. (a) and (c) Existing method DAZLE [13] that relies on the classification constraints localizes the weak discriminant attributes (e.g., "ground", "size small") to incorrect regions. (b) and (d) Our ALR module enhances the weak discriminant attributes so that the model can focus on all attributes to achieve correct localization. The samples are selected from AWA2 [18] and CUB [19], and words below the images are attributes.

decouple semantic vectors and global visual features into attribute vectors and regional visual features, respectively. Attribute localization maps the attributes to their corresponding visual regions by associating attribute vectors with regional visual features, thus facilitating visual-semantic associations at the attribute dimension. However, existing attribute localization methods heavily rely on classification constraints. This results in only a few attributes being correctly localized while other important attributes for different classes are neglected. Consequently, these methods fall short of fully uncovering the intrinsic semantic knowledge between visual and attribute features required for effective visual-semantic interactions. We will further discuss this issue by visualizing the attribute localization of an existing method on the ZSL benchmark datasets AWA2 [18] and CUB [19].

On the coarse-grained dataset (e.g., AWA2), the visual features among different classes are quite different, allowing classification to be primarily based on the head region and its corresponding strong discriminant attributes in the seen classes. Therefore, as shown in Figure 1(a), all attributes are localized within the head region that contains strong discriminant information. A similar challenge arises with the fine-grained dataset (e.g., CUB). Here, most attributes can be correctly localized because category determination requires the analysis of multiple regions and corresponding attributes. However, some weak discriminant attributes (those that appear in all or none of the seen classes) tend to be overlooked by the model, leading to incorrect attribute localization for these attributes. As shown in Figure 1(b), weak discriminant attributes (e.g., "eye color black") are localized in the incorrect regions. Such errors in attribute localization hinder effective visual-semantic interactions and diminish classification performance

in ZSL.

To enable ZSL models to accurately localize weak discriminant attributes, we introduce an attribute localization refinement (ALR) module. The ALR module consists of two main components: an attribute visual grouping constraint and an information entropy-based attribute regression constraint. The attribute visual grouping constraint organizes attributes according to their visual interrelations, guiding the model to focus on weak discriminant attributes by constraining the localized regions through group relationships. Meanwhile, the information entropy-based attribute regression constraint aims to further standardize the mapping value of semantic attributes. The ALR module can be flexibly plugged into existing methods to improve attribute localization accuracy. Qualitative results show that, when combined with the ALR module, existing methods achieve precise attribute localization. Extensive experiments also show that combined with the ALR module, existing attribute localization methods achieve state-of-the-art classification results across three ZSL benchmarks.

Our main contributions are the following.

• We have thoroughly reviewed and analyzed attribute localization for ZSL and have introduced the ALR module as a solution to the limitations faced by existing attribute localization methods. The module has a simple design and can be flexibly plugged into existing methods to improve attribute localization.

• The ALR module has a well-designed attribute visual grouping constraint and a weighted attribute regression constraint. These elements are crucial in directing the model's attention to weak discriminant attributes and standardizing attribute prediction values.

• Experiments show that our ALR module can solve existing problems in attribute localization and achieve correct attribute location, thus ensuring accurate visual-semantic interactions.

2 Related work

2.1 Zero-shot learning

ZSL aims to classify unseen classes using semantic knowledge transferred from seen classes [1, 3, 20–24]. The core principle involves establishing a mapping between the visual and semantic domains for seen classes and then transferring this semantic knowledge to classify unseen categories [5]. ZSL methods can be broadly categorized into two types based on their approach to visual and semantic domain interactions: embedding methods (visual to semantic mapping) [25–31] and generative methods (semantic to visual mapping) [32–40]. Although these methods have made some progress, they rely predominantly on aligning semantic vectors with global visual features to establish a visual-semantic association. However, this global mapping strategy fails to accurately capture these local attribute-region correspondences.

Therefore, some research in ZSL has shifted focus to local representation. Initial explorations of local representation employed object detection methods, which utilize a trained part detector to extract a fixed number of parts [8,10]. These approaches, however, are constrained by the need for additional and expensive annotation data required by part detectors. To better mine local features, the attention mechanism has been introduced to ZSL as an alternative to traditional object detection methods aimed at identifying local discriminative regions [9, 11, 41–43]. These semantic-guided attention methods attempt to discover limited discriminative visual regions using semantic vectors as a guide. Nonetheless, whether employing object detection or semantic-guided attention mechanisms, these strategies merely replace global visual representation focused on limited discriminative regions without ensuring a reliable association between attributes and their corresponding discriminant regions.

2.2 Attribute localization

To achieve an accurate association between visual features and semantic knowledge on an attribute dimension, the concept of attribute localization is proposed [12–15]. Unlike methods based on global alignment or local representation, attribute localization focuses on identifying each attribute-related region and utilizing the regional visual features corresponding to these attributes to facilitate visual-semantic interactions at the attribute level. Instead of aligning the global visual feature with a semantic vector, DAZLE [13] aligns each attribute-based regional visual feature with the attribute vector using an attribute embedding technique. Xu et al. [15] highlighted the importance of attribute localization in their research and introduced the APN model, which jointly learns discriminative global and local features using an attribute prototype vector. GEM [14] adds an attention transition module based on

visual-semantic attention, converting the attribute attention map into a gaze map, thereby mimicking the human gaze mechanism to improve attribute localization capabilities. TransZero [12] replaces the traditional attention structure by using the transformer structure [44] to perform attribute localization. These methods improve the accuracy of ZSL by implementing attribute localization. However, a lack of in-depth analysis and optimization in attribute localization leads to deficiencies in attribute localization. To this end, we analyze the limitations inherent in attribute localization and propose the ALR module as a solution.

3 Methodology

Notation. Suppose we have two sets of classes, $\mathcal{C}^s$ for the set of seen classes and $\mathcal{C}^u$ for the set of unseen classes. Samples corresponding to these classes are defined as $\boldsymbol{x} \in \mathcal{X}^s$ for seen classes and $\boldsymbol{x} \in \mathcal{X}^u$ for unseen classes, with their respective labels being $y \in \mathcal{Y}^s$ for seen and $y \in \mathcal{Y}^u$ for unseen classes. More specifically, let $\boldsymbol{z}^c = [z_1^c, \ldots, z_A^c]^{\mathrm{T}} = \phi(y^c)$ denote the semantic vector of class c comprising A attributes, where z_a^c represents the score with the a-th attribute in class c. Semantic vectors for all classes are available during both training and testing phases. The training set $\mathcal{D}^{\text{train}} = \{(\boldsymbol{x}, y)\}$ consists of samples from seen classes and their corresponding labels. Depending on the test settings in the testing set, ZSL is divided into two categories: CZSL and GZSL. The goal of CZSL is to predict image labels from unseen classes, i.e., $\boldsymbol{x} \in \mathcal{X}^u$ in the testing set $\mathcal{D}^{\text{test}} = \{\boldsymbol{x}\}$; conversely, GZSL aims to predict image labels from both seen and unseen classes, i.e., $\boldsymbol{x} \in \mathcal{X}^s \cup \mathcal{X}^u$ in the testing set $\mathcal{D}^{\text{test}} = \{\boldsymbol{x}\}$.

In addition, attribute localization methods typically use a pre-trained natural language model (e.g., the Glove model [45] trained on Wikipedia articles) to extract the attribute semantic vectors $\{\boldsymbol{v}_a\}_{a=1}^A$, which can be fine-tuned during the training process. Meanwhile, the input image i is divided into R grid cell regions of equal size, with the features of these regions being extracted using ResNet-101 [46] pre-trained on ImageNet [47]. Specifically, i^r denotes region r ($r \in R$) in input image i, and $\boldsymbol{f}_i^r = \boldsymbol{f}(i_r)$ denotes the feature of i^r.

3.1 Revisiting attribute localization

Existing attribute localization methods localize attributes to related visual regions and facilitate visual-semantic interactions at the attribute level by mapping regional visual features to attributes. These methods decompose the semantic vector into A attribute vectors (or attribute prototypes in APN [15]) and divide the image into R regions. Then, the attention method is used to identify visual regions most relevant to the attributes and generate an attribute attention map, thereby realizing attribute localization. Referring to [13], we revisit the attribute localization method.

First, for each attribute, the correlation between the attribute vectors and the regional visual features is calculated through the attention mechanism, resulting in the generation of the attribute attention map $\boldsymbol{m}_a = [m_a^1, \ldots, m_a^r, \ldots, m_a^R]^{\mathrm{T}}$, which can be formulated as

$$m_a^r = \text{softmax}\left(\boldsymbol{v}_a^{\mathrm{T}} \boldsymbol{W}_1 \boldsymbol{f}_i^r\right) = \frac{\exp\left(\boldsymbol{v}_a^{\mathrm{T}} \boldsymbol{W}_1 \boldsymbol{f}_i^r\right)}{\sum_{r'=1}^{R} \exp\left(\boldsymbol{v}_a^{\mathrm{T}} \boldsymbol{W}_1 \boldsymbol{f}_i^{r'}\right)}, \tag{1}$$

where $\boldsymbol{W}_1$ denotes a learnable matrix to measure the compatibility between each attribute vector and the visual feature of each region. The attribute attention map obtained here can also be used to visualize attribute localization.

After obtaining the attribute attention map, the attribute-based attention feature of the a-th attribute can be further computed:

$$\boldsymbol{h}_i^a = \sum_{r=1}^{R} m_a^r \boldsymbol{f}_i^r. \tag{2}$$

$\boldsymbol{h}_i^a$ represents the most relevant visual feature of the image i to the a-th attribute.

Next, the objective is to map the visual features into semantic vectors $\boldsymbol{z}_i = [z_i^1, \ldots, z_i^A]^{\mathrm{T}}$ and calculate the class score s_i^c (the probability that image i belongs to the class c). z_i^a which is considered the confidence of the a-th attribute in the image i can be obtained by calculating the correlation between $\boldsymbol{h}_i^a$ and $\boldsymbol{v}_a$:

$$z_i^a = \boldsymbol{v}_a^{\mathrm{T}} \boldsymbol{W}_2 \boldsymbol{h}_i^a, \tag{3}$$

where $\boldsymbol{W}_2$ denotes a learnable matrix. The confidence of each attribute is calculated to obtain the semantic vector $\boldsymbol{z}_i$. After obtaining the semantic vector $\boldsymbol{z}_i$, the class score is derived by computing the similarity between $\boldsymbol{z}_i$ and the semantic vector of each class. This process can be formulated as

$$s_i^c = \boldsymbol{z}_i \cdot \boldsymbol{z}^c. \tag{4}$$

Finally, the model employs a typical classification loss (e.g., cross-entropy loss) to ensure the image achieves the highest score with its corresponding class semantic vector. The model parameters are optimized by minimizing the cross-entropy loss between the prediction and the ground-truth label across the training set:

$$\mathcal{L}_{\mathrm{CE}}\left(\{s_i^c\}_{c\in\mathcal{C}^s\cup\mathcal{C}^u}\right) = -\sum_i \log p\left(s_i^{y_i}\right), \tag{5}$$

where $p\left(s_i^c\right)$ is the probability that image i belongs to class c, calculated by applying softmax to class scores s_i^c:

$$p\left(s_i^c\right) = \frac{\exp\left(s_i^c\right)}{\sum_{c'\in\mathcal{C}^s\cup\mathcal{C}^u}\exp(s_i^{c'})}. \tag{6}$$

Upon reviewing existing methods, it is evident that the model can achieve attribute localization by using the attention mechanism to assess the correlation between attributes and different regions. However, the optimization of model parameters depends mainly on the classification loss in the training set, which includes only seen classes. This approach renders the model highly sensitive to attributes with strong discriminative power in seen classes while ignoring the localization of other important attributes for unseen classes.

3.2 Attribute localization refinement

In this subsection, we introduce our ALR module. As shown in Figure 2, this module has an attribute visual grouping constraint and weighted attribute regression. The attribute visual grouping constraint enhances the model's attention to weak discriminant attributes by grouping them, thus enabling the model to accurately localize all attributes. Meanwhile, the weighted attribute regression, informed by attribute information entropy, standardizes the mapping value of semantic attributes. Our designed modules can be plugged into existing attribute localization methods.

3.2.1 *Attribute visual grouping constraint*

As shown in Figure 3, there are three relationships between the regions corresponding to the attributes: disjoint, joint, and overlap. Depending on the visual relationship, we classify overlapping attributes into the same group. For example, in dataset CUB, we organize some attributes into the head region group (e.g., "eye color black," "crown color red"), tail region group (e.g., "tail shape notched," "tail color black"), and entirety region group (e.g., "size small," "primary color buff"). The disjoint localization regions corresponding to attributes from different groups (e.g., head region group and tail region group) are constrained to separate, while the localization regions corresponding to overlapping attributes from the same group are encouraged to aggregate. However, for joint attributes from different groups (e.g., head region group and entirety region group), measuring the degree of joint relationship is challenging, so no constraints are imposed on these regions.

In existing methods, attribute attention maps are subjected to a softmax operation, i.e., $\sum_{r=1}^{R} m_a^r = 1$. The relationship between two attributes can be measured and calculated by the dot product between two attribute attention maps:

$$C_{a_1a_2} = \boldsymbol{m}_{a_1} \cdot \boldsymbol{m}_{a_2}. \tag{7}$$

When the localization regions of two attributes are disjoint, their attention maps exhibit different distributions, causing the value of the product $C_{a_1a_2}$ to be small, ideally approaching 0. On the contrary, when two attributes overlap, they share similar attention maps, and thus, the value of the product $C_{a_1a_2}$ is large, ideally approaching 1.

Therefore, to facilitate separation and aggregation, we constrain the value of the product $C_{a_1a_2}$ by grouping attributes and establishing an attribute-relationship guiding matrix $\boldsymbol{R}$. $\boldsymbol{R} = [R_{11}, \ldots, R_{1A}; \ldots; R_{A1}, \ldots, R_{AA}]$ is constructed according to the attribute grouping. The value of $R_{a_ia_j}$ equals 1, 0, and -1 when the a_i-th and a_j-th attributes are disjoint, joint, and overlap, respectively.

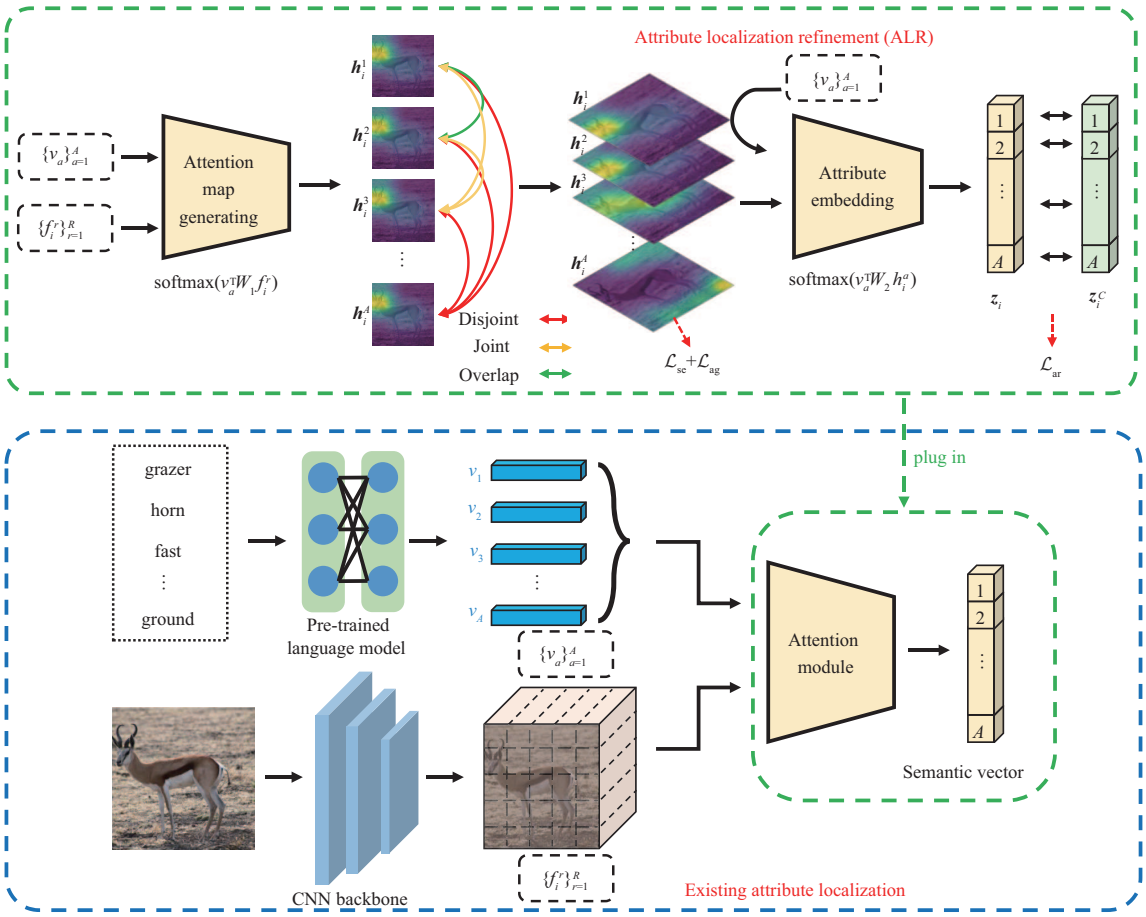

Figure 2 (Color online) Overview of our proposed method. The lower part is the existing attribute localization method, and the upper part is our pluggable module ALR.

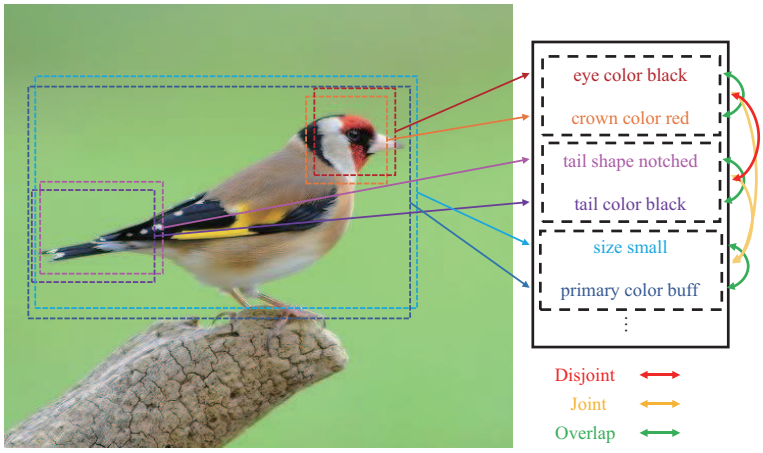

Figure 3 (Color online) Three types of relationship between visual regions corresponding to attributes: disjoint, joint, and overlap.

To separate localization regions corresponding to disjoint attributes, we can minimize the attention map product C calculated between disjoint attributes using the guiding matrix $\boldsymbol{R}$. Therefore, we design a separation loss $\mathcal{L}_{\rm se}$, which is formulated as follows:

$$\mathcal{L}_{\rm se} = \sum_{a_1=1}^{A} \sum_{a_2=1}^{A} R_{a_1 a_2} C_{a_1 a_2} \mathbb{I}_{[R_{a_1 a_2}]}, \tag{8}$$

where $\mathbb{I}_{[R_{a_1a_2}]}$ is an indicator function. When the a_1-th attribute and a_2-th attribute are disjoint, $R_{a_1a_2} = 1$ and $\mathbb{I}_{[R_{a_1a_2}]} = 1$, otherwise $\mathbb{I}_{[R_{a_1a_2}]} = 0$. Therefore, we can constrain disjoint attributes to achieve separation through separation loss.

Similarly, to facilitate the aggregation of overlapping localization regions, we aim to maximize the attention map product C calculated between overlapping attributes, i.e., minimize $-C$. Therefore, we introduce the aggregation loss $\mathcal{L}_{\mathrm{ag}}$, which can be formulated as

$$\mathcal{L}_{\mathrm{ag}} = \sum_{a_1=1}^{A} \sum_{a_2=1}^{A} \left(1 + R_{a_1a_2} C_{a_1a_2}\right) \mathbb{I}_{[-R_{a_1a_2}]}. \tag{9}$$

When the a_1-th attribute and a_2-th attribute overlap, $R_{a_1a_2} = -1$ and $\mathbb{I}_{[-R_{a_1a_2}]} = 1$, otherwise $\mathbb{I}_{[-R_{a_1a_2}]} = 0$. Therefore, we can achieve the aggregation of overlapping attributes by minimizing the aggregation loss (i.e., $1 - C_{a_1a_2}$).

Unlike overlap attributes and disjoint attributes, the joint regions between different joint attributes cannot be determined; thus, there are no additional constraints imposed on joint attributes.

Through the design of attribute grouping, separation loss, and aggregation loss, we can further refine attribute localization.

3.2.2 *Information entropy-based attribute regression constraint*

To further guide the model toward accurate visual-to-semantic embedding, we introduce a weighted attribute regression loss. Directly using the strong constraint of mean square error on all attributes easily weakens the classification ability of the model. Therefore, we assess the overall importance of attributes in classification by calculating information entropy, thereby realizing the distinction constraints on attributes.

The information entropy of the attribute a can be calculated by

$$H\left(z_a\right) = \sum_{i=1}^{c\epsilon\left(\mathcal{C}^s \cup \mathcal{C}^u\right)} P\left(z_a^i\right) \log P\left(z_a^i\right), \tag{10}$$

and $P(z_a^i) = \frac{z_a^i}{\sum_{c \in \mathcal{C}^s \cup \mathcal{C}^u} z_a^c}$ is the probability function of attribute a of class i.

Upon obtaining the information entropy of the a-th attribute, its importance to classification can be ranked (i.e., $R_a = \mathrm{Rank}(H\left(z_a\right))$). The coefficient of an attribute is further determined by the rank of its importance:

$$\lambda_a = \mathrm{e}^{-\frac{R_a}{A}}. \tag{11}$$

Thus, we can obtain the coefficient vector of the attributes $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_A]^{\mathrm{T}}$, and the final regression loss $\mathcal{L}_{\mathrm{ar}}$ is formulated as

$$\mathcal{L}_{\mathrm{ar}} = \left\|\boldsymbol{\lambda}\left(\psi\left(\boldsymbol{x}_i\right) - \boldsymbol{z}^c\right)\right\|_2^2 = \left\|\boldsymbol{\lambda}\left(\boldsymbol{z}_i - \boldsymbol{z}^c\right)\right\|_2^2, \tag{12}$$

where $\psi\left(\boldsymbol{x}_i\right)$ is overall the mapping function from image i to semantic vector $\boldsymbol{z}_i$.

Through the implementation of attribute visual grouping and attribute regression constraints, we can effectively refine attribute positioning.

3.3 Optimization and zero-shot prediction

To optimize the model, we need to minimize the final loss function that combines cross-entropy loss, separation loss, aggregation loss, and attribute regression loss:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{CE}} + \beta_{\mathrm{se}}\mathcal{L}_{\mathrm{se}} + \beta_{\mathrm{ag}}\mathcal{L}_{\mathrm{ag}} + \beta_{\mathrm{ar}}\mathcal{L}_{\mathrm{ar}}, \tag{13}$$

over the parameters of the attention model and attribute semantic vectors. Where β_{se}, β_{ag}, and β_{ar} are the weights to control their corresponding loss terms.

After training the complete model, we can first obtain the embedding features $\psi\left(\boldsymbol{x}_i\right)$ of the test sample $\boldsymbol{x}_i$. To predict the class of the test sample, we calculate the maximum class similarity score:

$$c^* = \arg\max_{c \in \mathcal{C}^u} \psi\left(\boldsymbol{x}_i\right) \boldsymbol{z}^c. \tag{14}$$

Table 1 Statistics of three benchmark datasets used in our experiments

Dateset	Attribute dimension	Seen class	Seen sample		Unseen class	Unseen sample
			Train for CZSL&GZSL	Test for GZSL		Test for CZSL&GZSL
AWA2	85	40	23527	5882	10	7913
CUB	312	150	7057	1764	50	2967
SUN	102	645	10320	2580	72	1440

In the GZSL setting, test images can originate from either seen or unseen classes. However, since only seen classes are available during the training phase, GZSL predictions tend to be biased toward seen classes [5, 48]. To address this issue, most existing visual semantic attention methods apply calibration to predict the test label (e.g., DAZLE [13]). Specifically, the GZSL classifier is defined as

$$c^* = \arg \max_{c \in \mathcal{C}^u \cup \mathcal{C}^s} \left(\psi \left(\boldsymbol{x}_i \right) \boldsymbol{z}^c + \mathbb{I}_{[c \in \mathcal{C}^u]} \right), \tag{15}$$

where $\mathbb{I}_{[c \in \mathcal{C}^u]}$ is an indicator function (i.e., it is 1 when $c \in \mathcal{C}^u$, otherwise -1).

4 Experiments

This section presents the datasets, evaluation protocol, implementation details, experimental results, ablative analysis, visualization of qualitative results, and hyperparameter analysis.

4.1 Datasets

We evaluate our method on three popular zero-shot learning benchmark datasets, including one coarse-grained dataset (AWA2 [18]) and two fine-grained datasets (CUB [19] and SUN [49]). Among them, AWA2 contains 37322 images from 50 animal categories with 85 attributes; CUB comprises 11788 images from 200 bird categories with 312 attributes; SUN consists of 14340 images from 717 scene classes with 102 attributes. We adopt the proposed split [5] to divide each class into seen and unseen classes on each dataset and report the dataset statistics in Table 1.

4.2 Evaluation protocol

We measure the performance of ZSL by testing the average per-class top-1 accuracy. For the CZSL setting, we predict unseen classes by computing the test sample accuracy, denoted by acc. In the GZSL setting, since the testing set consists of both seen and unseen samples, we need to evaluate top-1 accuracy in both seen (S) and unseen (U) classes, respectively. We then calculate their harmonic mean (defined as $H = (2 \times S \times U)/(S + U)$ [18]) to assess the performance of GZSL.

4.3 Implementation details

Our work focuses on improving the attribute localization of existing methods and proposes a pluggable module to improve the accuracy of these methods. To demonstrate the efficacy and versatility of the ALR module across different attribute localization methods, we conducted evaluations using DAZLE [13] and TransZero [12]. We kept most of the experimental settings from the original publications, implementing necessary adjustments, such as fine-tuning the loss function hyperparameters and replacing the attribute regression loss in TransZero with an information entropy-based attribute regression loss. Both DAZLE and TransZero employ ResNet101, pre-trained on ImageNet, as the CNN backbone for extracting the feature map without additional training. In DAZLE, the input image is resized to 224×224 pixels, yielding a $7 \times 7 \times 2048$ feature map from the last convolutional layer, which corresponds to a set of features from 7×7 regions. In TransZero, the input image is resized to 448×448 pixels, resulting in a feature map size of $14 \times 14 \times 2048$ in the last convolutional layer, corresponding to a set of features from 14×14 regions.

Based on the visual relationships among attributes in the AWA2, CUB, and SUN datasets, we conducted attribute grouping and established an attribute-relationship guiding matrix $\boldsymbol{R}$. For the CUB dataset, attributes can be directly divided into 8 groups based on head, belly, breast, wing, tail, leg, and entirety. The attributes within the first 7 groups are disjoint, while the 8th group intersects with the first 7 groups. As for the AWA2 dataset, attributes can be grouped according to different types of attributes,

Table 2 Results (%) of the state-of-the-art CZSL and GZSL modes on AWA2, CUB, and SUN datasets[a)]. The first part refers to embedding methods, the second part covers generative methods, the third part includes local representation methods, and the last part focuses on attribute localization methods

Method	AWA2				CUB				SUN			
	CZSL	GZSL			CZSL	GZSL			CZSL	GZSL		
	acc	*U*	*S*	*H*	acc	*U*	*S*	*H*	acc	*U*	*S*	*H*
SP-AEN (CVPR'18) [26]	58.8	23.3	*90.9*	37.1	55.4	34.7	70.6	46.6	59.2	24.9	38.6	30.3
IIR (ICCV'19) [25]	67.9	17.6	87.0	28.9	63.8	30.4	65.8	41.2	63.5	22.0	34.1	26.7
CE-GZSL (CVPR'21) [50]	70.4	63.1	78.6	*70.0*	*77.5*	63.9	66.8	65.3	63.3	48.8	38.6	*43.1*
TDCSS (CVPR'22) [51]	–	59.2	74.9	66.1	–	44.2	62.8	51.9	–	–	–	–
f-CLSWGAN (CVPR'18) [34]	68.2	57.9	61.4	59.6	57.3	43.7	57.7	49.7	60.8	42.6	36.6	39.4
f-VAEGAN (CVPR'19) [35]	*71.1*	57.6	70.6	63.5	61.0	48.4	60.1	53.6	64.7	45.1	38.0	41.3
HSVA (NIPS'21) [52]	–	59.3	76.6	66.8	62.8	52.7	58.3	55.3	63.8	48.6	**39.0**	**43.3**
MSDN (CVPR'22) [53]	70.1	62.0	74.5	67.7	76.1	68.7	67.5	68.1	65.8	52.2	34.2	41.3
SGMA (NIPS'19) [42]*	68.8	37.6	87.1	52.5	71.0	36.7	*71.3*	48.5	–	–	–	–
AREN (CVPR'19) [41]*	67.9	15.6	**92.9**	26.7	71.8	38.9	**78.7**	52.1	60.6	19.0	*38.8*	25.5
LFGAA (ICCV'19) [43]	68.1	27.0	93.4	41.9	67.6	36.2	80.9	50.0	61.5	18.5	40.0	25.3
APN (NIPS'20) [15]*	68.4	56.5	78.0	65.5	72.0	65.3	69.3	67.2	61.6	41.9	34.0	37.6
DAZLE (CVPR'20) [13]	67.9	60.3	75.7	67.1	66.0	56.7	59.6	58.1	59.4	52.3	24.3	33.2
DAZLE+ALR	70.3	*63.5*	75.8	69.1	68.9	63.6	57.5	60.4	61.9	47.1	27.8	35.0
TransZero (AAAI'22) [12]	70.1	61.3	82.3	70.2	76.8	*69.3*	68.3	*68.8*	*65.5*	*52.6*	33.4	40.8
TransZero+ALR	**71.2**	61.6	82.3	**70.5**	**78.8**	**70.4**	69.0	**69.7**	**66.2**	**52.7**	34.0	41.3

a) The best and the second-best results are in bold and italic, respectively. The symbol '–' indicates no available results. The symbol '*' denotes end-to-end methods. The underline indicates that methods enhanced with the ALR module achieve better performance.

such as color, texture, shape, body parts, habits, and environment. Different attribute groups within the same type are disjoint, like different colors, 'white' and 'black,' while attributes from different types are joint. Similarly, for the SUN dataset, attributes are grouped according to function, characteristics, spatial environment, and local object. On average, it only takes 6 min to complete the grouping of a dataset.

4.4 Comparison with state-of-the-art methods

We evaluated the proposed ALR module on DAZLE and TransZero models to demonstrate its effectiveness and compatibility across different attribute localization frameworks. In addition to comparing it with current popular attribute localization methods, we benchmarked it against state-of-the-art (SoTA) embedding models, generative models, and visual attention methods. Baseline results were obtained using the official code.

Table 2 [12,13,15,25,26,34,35,41–43,50–53] demonstrates the effectiveness of the proposed ALR module when applied to DAZLE and TransZero, leading to improved performance (i.e., DAZLE+ALR and TransZero+ALR), thus showcasing its flexible plug-in function. Notably, TransZero+ALR outperforms SoTA, highlighting the ability of ALR to enhance classification performance through optimized attribute localization. However, we observed that the performance improvement provided by our module on the SUN dataset is less significant than on other datasets. The main reason is that the relationships between visual regions corresponding to attributes are mostly joint; therefore, the optimization of attribute localization and the advantages of our module are limited.

4.5 Ablation study

To gain further insight into the ALR module, we conducted experiments to evaluate the effects of the attribute visual grouping constraint (i.e., $\mathcal{L}_{\rm se}+\mathcal{L}_{\rm ag}$) and the information entropy-based attribute regression constraint (i.e., $\mathcal{L}_{\rm ar}$). As shown in Table 3, the attribute visual grouping constraint improves the acc/H over the baseline (DAZLE) by 1.9%/1.1% on AWA2 and 1.0%/1.0% on CUB. When only the information entropy-based attribute regression constraint is applied, the acc/H improved by 0.3%/1.2% on AWA2 and 0.2%/0.7% on CUB. The complete ALR module improves acc/H by 2.4%/2.0% on AWA2 and 2.9%/2.3% on CUB.

Table 3 Results (%) of CZSL and GZSL ablation studies on AWA2 and CUB. The baseline is DAZLE. We analyze the performance of each component of our module

Method	AWA2				CUB			
	acc	U	S	H	acc	U	S	H
baseline	67.9	60.3	75.7	67.1	66.0	56.7	59.6	58.1
baseline+ALR ($\mathcal{L}_{\rm se}+\mathcal{L}_{\rm ag}$)	69.8	63.0	74.3	68.2	67.0	55.0	63.9	59.1
baseline+ALR ($\mathcal{L}_{\rm ar}$)	68.2	61.3	76.9	68.3	66.2	57.0	60.9	58.8
baseline+ALR ($\mathcal{L}_{\rm se}+\mathcal{L}_{\rm ag}+\mathcal{L}_{\rm ar}$)	70.3	63.5	75.8	69.1	68.9	63.7	57.4	60.4

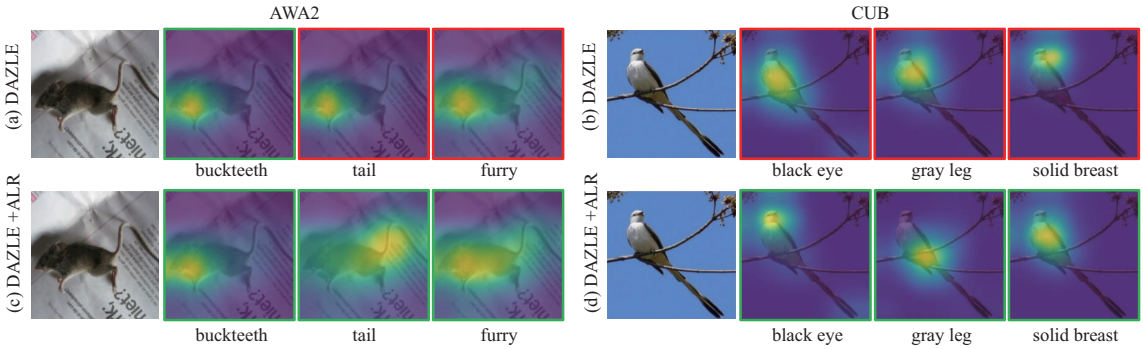

Figure 4 (Color online) Attribute localization results. Attribute attention map for (a) and (b) existing method DAZLE and (c) and (d) method with our ALR module on AWA2 and CUB. The image with the red box indicates incorrect localization, and the image with the green box indicates correct localization.

4.6 Qualitative results

To further demonstrate the effectiveness of the ALR module in improving attribute localization and representation compared to existing methods, we present attribute localization and t-SNE visualizations.

4.6.1 *Visualization of attribute localization*

To demonstrate improvements in attribute localization capabilities brought by the ALR module, we visualized the attribute attention map of the DAZLE and DAZLE+ALR models. As illustrated in Figure 1, the original DAZLE model occasionally fails to accurately localize attributes with weak discrimination attributes (for example, the "ground" attribute in AWA2 and the "eye color black" attribute in CUB). However, when combined with the ALR module, the attribute localization ability is significantly improved, with attributes being accurately localized to their corresponding regions. Moreover, we conducted extensive visualization experiments (shown in Figure 4) to further validate the effectiveness of our module in mitigating the limitations posed by existing attribute localization methods.

4.6.2 *t-SNE visualization*

To demonstrate the impact of attribute localization optimization on improving visual feature representation and facilitating knowledge transfer, we present t-SNE [54] visualizations of visual features from unseen classes selected from the CUB and AWA2 datasets. These features were learned by the ResNet101 backbone, DAZLE, and DAZLE+ALR. Figure 5 demonstrates that the visual features extracted by DAZLE show marked improvements over those extracted by the CNN backbone. Furthermore, the quality of visual features is further improved with the integration of the ALR module. These results highlight the effectiveness of attribute localization optimization in enhancing feature representation and knowledge transfer.

4.7 Hyperparameter analysis

Our work centers on three pivotal factors: the weights of separation loss, aggregation loss, and attribute regression loss, denoted by $\beta_{\rm se}$, $\beta_{\rm ag}$, and $\beta_{\rm ar}$, respectively. To determine their effect on visual-semantic interactions, we conducted experiments using the baseline DAZLE enhanced with the ALR module. We experimented with a wide range of values for $\beta_{\rm se}$, $\beta_{\rm ag}$, and $\beta_{\rm ar}$ and evaluated their performance on CUB and AWA2. The results, shown in Figure 6, reveal that excessively high or low values of $\beta_{\rm se}$, $\beta_{\rm ag}$, and $\beta_{\rm ar}$

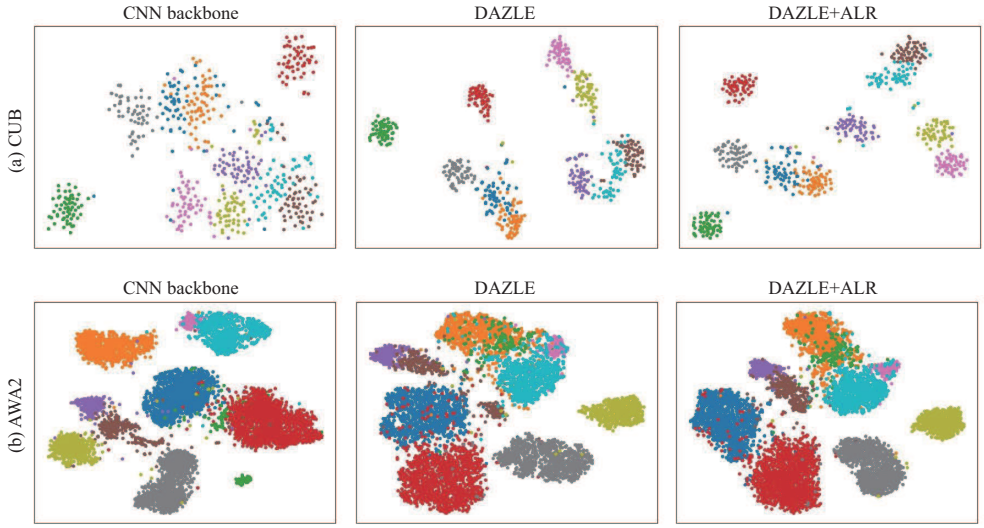

Figure 5 (Color online) t-SNE visualizations of visual features for unseen classes selected from (a) CUB and (b) AWA2, learned by the CNN backbone, DAZLE, and DAZLE+ALR. The 10 colors represent 10 different unseen classes selected from CUB and AWA2.

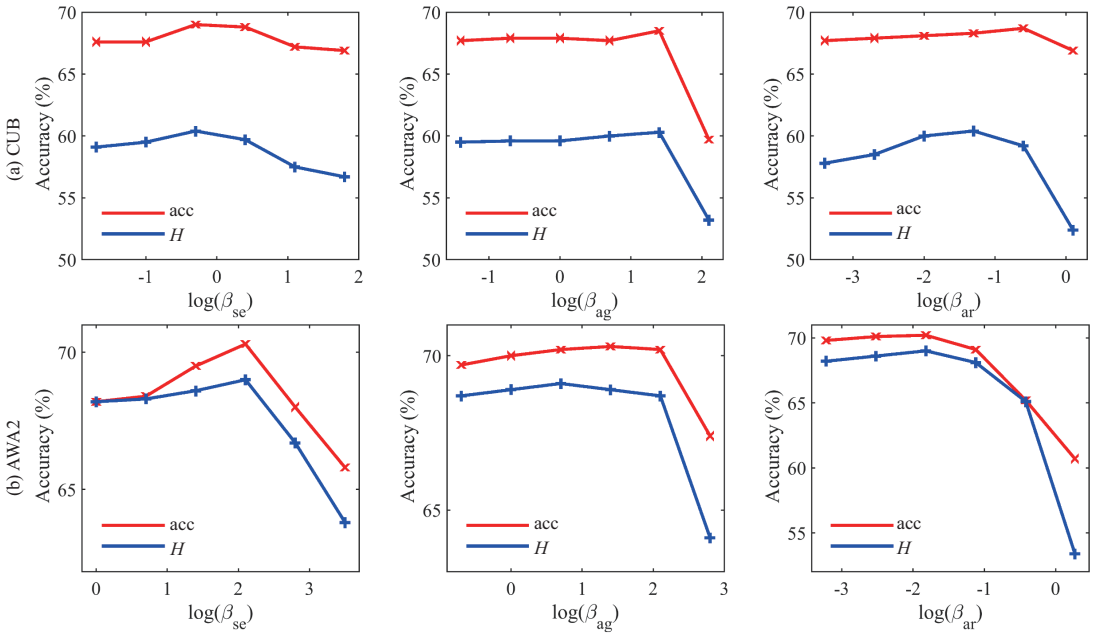

Figure 6 (Color online) Effects of $\beta_{\rm se}$, $\beta_{\rm ag}$, and $\beta_{\rm ar}$ on (a) CUB and (b) AWA2.

tend to diminish the evaluation metrics. This decline can be attributed to the fact that high loss values diminish the relative contribution of the classification loss, while low values lessen their overall impact. Optimal results were achieved when we set the combination coefficients of $\beta_{\rm se}$, $\beta_{\rm ag}$, and $\beta_{\rm ar}$ to (125, 5, 0.015) and (0.5, 25, 0.25) on CUB and AWA2, respectively.

4.8 Generality analysis

To show the generality of our ALR, we extracted feature representations from DAZLE enhanced with ALR (i.e., $\sum_{a=1}^{A} \boldsymbol{h}_i^a$, the sum of found visual features most relevant to the attributes) and applied them to popular generative models, such as f-CLSWGAN [34] and TF-VAEGAN [55]. The results, presented in Table 4, show that our ALR module effectively boosts the performances of these generative models on two datasets. Specifically, ALR improved the performance of f-CLSWGAN in terms of acc/H by 7.1%/6.1% on CUB and 1.5%/2.7% on AWA2. Similarly, improvements of 4.2%/2.4% on CUB and 1.3%/2.4% on

Table 4 Results of various generative models with visual features extracted from ALR

Methods	CUB				AWA2			
	CZSL	GZSL			CZSL	GZSL		
	acc	Au	As	Ah	acc	Au	As	Ah
f-CLSWGAN [34]	57.3	43.7	57.7	49.7	68.2	57.9	61.4	59.6
f-CLSWGAN [34] + ALR	64.4	50.3	62.7	55.8	69.7	62.4	62.1	62.3
TF-VAEGAN [55]	64.9	52.8	64.7	58.1	72.2	59.8	75.1	66.6
TF-VAEGAN [55] + ALR	69.1	59.1	62.0	60.5	73.5	58.7	83.5	69.0

AWA2 were achieved for TF-VAEGAN with the integration of LAR. These results indicate that our ALR can discover semantically relevant visual features, helping generative models to better generate transferable visual features.

5 Conclusion

In this paper, we illustrate that traditional attribute localization methods rely on classification constraints in seen classes, which leads to neglecting the accurate localization of some important attributes for unseen classes. Furthermore, to accurately localize all attributes, we propose the ALR module, which enhances the weak discriminant attributes by grouping them and implements weighted attribute regression to standardize the mapping values of semantic attributes. This module can be flexibly plugged into current attribute localization methods to enhance attribute localization capabilities, thereby improving classification performance. Our experiments across several widely used datasets demonstrated the effectiveness of the proposed module. When used in conjunction with existing methods, it helps mitigate localization errors and achieve SoTA results.

Acknowledgements This work was supported by National Key R&D Program of China (Grant No. 2022YFC3301000).

References

1 Lampert C H, Nickisch H, Harmeling S. Learning to detect unseen object classes by betweenclass attribute transfer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009. 951–958
2 Lampert C H, Nickisch H, Harmeling S. Attribute-based classification for zero-shot visual object categorization. IEEE Trans Pattern Anal Mach Intell, 2013, 36: 453–465
3 Palatucci M, Pomerleau D, Hinton G E, et al. Zero-shot learning with semantic output codes. In: Proceedings of Advances in Neural Information Processing Systems, 2009
4 Pourpanah F, Abdar M, Luo Y X, et al. A review of generalized zero-shot learning methods. IEEE Trans Pattern Anal Mach Intell, 2023, 45: 4051–4070
5 Xian Y Q, Schiele B, Akata Z. Zero-shot learning-the good, the bad and the ugly. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 4582–4591
6 Chen Z, Luo Y D, Qiu R H, et al. Semantics disentangling for generalized zero-shot learning. In: Proceedings of the IEEE International Conference on Computer Vision, 2021. 8692–8700
7 Ye Y L, He Y K, Pan T J, et al. Alleviating domain shift via discriminative learning for generalized zero-shot learning. IEEE Trans Multimedia, 2022, 24: 1325–1337
8 Elhoseiny M, Zhu Y Z, Zhang H, et al. Link the head to the "beak": zero shot learning from noisy text description at part precision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017
9 Xie G S, Liu L, Zhu F, et al. Region graph embedding network for zero-shot learning. In: Proceedings of the European Conference on Computer Vision, 2020. 562–580
10 Yang S Q, Wang K, Herranz L, et al. On implicit attribute localization for generalized zero-shot learning. IEEE Signal Process Lett, 2021, 28: 872–876
11 Yu Y L, Ji Z, Fu Y W, et al. Stacked semantics-guided attention model for fine-grained zero-shot learning. In: Proceedings of Advances in Neural Information Processing Systems, 2018
12 Chen S M, Hong Z M, Liu Y, et al. TransZero: attribute-guided transformer for zero-shot learning. In: Proceedings of Association for the Advancement of Artificial Intelligence, 2022
13 Huynh D, Elhamifar E. Fine-grained generalized zero-shot learning via dense attribute-based attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020. 4483–4493
14 Liu Y, Zhou L, Bai X, et al. Goal-oriented gaze estimation for zero-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021. 3794–3803
15 Xu W J, Xian Y Q, Wang J N, et al. Attribute prototype network for zero-shot learning. In: Proceedings of Advances in Neural Information Processing Systems, 2020. 21969–21980
16 Chen S M, Hong Z M, Hou W J, et al. TransZero++: cross attribute-guided transformer for zero-shot learning. IEEE Trans Pattern Anal Mach Intell, 2023, 45: 12844–12861
17 Chen S M, Hong Z M, Xie G S, et al. GNDAN: graph navigated dual attention network for zero-shot learning. IEEE Trans Neural Netw Learn Syst, 2024, 35: 4516–4529
18 Xian Y Q, Lampert C H, Schiele B, et al. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. IEEE Trans Pattern Anal Mach Intell, 2018, 41: 2251–2265
19 Wah C, Branson S, Welinder P, et al. The caltech-UCSD birds-200-2011 dataset. 2011. https://authors.library.caltech.edu/27452/

20 Fu Y W, Xiang T, Jiang Y G, et al. Recent advances in zero-shot recognition: toward data-efficient understanding of visual content. IEEE Signal Process Mag, 2018, 35: 112–125
21 Ji Z, Wang H R, Yu Y L, et al. A decadal survey of zero-shot image classification (in Chinese). Sci Sin Inform, 2019, 49: 1299–1320
22 Feng Y G, Yu J, Sang J T, et al. Survey on knowledge-based zero-shot visual recognition (in Chinese). J Software, 2021, 32: 370–405
23 Wang W, Zheng V W, Yu H, et al. A survey of zero-shot learning. ACM Trans Intell Syst Technol, 2019, 10: 1–37
24 Ji Z, Wang Q, Cui B Y, et al. A semi-supervised zero-shot image classification method based on soft-target. Neural Networks, 2021, 143: 88–96
25 Cacheux Y L, Borgne H L, Crucianu M. Modeling inter and intra-class relations in the triplet loss for zero-shot learning. In: Proceedings of the IEEE International Conference on Computer Vision, 2019. 10333–10342
26 Chen L, Zhang H W, Xiao J, et al. Zero-shot visual recognition using semantics-preserving adversarial embedding networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018. 1043–1052
27 Liu S C, Long M S, Wang J M, et al. Generalized zero-shot learning with deep calibration network. In: Proceedings of Advances in Neural Information Processing Systems, 2018
28 Romera-Paredes B, Torr P. An embarrassingly simple approach to zero-shot learning. In: Proceedings of International Conference on Machine Learning, Lille, 2015. 2152–2161
29 Zhang L, Xiang T, Gong S G. Learning a deep embedding model for zero-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017
30 Yu Y L, Ji Z, Guo J C, et al. Zero-shot learning via latent space encoding. IEEE Trans Cybern, 2019, 49: 3755–3766
31 Kong X, Gao Z D, Li X F, et al. En-compactness: self-distillation embedding & contrastive generation for generalized zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022
32 Chen S M, Wang W J, Xia B H, et al. FREE: feature re-finement for generalized zero-shot learning. In: Proceedings of the IEEE International Conference on Computer Vision, 2021. 122–131
33 Verma V K, Arora G, Mishra A, et al. Generalized zero-shot learning via synthesized examples. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 4281–4289
34 Xian Y Q, Lorenz T, Schiele B, et al. Feature generating networks for zero-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 5542–5551
35 Xian Y Q, Sharma S, Schiele B, et al. f-VAEGAN-D2: a feature generating framework for any-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. 10275–10284
36 Zhu Y Z, Elhoseiny M, Liu B C, et al. A generative adversarial approach for zero-shot learning from noisy texts. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 1004–1013
37 Yu Y L, Ji Z, Han J G, et al. Episode-based prototype generating network for zero-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020. 14035–14044
38 Li J J, Jing M M, Lu K, et al. Leveraging the invariant side of generative zero-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. 7394–7403
39 Ji Z, Yan J T, Wang Q, et al. Triple discriminator generative adversarial network for zero-shot image classification. Sci China Inf Sci, 2021, 64: 120101
40 Chen S M, Chen S H, Hou W J, et al. EGANS: evolutionary generative adversarial network search for zero-shot learning. IEEE Trans Evol Computat, 2024, 28: 582–596
41 Xie G S, Liu L, Ji X B, et al. Attentive region embedding network for zero-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. 9384–9393
42 Zhu Y Z, Xie J W, Tang Z Q, et al. Semantic-guided multi-attention localization for zero-shot learning. In: Proceedings of Advances in Neural Information Processing Systems, 2019
43 Liu Y, Guo J S, Cai D, et al. Attribute attention for semantic disambiguation in zero-shot learning. In: Proceedings of the IEEE International Conference on Computer Vision, 2019. 6698–6707
44 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Proceedings of Advances in Neural Information Processing Systems, 2017
45 Pennington J, Socher R, Manning C D. GloVe: global vectors for word representation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2014. 1532–1543
46 He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 770–778
47 Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009. 248–255
48 Chao W L, Changpinyo S, Gong B Q, et al. An empirical study and analysis of generalized zeroshot learning for object recognition in the wild. In: Proceedings of European Conference on Computer Vision, 2016. 52–68
49 Patterson G, Hays J. SUN attribute database: discovering, annotating, and recognizing scene attributes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012. 2751–2758
50 Han Z Y, Fu Z Y, Chen S, et al. Contrastive embedding for generalized zero-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021. 2371–2381
51 Feng Y G, Huang X W, Yang P B, et al. Non-generative generalized zero-shot learning via task-correlated disentanglement and controllable samples synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022. 9346–9355
52 Chen S M, Xie G S, Liu Y, et al. HSVA: hierarchical semantic-visual adaptation for zero-shot learning. In: Proceedings of Advances in Neural Information Processing Systems, 2021. 16622–16634
53 Chen S M, Hong Z M, Xie G S, et al. MSDN: mutually semantic distillation network for zero-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022. 7612–7621
54 van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res, 2008, 9: 2579–2605
55 Narayan S, Gupta A, Khan F S, et al. Latent embedding feedback and discriminative features for zero-shot classification. In: Proceedings of European Conference on Computer Vision, 2020